The SAGE Encyclopedia of

Educational Research, Measurement, and Evaluation

Edited by

Bruce B. Frey

The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Editorial Board

The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation

Editor

Bruce B. Frey
University of Kansas

SAGE reference

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Los Angeles
London
New Delhi
Singapore
Washington DC
Melbourne

FOR INFORMATION:

SAGE Publications, Inc.

2455 Teller Road

Thousand Oaks, California 91320

E-mail: order@sagepub.com

SAGE Publications Ltd.

1 Oliver’s Yard

55 City Road

London, EC1Y 1SP

United Kingdom

SAGE Publications India Pvt. Ltd.

B 1/I 1 Mohan Cooperative Industrial Area Mathura Road, New Delhi 110 044

India

SAGE Publications Asia-Pacific Pte. Ltd.

3 Church Street

#10-04 Samsung Hub

Singapore 049483

Library of Congress Cataloging-in-Publication Data Names: Frey, Bruce B., editor.

Title: The Sage encyclopedia of educational research, measurement, and evaluation/editor, Bruce B. Frey, University of Kansas.

Description: Thousand Oaks, California : SAGE Publications, Inc., [2018] | Includes bibliographical references and index.

Identifiers: LCCN 2017042016 | ISBN 978-1-5063-2615-3 (hardcover : alk. paper) Subjects: LCSH: Education—Research—Methodology.

Classification: LCC LB1028 .S139 2018 | DDC 370.72—dc23 LC record available at https://lccn.loc.gov/2017042016

Printed in the United States of America.

Acquisitions Editor: Andrew Boney Developmental Editor: Shirin Parsavand Assistant Editor: Jordan Enobakhare Reference Systems Manager: Leticia Gutierrez Production Editor: David C. Felts Copy Editor: Hurix Digital Typesetter: Hurix Digital Proofreaders: Lawrence W. Baker, Barbara Coster, Eleni Georgiou, Sally Jaskold Indexer: Joan Shapiro

Cover Designer: Anupama Krishnan Marketing Manager: Jennifer Bourque

Contents

List of Entries

Reader’s Guide

Assessment

Assessment Issues

Assessment Methods

Computerized Adaptive Testing
Confidence Interval
Curriculum-Based Assessment
Diagnostic Tests
Difficulty Index
Discrimination Index
English Language Proficiency Assessment
Formative Assessment
Intelligence Tests
Interquartile Range
Minimum Competency Testing
Mood Board
Personality Assessment
Power Tests
Progress Monitoring
Projective Tests
Psychometrics
Reading Comprehension Assessments
Screening Tests
Self-Report Inventories
Sociometric Assessment
Speeded Tests
Standards-Based Assessment
Summative Assessment
Technology-Enhanced Items
Test Battery
Testing, History of
Tests
Value-Added Models
Written Language Assessment

Classroom Assessment

Authentic Assessment
Backward Design
Bloom’s Taxonomy
Classroom Assessment
Constructed-Response Items
Curriculum-Based Measurement

Item Response Theory

Reliability

Scores and Scaling

Standardized Tests

Validity

Cognitive and Affective Variables

Data Visualization Methods

Disabilities and Disorders

Distributions

Educational Policies

Evaluation Concepts

Evaluation Designs

Human Development

Instrument Development

Organizations and Government Agencies

Joint Committee on Standards for Educational Evaluation
National Council on Measurement in Education
National Science Foundation
Office of Elementary and Secondary Education
Organisation for Economic Cooperation and Development
Partnership for Assessment of Readiness for College and Careers
Smarter Balanced Assessment Consortium
Teachers' Associations
U.S. Department of Education
World Education Research Association

Professional Issues

Accountability
Certification
Classroom Observations
Compliance
Confidentiality
Conflict of Interest
Data-Driven Decision Making
Diagnostic and Statistical Manual of Mental Disorders
Educational Researchers, Training of
Ethical Issues in Educational Research
Ethical Issues in Evaluation
Ethical Issues in Testing
Evaluation Consultants
Federally Sponsored Research and Programs
Framework for Teaching
Guiding Principles for Evaluators
Licensure
Professional Development of Teachers
Professional Learning Communities
School Psychology
Standards for Educational and Psychological Testing
Teacher Evaluation
Teachers' Associations

Publishing

Qualitative Research

Research Concepts

Research Designs

Research Methods

Research Tools

Social and Ethical Issues

Social Network Analysis

Statistics

Bayesian Statistics

Statistical Analyses

Statistical Concepts

Statistical Models

Teaching and Learning

Theories and Conceptual Frameworks

Theories and Conceptual Frameworks

Threats to Research Validity

About the Editor

Bruce B. Frey, Ph.D.,

is an award-winning researcher, teacher, and professor of educational psychology at the University of Kansas. He is the author of *There's a Stat for That!, Modern Classroom Assessment,* and *100 Questions (and Answers) about Tests and Measurement* for SAGE and associate editor of SAGE's *Encyclopedia of Research Design*. He also wrote *Statistics Hacks* for O'Reilly Media. His primary research interests include classroom assessment, instrument development, and program evaluation. In his spare time, Bruce leads a secret life as Professor Bubblegum, host of *Echo Valley*, a podcast that celebrates bubblegum pop music of the late 1960s. The show is wildly popular with the young people.

Contributors

Tineke A. Abma
VU University Medical Center
Samuel E. Abrams
National Center for the Study of Privatization in Education
Phillip L. Ackerman
Georgia Institute of Technology
Robert A. Ackerman
University of Texas at Dallas
Christopher M. Adams
Fitchburg State University
Allison Jennifer Ames
James Madison University
Ji An
University of Maryland, College Park
Samantha F. Anderson
University of Notre Dame
Heidi Arnouts
University of Antwerp
Raman Arora
Johns Hopkins University
Diana J. Arya
University of California, Santa Barbara
Patricia Teague Ashton
University of Florida
Tony Attwood
The Minds and Hearts Clinic
Marilyn M. Ault
University of Kansas
Karen Badger
University of Kentucky
Alison L. Bailey
University of California, Los Angeles
Ella G. Banda
University of Massachusetts Amherst
Brian R. Barber

Kent State University

Nicole Barnes
Montclair State University

Gail Vallance Barrington
Barrington Research Group, Inc.

Jordan R. Bass
University of Kansas

A. Alexander Beaujean
Baylor University

Danielle M. Becker
University of Minnesota

Thomas J. Beckman
Mayo Clinic

Eric T. Beeson
Northwestern University

Linda S. Behar-Horenstein
University of Florida

John Bell
Michigan State University

Aarti Bellara
University of Connecticut

Nicholas F. Benson
Baylor University

Peter M. Bentler
University of California, Los Angeles

Dale E. Berger
Claremont Graduate University

Jacquelyn A. Bialo
Georgia State University

Magdalena Bielenia-Grajewska
University of Gdansk

Scott Bishop
Questar Assessment Inc.

Bruce E. Blaine
St. John Fisher College

Shane D. Blair
University of Kansas

Lorraine Blatt
Johns Hopkins University

Joseph K. Blitzstein
Linda Dale Bloomberg
Teachers College, Columbia University
David Bloome
The Ohio State University
Clive R. Boddy
Middlesex University
Christian Bokhove
University of Southampton
Stella Bollmann
Ludwig Maximilians University
Edward L. Boone
Virginia Commonwealth University
Marc H. Bornstein
National Institute of Child Health and Human Development
Robert Boruch
University of Pennsylvania
Frank A. Bosco Jr.
Virginia Commonwealth University
Alex J. Bowers
Teachers College, Columbia University
Richard E. Boyatzis
Case Western Reserve University
Michelle L. Boyer
University of Massachusetts Amherst
Nancy N. Boyles
Southern Connecticut State University
Laine Bradshaw
University of Georgia
Alisa Palmer Branham
University of Kansas
Mason Lee Branham
University of South Carolina
Michael T. Brannick
University of South Florida
Markus Brauer
University of Wisconsin-Madison
Robert L. Brennan
University of Iowa

Tonya Breymier

Indiana University East

Ali S. Brian

University of South Carolina

Sharon Brisolara

Shasta College

Angela Broaddus

Benedictine College

Alex Brodersen

University of Notre Dame

Jeanne Brooks-Gunn

Teachers College, Columbia University

James Dean Brown

University of Hawaii at Manoa

Julie C. Brown

University of Minnesota, Twin Cities

Mary T. Brownell

University of Florida

Alan W. Brue

Capella University

Jennifer A. Brussow

University of Kansas

Kelley Buchheister

University of Nebraska–Lincoln

Frederick Burrack

Kansas State University

Gervase R. Bushe

Simon Fraser University

Andrew C. Butler

University of Texas at Austin

Yuko Goto Butler

University of Pennsylvania

Li Cai

University of California, Los Angeles

Meghan K. Cain

University of Notre Dame

Gregory L. Callan

Ball State University

Mitchell Campbell

University of Wisconsin-Madison
Emily Cantwell
University of Kansas
Kimberly Capp
Nova Southeastern University
Nicole Mittenfelner Carl
University of Pennsylvania
Thomas R. Carretta
Air Force Research Laboratory
Arthur Charpentier
Université du Québec à Montréal
Walter Chason
University of South Florida
Helen L. Chen
Stanford University
Jie Chen
University of Kansas
Yi-Hsin Chen
University of South Florida
Ying Cheng
University of Notre Dame
Yuk Fai Cheong
Emory University
Monaliza M. Chian
University of California, Santa Barbara
Shriniwas Chinta
South West Sydney Area Health Service
Mary M. Chittooran
Saint Louis University
Theodore J. Christ
University of Minnesota
Christine Ann Christle
University of South Carolina
Amy Clark
University of Kansas
Nathan H. Clemens
University of Texas at Austin
Hamish Coates
University of Melbourne
Jill S. M. Col

Jill S. M. Coleman
Ball State University
Zachary K. Collier
University of Florida
Eric Alan Common
University of Kansas
Zachary Conrad
Kansas Multi-Tier System of Supports
Bryan G. Cook
University of Hawaii at Manoa
Kyrsten M. Costlow
National Institute of Child Health and Human Development
Matthew Gordon Ray Courtney
Melbourne Graduate School of Education
Nelson Cowan
University of Missouri
Jana Craig-Hare
University of Kansas
Bonnie Cramond
University of Georgia
Kent J. Crippen
University of Florida
Kevin Crouse
Rutgers Graduate School of Education
Toni Crouse
University of Kansas
Zhongmin Cui
ACT, Inc.
Steven Andrew Culpepper
Arthur J. Cunningham
St. Olaf College
Nicholas A. Curtis
James Madison University
Joshua A. Danish
Indiana University Bloomington
Cynthia S. Darling-Fisher
University of Michigan
Judith Davidson
University of Massachusetts Lowell
Jennifer Davidtz

Jennifer Davidtz
Nova Southeastern University
Larry Davis
Educational Testing Service
Michael E. Dawson
University of Southern California
Anita B. Delahay
Carnegie Mellon University
Christine E. DeMars
James Madison University
Leah Dembitzer
Rutgers, The State University of New Jersey
Angelo S. DeNisi
Tulane University
Justin A. DeSimone
University of Cincinnati
Maria DeYoreo
Duke University
Ronli Diakow
New York City Department of Education
Kathryn Doherty Kurtz
University of Massachusetts Boston
Thurston A. Domina
University of North Carolina at Chapel Hill
Ashley Donohue
Baylor University
Neil J. Dorans
Educational Testing Service
John F. Dovidio
Yale University
Lyman L. Dukes III
University of South Florida St. Petersburg
Danielle N. Dupuis
University of Minnesota
Stephanie Dyson Elms
University of Kansas
John Joseph Dziak
Pennsylvania State University
Meghan Ecker-Lyster
University of Kansas

University of Kansas

Julianne Michelle Edwards
Azusa Pacific University

Anna J. Egalite
North Carolina State University

Thorlene Egerton
The University of Melbourne

Valeisha M. Ellis
Spelman College

Susan E. Embretson
Georgia Institute of Technology

Amy S. Gaumer Erickson
University of Kansas

Eduardo Estrada
Camilo José Cela University

Kimberly Ethridge
Nova Southeastern University

Howard T. Everson
SRI International

Leandre R. Fabrigar
Queen’s University

Carl Francis Falk
Michigan State University

Fen Fan
University of Massachusetts Amherst

Anna C. Faul
University of Louisville

John M. Ferron
University of South Florida

David Fetterman
Fetterman & Associates

W. Holmes Finch
Ball State University

Roger Fischer
Montana State University

Helenrose Fives
Montclair State University

Sara A. Florence
Nova Southeastern University

Timothy Franz

Timothy Franz
St. John Fisher College
Bruce B. Frey
University of Kansas
Alon Friedman
University of South Florida
Catherine O. Fritz
University of Northampton
Kyra N. Fritz
University of Louisville
John Mark Froiland
Pearson
Danling Fu
University of Florida
Matthew B. Fuller
Sam Houston State University
Gavin W. Fulmer
University of Iowa
Joseph Calvin Gagnon
University of Florida
April Galyardt
University of Georgia
Copelan Gammon
National Institutes of Health
Alejandra Garcia
University of Massachusetts Amherst
Andrea M. Garcia
University of Kansas
Rachel Darley Gary
University of Massachusetts Amherst
Pat J. Gehrke
University of South Carolina
Nicholas W. Gelbar
University of Connecticut
Claudia A. Gentile
NORC at the University of Chicago
Elizabeth T. Gershoff
University of Texas at Austin
Dean R. Gerstein
Independent Scholar

Iman Ghaderi
University of Arizona
Graham R. Gibbs
University of Huddersfield
Drew Gitomer
Rutgers Graduate School of Education
Lina Goldenberg
University of Kansas
Samantha B. Goldstein
National Institute of Child Health and Human Development
Juana Gómez-Benito
University of Barcelona
Roland H. Good
Dynamic Measurement Group
Jacqueline D. Goodway
The Ohio State University
Brian S. Gordon
University of Kansas
Chad M. Gotch
Washington State University
Bruce Granshaw
Victoria Universtiy
Judith L. Green
University of California, Santa Barbara
Jennifer C. Greene
University of Illinois at Urbana-Champaign
Judith M. S. Gross
University of Kansas
Fei Gu
McGill University
Cassandra Guarino
University of California, Riverside
Daniel B. Hajovsky
University of South Dakota
Kevin A. Hallgren
University of Washington
Marc Hallin
Université libre de Bruxelles
Lawrence C. Hamilton

University of New Hampshire
K. Chris Han
Graduate Management Admission Council
Carl B. Hancock
The University of Alabama
Gregory R. Hancock
University of Maryland, College Park
Maggie Quinn Hannan
University of Pittsburgh
Brenda Hannon
Texas A&M University-Kingsville
David M. Hansen
University of Kansas
Shlomo Hareli
University of Haifa
Lisa L. Harlow
University of Rhode Island
Jeffrey R. Harring
University of Maryland, College Park
Heather D. Harris
James Madison University
Judith R. Harrison
Rutgers, The State University of New Jersey
Jessica P. Harvey
Southern Illinois University Edwardsville
Richard D. Harvey
Saint Louis University
John D. Hathcoat
James Madison University
Clifford E. Hauenstein
Georgia Institute of Technology
Ellen Hazelkorn
Dublin Institute of Technology
Dan He
University of Kansas
Lauren M. Henry
National Institute of Child Health and Human Development
Socorro Herrera
Kansas State University

Michael Herriges
University of Minnesota
Salome Heyward
Salome Heyward and Associates
Tyler Hicks
University of Kansas
M. Dolores Hidalgo
University of Murcia
Rana S. Hinman
University of Melbourne
John M. Hintze
University of Massachusetts Amherst
Tyrell Hirchert
University of Northern Colorado
David C. Hoaglin
University of Massachusetts Medical School
Michael F. Hock
University of Kansas
Janice A. Hogle
University of Wisconsin-Madison
Søren Holm
University of Manchester
S. Jeanne Horst
James Madison University
Jessica Hoth
Universität Vechta
Carrie R. Houts
Vector Psychometric Group, LLC
Kenneth R. Howe
University of Colorado Boulder
Lindsay Till Hoyt
Fordham University
Mei Hoyt
University of North Texas
Hsiu-Fang Hsieh
Fooyin University
Bo Hu
University of Kansas
Anne Corinne Huggins-Manley
University of Florida

University of Florida
Ben P. Hunter
University of Kansas School of Medicine-Wichita
R. Shane Hutton
Vanderbilt University
Dragos Iliescu
University of Bucharest
Paul B. Ingram
Texas Tech University
Dianne Nutwell Irving
Georgetown University
S. Earl Irving
Kia Eke Panuku, The University of Auckland
Dan Ispas
Illinois State University
Jessica N. Jacovidis
James Madison University
Justin Jager
Arizona State University
Lilli Japec
Statistics Sweden
Gerard Michael Jellig
University of Pennsylvania
Patricia A. Jenkins
Albany State University
Rebecca Jesson
University of Auckland
Jennifer L. Jewiss
University of Vermont
Hong Jiao
University of Maryland, College Park
Daniela Jiménez
Pontificia Universidad Católica de Chile
Maria Jimenez-Buedo
Universidad Nacional de Educación a Distancia
Adam Michael Johansen
University of Warwick
Jeffrey P. Johnson
Educational Testing Service
Matthew S. Johnson

Matthew S. Johnson
Teachers College, Columbia University
Paul E. Johnson
University of Kansas
Robert L. Johnson
University of South Carolina
Tessa Johnson
University of Maryland, College Park
Natalie D. Jones
Claremont Graduate University
Nathan D. Jones
Boston University
Seang-Hwane Joo
University of South Florida
Jeanette Joyce
Rutgers, The State University of New Jersey
Diana Joyce-Beaulieu
University of Florida
George Julnes
University of Baltimore
Hyun Joo Jung
University of Massachusetts Amherst
Uta Jüttner
Hochschule Luzern
David Kahle
Baylor University
Irene Kaimi
Plymouth University
Matthew P. H. Kan
Queen’s University
Jeffrey D. Karpicke
Purdue University
Arunprakash T. Karunanithi
University of Colorado Denver
Meagan Karvonen
University of Kansas
Kentaro Kato
Benesse Educational Research and Development
Daniel Katz
University of California, Santa Barbara

University of California, Santa Barbara
Ian Katz
Saint Louis University
Irvin R. Katz
Educational Testing Service
Mira B. Kaufman
National Institute of Child Health and Human Development
Walter Keenan
University of Connecticut
Harrison J. Kell
Educational Testing Service
Jessie Kember
University of Minnesota
Ana H. Kent
Saint Louis University
Ryan J. Kettler
Rutgers, The State University of New Jersey
Haeyoung Kim
Korea University
Hyung Jin Kim
University of Iowa
Hyung Won Kim
University of Texas Rio Grande Valley
Minkyoung Kim
Indiana University Bloomington
Yoon Jeon Kim
Massachusetts Institute of Technology
Neal Kingston
University of Kansas
Allyson J. Kiss
University of Minnesota
Karla Kmetz-Morris
University of South Florida St. Petersburg
Olga Korosteleva
California State University, Long Beach
Rachel Elizabeth Kostura Polk
University of Kansas
Laura M. B. Kramer
University of Kansas
Parvati Krishnamurty

Parvati Krishnamurty
NORC at the University of Chicago

Jeffrey D. Kromrey
University of South Florida

Ivar Krumpal
University of Leipzig

B. Venkatesh Kumar
Tata Institute of Social Sciences

Swapna Kumar
University of Florida

Michael Kung
University of Florida

Lori Kupczynski
Texas A&M University-Kingsville

Carrie La Voy
University of Kansas

Chi Yan Lam
Queen’s University

Kathleen Lynne Lane
University of Kansas

Hongling Lao
University of Kansas

Lotta C. Larson
Kansas State University

Norman J. Lass
West Virginia University

Brandon LeBeau
University of Iowa

James M. LeBreton
Pennsylvania State University

Thomas Ledermann
Utah State University

Kerry Lee
University of Auckland

Lina Lee
University of New Hampshire

Lisa Lee
NORC at the University of Chicago

Won-Chan Lee
University of Iowa

Walter L. Leite
University of Florida
Hildie Leung
The Hong Kong Polytechnic University
Janet Tsin-yee Leung
The Hong Kong Polytechnic University
Daniel Lewis
Pacific Metrics Corporation
Chen Li
University of Maryland, College Park
Isaac Y. Li
University of South Florida
Jianqiang Liang
The Hong Kong Polytechnic University
Rosemary Luyin Liang
The Hong Kong Polytechnic University
Dandan Liao
University of Maryland, College Park
Gregory Arief D. Liem
National Institute of Education, Nanyang Technological University
Nicholas K. Lim
Spalding University
TickMeng Lim
Open University Malaysia
Li Lin
The Hong Kong Polytechnic University
Sheila K. List
Virginia Commonwealth University
Haiyan Liu
University of Notre Dame
Jingchen Liu
Columbia University
Shengtao Liu
Hunan University
Xiaofeng Steven Liu
University of South Carolina
Yang Liu
University of California, Merced
Sarah Lockenvitz

Missouri State University
Jill Hendrickson Lohmeier
University of Massachusetts Lowell
Stephen W. Loke
University of Kansas Medical Center
Patricia D. López
San José State University
Sue Lottridge
Pacific Metrics Corporation
Patricia A. Lowe
University of Kansas
Richard M. Luecht
University of North Carolina at Greensboro
Lars Lyberg
Lyberg Survey Quality Management Inc.
Cecilia Ma
The Hong Kong Polytechnic University
David P. MacKinnon
Arizona State University
Joseph Madaus
University of Connecticut
Patrick Mair
Harvard University
Matthew C. Makel
Duke University
Christoforos Mamas
University of California, San Diego
Gregory J. Marchant
Ball State University
Michael O. Martin
Boston College
Phillip K. Martin
University of Kansas School of Medicine–Wichita
Julie Masterson
Missouri State University
Andrew Maul
University of California, Santa Barbara
Brendan Maxcy
Indiana University–Purdue University Indianapolis

Joseph A. Maxwell
George Mason University
Scott E. Maxwell
University of Notre Dame
Rebecca Mazur
University of Massachusetts Amherst
Dominica McBride
Become
Michael A. McDaniel
Virginia Commonwealth University
Andrew McEachin
RAND Corporation
Elizabeth H. McEneaney
University of Massachusetts Amherst
Ryan J. McGill
College of William and Mary
Jamie C. McGovern
University of Kansas Medical Center
Mary L. McHugh
Angeles College
Alex McInturff
University of California, Berkeley
James McLeskey
University of Florida
Miles Allen McNall
Michigan State University
David E. Meens
University of Colorado Boulder
Valerie Meier
University of California, Santa Barbara
Daryl F. Mellard
University of Kansas
Krystal Mendez
University of Kansas
Sylvia L. Mendez
University of Colorado Colorado Springs
Natalja Menold
GESIS–Leibniz Institute for the Social Sciences
Margaret Kristin Merga

Murdoch University
Gabriel J. Merrin
University of Illinois at Urbana-Champaign
Craig A. Mertler
Arizona State University
Audrey Michal
Northwestern University
Milica Miočević
Arizona State University
Gregory Mitchell
University of Virginia
Monica Morell
University of Maryland, College Park
David Morgan
Spalding University
Demetri L. Morgan
Loyola University Chicago
Grant B. Morgan
Baylor University
Carl N. Morris
Harvard University
Peter E. Morris
Lancaster University and University of Northampton
Kristin M. Morrison
Georgia Institute of Technology
Wilfridah Mucherah
Ball State University
Jamie R. Mulkey
Caveon, LLC
Ina V. S. Mullis
Boston College
Karen D. Multon
University of Kansas
Sohad Murrar
University of Wisconsin-Madison
Angela K. Murray
University of Kansas
Brittany Murray
University of North Carolina at Chapel Hill
Dorothy J. Musselwhite

Dorothy J. Musselwhite
University of Georgia

Jessica Namkung
University of Nebraska–Lincoln

Oksana Naumenko
University of North Carolina at Greensboro

Mario A. Navarro
Claremont Graduate University

Kelli L. Netson
University of Kansas School of Medicine–Wichita

Kirsten Newell
University of Minnesota

Joan Newman
University at Albany

Anh Andrew Nguyen
Queen’s University

Thu Suong Nguyen
Indiana University–Purdue University Indianapolis

Joseph R. Nichols
Saint Louis University

Nicole M. Nickens
University of Central Missouri

Kyle Nickodem
University of Minnesota

Kathleen H. Nielsen
University of Washington

Christopher R. Niileksela
University of Kansas

Kim Nimon
University of Texas at Tyler

Patricia M. Noonan
University of Kansas

Anthony Odland
Sanford Health

Laura O’Dwyer
Boston College

Insu Paek
Florida State University

Qianqian Pan
University of Kansas

University of Kansas
Ming Fai Pang
The University of Hong Kong
Eugene T. Parker
University of Kansas
Sarah Parsons
University of Southampton
Tracy Paskiewicz
University of Massachusetts Boston
Meagan M. Patterson
University of Kansas
Michael Quinn Patton
Utilization-Focused Evaluation
Trena M. Paulus
University of Georgia
Phillip D. Payne
Kansas State University
Melissa Pearrow
University of Massachusetts Boston
Mark Pedretti
Claremont Graduate University
Beverly Pell
University of Kansas
Peng Peng
University of Nebraska–Lincoln
Marianne Perie
University of Kansas
Laura Pevytoe
National Institutes of Health
Lia Plakans
University of Iowa
Anthony Jason Plotner
University of South Carolina
Jonathan A. Plucker
Johns Hopkins University
Kelvin Terrell Pompey
University of South Carolina
Michael I. Posner
University of Oregon
Dmitriy Poznyak

Dmitriy Poznyak
Mathematica Policy Research
Ludmila N. Praslova
Vanguard University of Southern California
Christopher Prickett
Texas A&M University, College Station
Susan Prion
University of San Francisco
Joshua N. Pritikin
Virginia Commonwealth University
Ana Puig
University of Florida
Elisabeth M. Pyburn
James Madison University
Patrick Radigan
University of Colorado Colorado Springs
Kelsey Ragan
Texas A&M University
Sharon F. Rallis
University of Massachusetts Amherst
Jennifer Randall
University of Massachusetts Amherst
David L. Raunig
ICON Clinical Research
Sharon M. Ravitch
University of Pennsylvania
Jason Ravitz
Google Inc
Randall Reback
Barnard College
Lynne M. Reder
Carnegie Mellon University
Malcolm James Ree
Our Lady of the Lake University
Charles M. Reigeluth
Indiana University Bloomington
Sally M. Reis
University of Connecticut
Rachel L. Renbarger
Baylor University

Baylor University
Matthew R. Reynolds
University of Kansas
Melissa N. Richards
National Institute of Child Health and Human Development
John T. E. Richardson
The Open University, UK
Robert D. Ridge
Brigham Young University
Rigoberto Rincones-Gómez
University of North Carolina Wilmington
Joseph A. Rios
Educational Testing Service
Francisco L. Rivera-Batiz
Columbia University
Andrew T. Roach
Georgia State University
L. Danielle Roberts-Dahm
University of South Florida St. Petersburg
Michael C. Rodriguez
University of Minnesota
Liliana Rodríguez-Campos
University of South Florida
Mary Roduta Roberts
University of Alberta
Bradley D. Rogers
University of South Carolina
Jonathan D. Rollins
University of North Carolina at Greensboro
Jeanine Romano
American Board of Pathology
Benjamin D. Rosenberg
Chapman University
Jeffrey N. Rouder
University of Missouri
Amber Rowland
University of Kansas
David J. Royer
University of Kansas
Donald B. Rubin

Harvard University
Cort W. Rudolph
Saint Louis University
Jennifer Lin Russell
University of Pittsburgh
Tonya Rutherford-
Hemming
University of North Carolina at Greensboro
Thomas G. Ryan
Nipissing University
Falak Saffaf
Saint Louis University
Stephen Keith Sagarin
Berkshire Waldorf School
Johnny Saldaña
Arizona State University
Asmalina Saleh
Indiana University Bloomington
Courtney B. Sanders
James Madison University
Massimiliano Sassoli de Bianchi
Laboratorio di Autoricerca di Base
Dorothea Schaffner
University of Applied Science Lucerne
Anne M. Schell
Occidental College
Heather Schmitt
Michigan State University
Stephanie Schmitz
University of Northern Iowa
Rachel Watkins Schoenig
Cornerstone Strategies, LLC
Ryan W. Schroeder
University of Kansas School of Medicine–Wichita
Michael A. Seaman
University of South Carolina
Kathleen Sexton-Radek
Elmhurst College
Nichola Shackleton

University of Auckland
Priti Shah
University of Michigan
Sarah Shannon
University of Washington
Can Shao
University of Notre Dame
Daniel Tan-lei Shek
The Hong Kong Polytechnic University
Mark D. Shermis
University of Houston–Clear Lake
Galit Shmueli
National Tsing Hua University
Nicholas J. Shudak
University of South Dakota
Boaz Shulruf
University of New South Wales
Vivian Shyu
University of Colorado Denver
Jason T. Siegel
Claremont Graduate University
Satoko Siegel
Western University of Health Sciences
Timothy C. Silva
Claremont Graduate University
Julius Sim
Keele University
Lucinda Simmons
Elmhurst College
Stephen G. Sireci
University of Massachusetts Amherst
Julie Slayton
University of Southern California
Stephanie Snidarich
University of Minnesota
Brian Song
California State University, Long Beach
Nancy Butler Songer
Drexel University

J. E. R. Staddon
Duke University and the University of York
Laura M. Stapleton
University of Maryland, College Park
Rachel M. Stein
University of California, Santa Barbara
Douglas Steinley
University of Missouri
Robert J. Sternberg
Cornell University
David W. Stewart
Loyola Marymount University
David W. Stockburger
Missouri State University
Vera Lynne Stroup-Rentier
Kansas State Department of Education
Richard R. Sudweeks
Brigham Young University
Cindy Suurd Ralph
Queen’s University
Sruthi Swami
University of California, Santa Barbara
Cara N. Tan
Claremont Graduate University
Ser Hong Tan
National Institute of Education
Michael Tang
University of Colorado Denver
Maciej Taraday
Jagiellonian University
Charlotte Tate
San Francisco State University
Mohsen Tavakol
The University of Nottingham, School of Medicine
Mark S. Teachout
University of the Incarnate Word
David Teira
Universidad Nacional de Educación a Distancia
Lee Teitel

Harvard University
Alexandru C. Telea
University of Groningen
Michael S. Ternes
University of Kansas
Jordan Thayer
University of Minnesota
Lori A. Thombs
University of Missouri
W. Jake Thompson
University of Kansas
Martha L. Thurlow
University of Minnesota
Gail Tiemann
University of Kansas
Rebecca Tipton
Baylor University
Sara Tomek
The University of Alabama
David Torres Irribarra
Pontificia Universidad Católica de Chile
Meng-Jung Tsai
National Taiwan University of Science & Technology
Kayla Tureson
Sanford Health
Yvonne H. M. van den Berg
Behavioural Science Institute, Radboud University
Thomas I. Vaughan-Johnston
Queen’s University
Frank R. Vellutino
University at Albany
Aldert Vrij
University of Portsmouth
Jonathan Wai
Duke University Talent Identification Program
Breanna A. Wakar
Mathematica Policy Research
Zachary Walker
National Institute of Education
[illegible]

Jacqueline Remondet Wall
American Psychological Association
Ryan W. Walters
Creighton University
Lisi Wang
University of Texas at Austin
Xi Wang
University of Massachusetts Amherst
Yan Wang
University of South Florida
Emily Ward
Baylor University
Jackie Waterfield
Keele University
Kathryn Weaver
University of New Brunswick
David J. Weiss
University of Minnesota
Craig Stephen Wells
University of Massachusetts Amherst
Jenny C. Wells
University of Hawaii at Manoa
Megan E. Welsh
University of California, Davis
Brian C. Wesolowski
University of Georgia
Colin P. West
Mayo Clinic
David Westfall
Emporia State University
Anna Wieczorek-Taraday
Nencki Institute of Experimental Biology
Christina Wikström
Umeå university
Magnus Wikström
Umeå university
Immanuel Williams
Rutgers, The State University of New Jersey
Pamela Williamson
University of North Carolina at Greensboro

University of North Carolina at Greensboro

Linda Wilmshurst
The Center for Psychology

Stefanie A. Wind
The University of Alabama

Steven L. Wise
Northwest Evaluation Association

Sara E. Witmer
Michigan State University

James Wollack
University of Wisconsin-Madison

Kenneth K. Wong
Brown University

Rebecca H. Woodland
University of Massachusetts Amherst

Heather H. Woodley
New York University

Annette Woods
Queensland University of Technology

Florence Wu
The Hong Kong Polytechnic University

Jing Wu
The Hong Kong Polytechnic University

Yi-Fang Wu
ACT, Inc.

Gongjun Xu
University of Minnesota

Inbal Yahav
Bar Ilan University

Ji Seung Yang
University of Maryland, College Park

Yang Lydia Yang
Kansas State University

Yanyun Yang
Florida State University

Brandon W. Youker
Grand Valley State University

Jing Yu
University of California, Santa Barbara

Elizabeth R. Zell

Elizabeth R. Zell
Stat-Epi Associates, Inc.
April L. Zenisky
University of Massachusetts Amherst
Hao Helen Zhang
University of Arizona
Kun Zhang
Carnegie Mellon University
Zhiyong Zhang
University of Notre Dame
Fei Zhao
The Citadel
Chunmei Zheng
Pearson
Xiaodi Zhou
University of Georgia
Qingqing Zhu
University of Kansas
Xiaoqin Zhu
The Hong Kong Polytechnic University
Michele F. Zimowski
NORC at the University of Chicago
Bruno D. Zumbo
University of British Columbia

Introduction

Educational Research, Measurement, and Evaluation

The title of this encyclopedia seems to cover at least three distinct fields, each of which could easily fill several volumes of its own. Indeed, there are several fine reference works that cover social science research methods, statistics, assessment and measurement, and program evaluation. Some of the nicer examples are already published by SAGE, in fact. Several encyclopedias cover the field of education in general, with a moderate amount of methodological entries, as well. However, these areas, in terms of their methods, activities, and typical variables of interest, are strongly entwined. Program evaluators use methods developed by educational researchers and develop instruments following measurement best practice. Many, maybe most, of the basic measurement statistics and foundational measurement principles were first conceived and developed by educational researchers. Program evaluators tend to be educational researchers, as well, and, of course, use the tools of social science research to reach their conclusions.

The strongest evidence, perhaps, of the symbiotic relationship among educational research, measurement, and program evaluation can be found in the faculty and courses at research universities. It is often the same professors who teach the research courses, the assessment courses, and the evaluation courses. The same is true of the scholarly journals in educational research, measurement, and evaluation. These journals routinely publish studies across all three areas.

Producing a single encyclopedia that covers the wide range of topics across these connected areas allows for a unique contextual dimension that promotes deeper understanding and allows for more effective learning. In addition to reference works, there are textbooks, handbooks, monographs, and other publications focused on various aspects of educational research, measurement, and evaluation, but to date, there exists no major reference guide for students, researchers, and grant writers new to the field or a particular methodology. The encyclopedia fills that gap. This is the first comprehensive A-to-Z reference work that fully explores methods specific to educational research, assessment, measurement, and evaluation. *The SAGE Encyclopedia of Educational Research,* *Measurement, and Evaluation* is comprehensive and integrates the three methodological areas of scholarship in the science of education. In an era of constant changes in state and federally driven curricular standards and high-stakes testing, a growing need for innovative instructional methods, increased reliance on data-driven decision-making and calls for accountability in research, a shared understanding of the methods of educational research, measurement, and evaluation is more important than ever.

Making an Encyclopedia

A project of this size takes many people and a long time. Once a publisher, like SAGE, realizes there is a need for a multi-volume reference work like this, they choose an editor, like me. I was likely chosen because I've taught and published across these areas of educational research, measurement, and evaluation.

My first step was to recruit a world-class group of expert advisors, leaders in their field who teach and publish in educational research, measurement and evaluation. I was fortunate to form an Advisory Board of these five nice and wise folks:

- Dr. Rebecca Woodland, University of Massachusetts Amherst
- Dr. Neal Kingston, University of Kansas
- Dr. Jill Lohmeier, University of Massachusetts Lowell
- Dr. William Skorupski, University of Kansas
- Dr. Jonathan Templin, University of Kansas

Together, we began to shape the encyclopedia. What topics or broad categories should be covered? We wanted the emphasis to be on methodology in educational research, measurement, and evaluation, but we also wanted the encyclopedia to cover important theories and common research variables. What entries should be included? Encyclopedia publishers call entry titles "headwords," and choosing these headwords was critical. In a four-volume encyclopedia there is only room for so many headwords (about 700), each headword can only be a certain number of words (about 500 to 3,000), and the right length of each headword varies depending on the importance of the entry. At the end of this process, we identified these broad topics as a framework for what belongs in the encyclopedia:

Assessment

Cognitive and Affective Variables
Data Visualization Methods
Disabilities and Disorders
Distributions
Educational Policies
Evaluation Concepts
Evaluation Designs
Human Development
Instrument Development
Organizations and Government Agencies
Professional Issues
Publishing
Qualitative Research
Research Concepts
Research Designs
Research Methods
Research Tools
Social and Ethical Issues
Social Network Analysis
Statistics
Teaching and Learning
Theories and Conceptual Frameworks
Threats to Research Validity

The reader's guide, near this introduction, lists all the entries, grouped by these categories, so you can find what you want quickly. Based on these topics, we began identifying entries that an encyclopedia of educational research, measurement, and evaluation should include. We then needed to find hundreds of experts to write the 691 entries in this work. The advisory board suggested names and reserved some entries for themselves, I took a few for myself and began identifying potential authors, and, in what turned out to be a smart move, I hired a bright doctoral student, Alan Nong, to help with the search process. For some entries, there were clear leaders in the field, or authors of key studies to recruit. For other general entries, we searched education, educational psychology, curriculum, and statistics departments at universities throughout the globe. We have some of the top scholars in their field among the authors of these entries.

Acknowledgments

Among the experts who authored entries for the encyclopedia are Robert Brennan writing about testing, Craig Mertler writing about standardized scores, Bruno Zumbo writing about psychometrics, Joseph Maxwell writing about mixed methods research and qualitative data analysis, Laura O’Dwyer and Donald Rubin writing about experimental design, Peter Bentler writing about structural equation modeling, and David Fetterman and Michael Quinn Patton writing about evaluation. And literally hundreds of others were willing to contribute, for which I am grateful. You can see a list of all contributors at the front of this encyclopedia. About two years was spent in the recruitment and writing process and the initial stages of review and editing before the encyclopedia was copyedited. I am so impressed with the quality and value of the work done by our authors. Thanks, sincerely, to all of them.

I am most grateful for the fine work and guidance of the friendly people at SAGE. It was Jim Brace-Thompson who, as acquisition editor, promoted the idea to SAGE and first got me involved, and Andrew Boney who took the project on, worked with layout and design, and kept things running smoothly. Jordan Enobakhare helped get things going. Anna Villaseñor taught me how to use the remarkable software system that lets people make encyclopedias. Most of the heavy lifting during the first year or two of this process was done by Susan Moskowitz, who somehow handled the recruitment and enlistment of hundreds of scholars, shepherded their submissions, and worked with their schedules. Shirin Parsavand agreed to be the developmental editor and, though I don’t think we’ve ever met (or been at the same conference at the same time), it seems like we shared an office and this has truly been a team effort with her. As we neared the finish line, David Felts, the production editor for the encyclopedia, took the reins to move us through the final gates. Thank you, SAGE editors and staff, as you all are really first rate and so good at what you do. Finally, I am so thankful for the material and moral support of my wife, Bonnie Johnson. She is a finer scholar than I.

Bruce B. Frey

A

Fei Zhao Fei Zhao Zhao, Fei

a Parameter *a* parameter

1

3

a Parameter

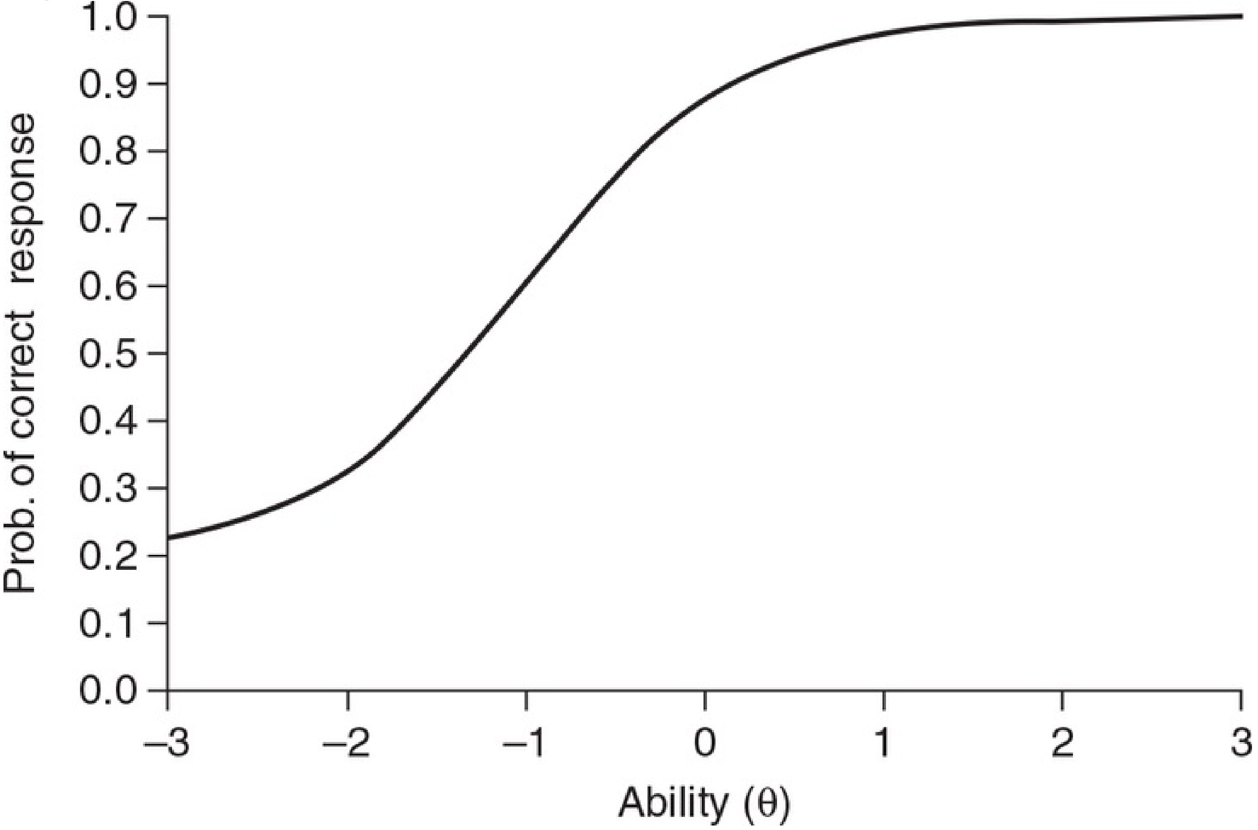
The *a* parameter, or the discrimination parameter, is one of the key item parameters in many item response theory (IRT) models. This entry discusses how the *a* parameter is defined and interpreted. For this discussion, properties of the *a* parameter are introduced in the unidimensional IRT framework in which items in a test measure one and only one construct, with a focus on dichotomous item responses. Realistically, items in a test are considered to be unidimensional as long as a single construct accounts for a substantial portion of the total score variance.

Unidimensional IRT framework is the focus of this entry because it is the foundation of its multidimensional counterpart, and basic principles in unidimensional IRT framework can be straightforwardly interpreted in the multidimensional context. In addition, this discussion concentrates on dichotomous item response models because they can be considered as the special cases of polytomous models, and when item responses become binary, polytomous models reduce to dichotomous models.

Defining the *a* Parameter

Using a mathematical formula, the IRT theory defines the probability of an examinee’s correct response to an item as a function of the latent ability of the examinee and that item’s properties. This function, the item characteristic curve (ICC), is also referred to as the item response function. An ICC defines a smooth nonlinear relationship between latent trait constructs (θ) and probability of a correct response. If assumptions are met, the ICCs can be stable over groups of examinees, and the θ scale also can be stable even when the test includes different items. A graphical representation of an ICC is given in Figure 1.

Figure 1 Example of an ICC



The generic mathematics function for Figure 1 is shown in Equation 1.

$$P(x_i = 1 \mid \theta) = c_i + (1 - c_i)\frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))},$$

where θ is the examinee ability parameter, c_i is often referred to as the pseudo-guessing or lower asymptote parameter with a value typically between 0 and 0.25, b_i is the location or difficulty parameter, and a_i is the discrimination or slope parameter. The parameter a_i indicates the steepness of ICC at $\theta = b$, where probability of correctly answering an item changes most rapidly. The logistic function presented in Equation 1 is called the three-parameter logistic (3PL) model, which was presented by Allan Birnbaum in a pioneering work by Frederic Lord and Melvin Novick in 1968. The two-parameter logistic (2PL; Equation 2) and one-parameter logistic (1PL) model (Equation 3) can be considered as special cases of the 3PL IRT model. As indicated in the corresponding formula, the 2PL model only has the *a* and the *b* parameter, whereas the 1PL model only has the *b* parameter and the *a* parameter is fixed.

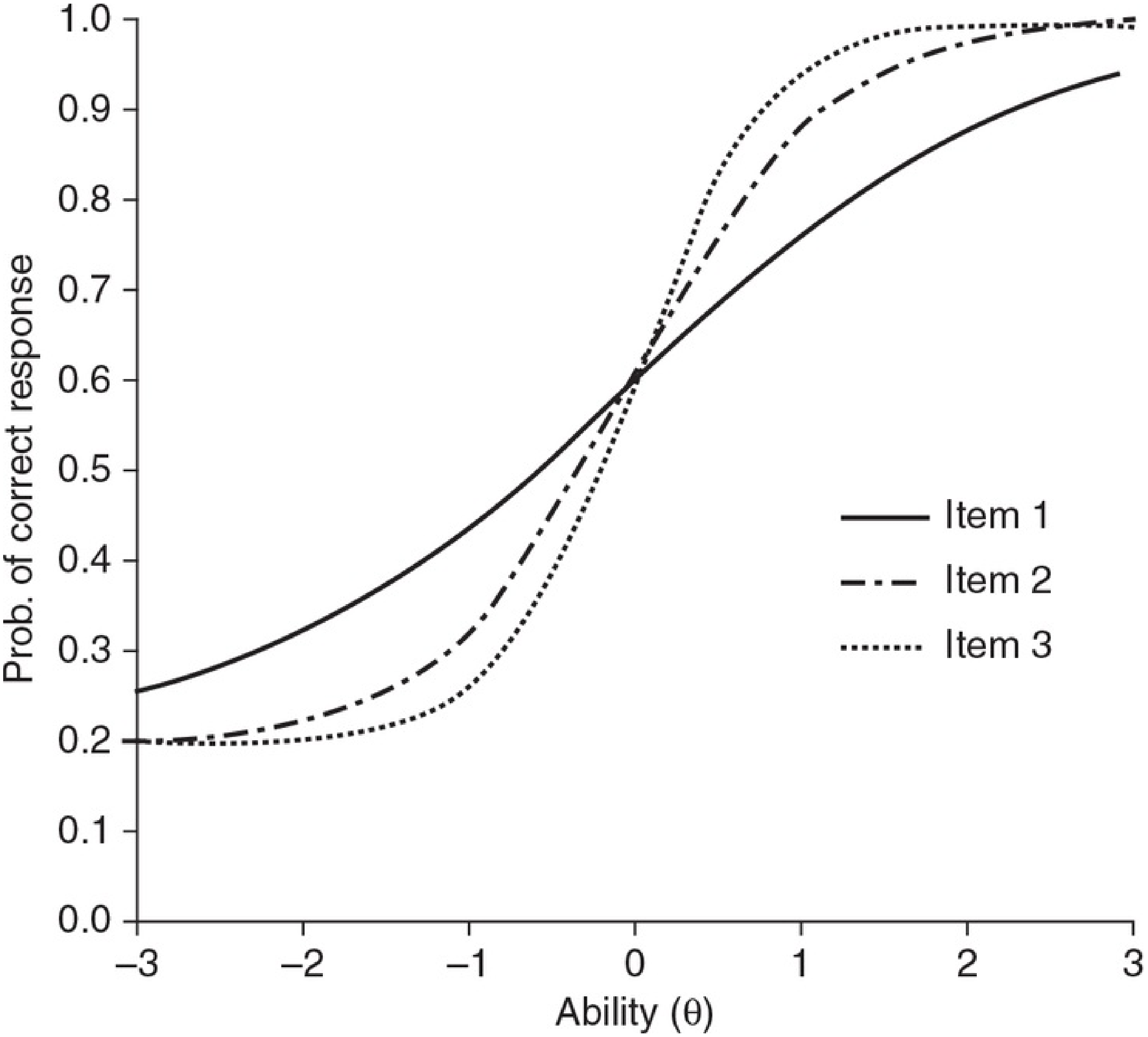
$$P(x_i = 1 \mid \theta) = \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))},$$

$$P(x_i = 1 \mid \theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}.$$

Interpreting the *a* Parameter

Figure 2 represents ICCs for three dichotomous items under the unidimensional 3PL IRT model. Among these 3 items, all *b*’s = 0 and all *c*’s = 0.2, but *a* values differ: a_1 = 0.5, a_2 = 1, and a_3 = 1.5.

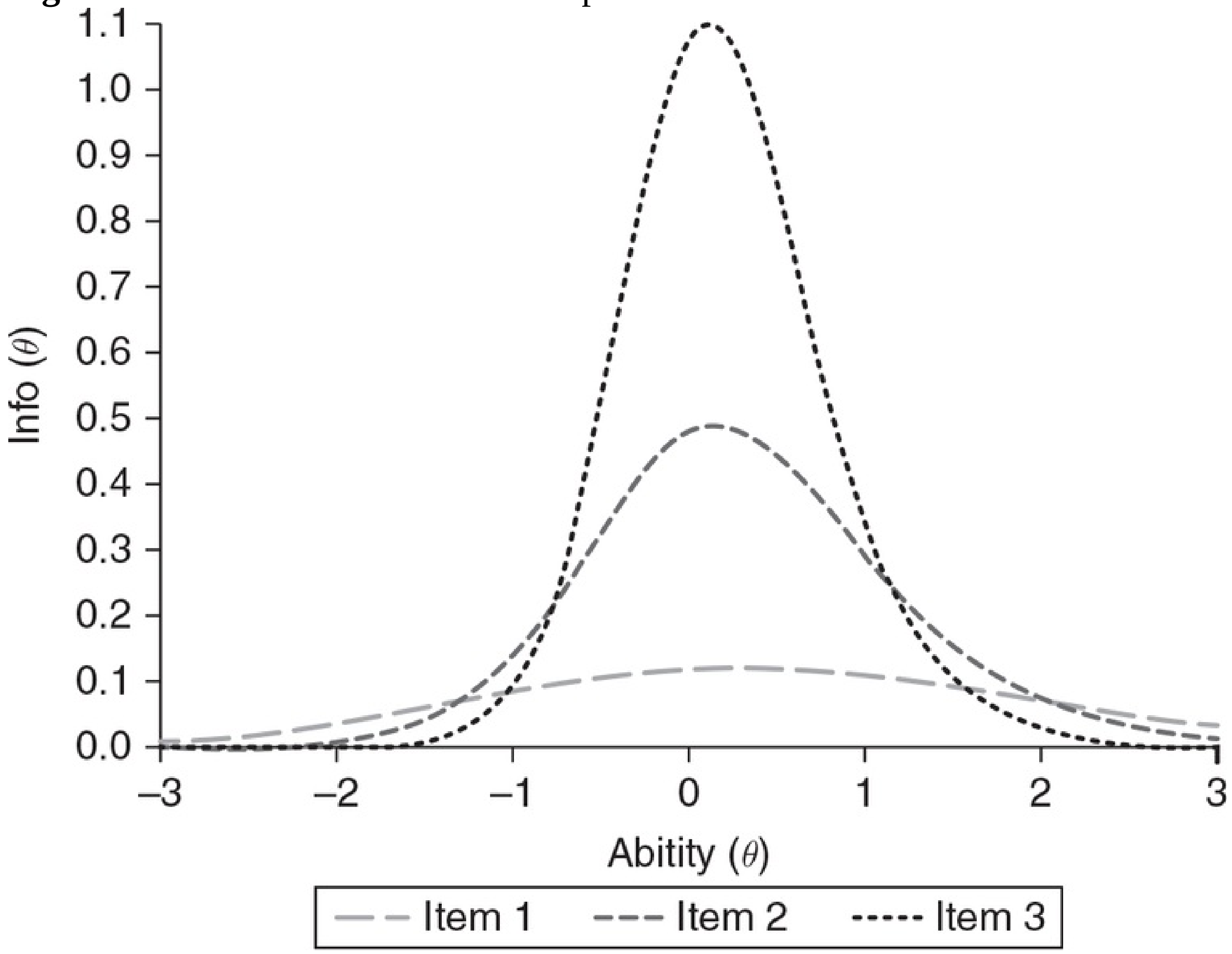
Figure 2 ICCs for three example items



As shown in Figure 2, the slope of Item 1, at the location where examinees' abilities are about the same as the item's difficulty ($\theta = b = 0$), is the flattest among the three items, whereas the slope of Item 3 at the same location is the steepest. Therefore, Item 1 has the lowest discrimination value among the three and Item 3 has the highest. If identifying two examinees of whom one has an ability larger than zero and one smaller than zero on the ability axis, the difference between the probabilities of the two students answering Item 3 correctly will be greater than Item 1 or 2. It is therefore easier to discriminate between the two examinees using Item 3, compared to Item 1 or 2. With all else equal, Item 3 can be concluded as more desirable because it can effectively distinguish among examinees differing in ability.

Using the item information function, similar information can be verified via a different angle. Generally speaking, information stands for precision. If the amount of information is large at a given ability level, an examinee whose true ability at that level can be estimated with the greatest precision. Figure 3 shows the item information functions for the 3 items previously shown in Figure 2. As shown in Figure 3, when the *b* parameter and the *c* parameter are the same across items as in the example, Item 3 has the largest item information because the *a* parameter associated with Item 3 has the largest value among the 3 items.

Figure 3 IIFs for the same three example items



The *a* parameter interpretation in the 2PL model is very similar to its interpretation in the 3PL model. However, because the location parameter for the 2PL model indicates the ability level at which examinees have a 50% chance of answering an item correctly, the *a* parameter indicates the item discrimination information specifically at this ability level. The item discriminations are considered equal across items in the 1PL model because all *a* parameters are fixed. Therefore, if a group of items' ICCs are presented together under the 1PL model, the curves will not cross each other. Overall, items can be maximally informative at any part of the ability (θ) continuum, but interpretation of the *a* parameter will be most meaningful when interpreted in conjunction with the *b* and/or *c* parameter for the same item.

The item responses are said to be polytomous when more than two categories exist. Representative models in this group include graded response model, generalized partial credit model, and nominal response model. For polytomous item responses, the probability of an examinee reaching a score category can be described by one of these polytomous IRT models. Interpretation of the discrimination parameter in polytomous response models is very similar to its counterparts in the binary response models arena because polytomous item response models are developed by extending the general underlying IRT premise to items scored in two or more categories.

Fei Zhao

See also *b* Parameter; *c* Parameter; Item Response Theory

Further Readings

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques. New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397–479). Reading, MA: MIT Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), 29–51.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159–176.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4 part 2), 100.

David Morgan David Morgan Morgan, David

ABA Designs

ABA designs

3

6

ABA Designs

ABA designs, also known as *reversal designs,* are among a family of single-case experimental designs most often used by behavioral scientists and educators to evaluate the effectiveness of clinical or educational interventions. This entry first describes ABA designs and provides an example, then discusses phase changes in ABA designs, how ABA designs are used to identify treatment effects, and the limitations of ABA designs.

In a typical ABA design, a relevant dependent variable, such as frequency of tantrums, self-injurious behaviors, or correct academic responses, is measured continuously over some period of time for a single participant. Observation and measurement of this behavior initially occurs under a *baseline* condition (A in the ABA sequence), in which no independent variable, or treatment, is presented. During this baseline condition, the behavior of interest is assumed to be occurring at its natural level, prior to introduction of the independent variable. After this behavior has demonstrated stability, showing no discernible upward or downward trend during baseline, the treatment or intervention phase (B) is introduced, and measurement of the dependent variable continues. Finally, a return to baseline (A) is programmed, allowing for assessment of the dependent variable once again in the absence of treatment.

The ABA design can be seen as a formalization of the common “before and after” observations that many of us make of ourselves in response to changes in diet, exercise routines, and other efforts at self-improvement. Such designs are common experimental methods in the natural sciences and were advocated by Claude Bernard, the father of experimental medicine. The general logic of single case experimentation was adopted by psychologist B. F. Skinner as a powerful case experimentation was adopted by psychologist B. F. Skinner as a powerful method for the continuous analysis of learning processes in real time. Beginning in the late 1930s, Skinner pioneered a branch of natural science he called the experimental analysis of behavior, whose products include the contemporary profession of applied behavior analysis.

Applied behavior analysts have been longtime proponents and practitioners of ABA designs, as their work involves the application of basic learning principles to socially significant behavior across many domains, including schools, workplaces, and the home. Behavior analysts utilize single-case ABA designs in order to evaluate the effectiveness of treatments, especially in the area of autism and developmental disabilities. ABA designs are experimental designs that support causal inferences, and the data produced by such designs contribute to our knowledge of evidence-based interventions in the behavioral sciences.

A hallmark of the ABA design is its focus on the behavior of the individual. Behavior analysts consider single-case designs and continuous measurement superior to large-scale group designs in resolving the nuances of moment-to-moment behavior–environment interactions that are often the target of such interventions. Behavior often responds both quickly and dramatically to changes in environmental variables, and the ABA design is a powerful method for assessing these changes. In fact, use of ABA designs can often identify environmental events that correspond to changes in behavior with as much regularity as the lights turn off and on in a room with each flip of a light switch. Each change in condition, from baseline to treatment, or from treatment to baseline, becomes an opportunity to observe a functional relationship between the treatment and the behavioral variable. In addition, further phase changes can be programmed to replicate this functional relationship; thus, ABA designs have intraparticipant replication built into them, and this is an extremely important criterion for developing evidence-based practices.

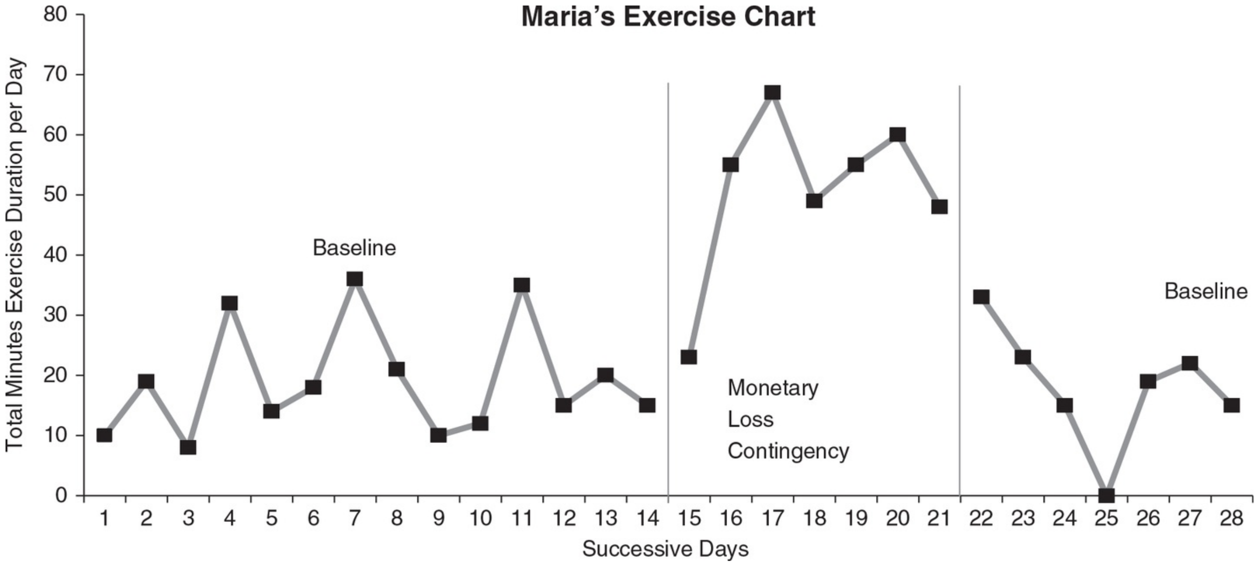
Example of ABA Design

The essential logic of the ABA designs is fairly intuitive and can be depicted in a hypothetical example. Maria, a successful CEO of a marketing firm, having learned that she is at risk for developing heart disease due to family history, has decided to initiate a regular exercise regimen. She joins a local gym and, being an observant and detail-oriented person, keeps track of her total time spent exercising for 2 weeks. Maria’s interest is in increasing her overall duration of cardiovascular exercise, not any specific kind of workout, so she varies her cardiovascular exercise, not any specific kind of workout, so she varies her exercise routines (e.g., swimming, riding a stationary bike, walking on a treadmill) and times of her workouts, adding up the total time for each week.

At the end of a 2-week period, Maria sees that her exercise duration varies from day to day, but it doesn’t seem impressive to her, and she would like to set her goals much higher. Like most of us, Maria lives a very busy life, raising two children and putting in the long hours of an executive officer, and realizes that increasing her exercise amount is going to be a challenge. In order to provide additional motivation, Maria recruits the help of a close friend. She writes her friend a check for a very large amount, more than she would feel comfortable losing, and tells her friend that if she (Maria) doesn’t increase her total exercise duration by at least 20% by the end of the next week, her friend is to give the money to an organization for which Maria has nothing but contempt. In essence, Maria is using a potentially aversive consequence, losing a sizable amount of money, in order to motivate her to increase her exercise. Although self-imposed, she is manipulating an independent variable (threat of lost money) in order to alter measurable aspects of a dependent variable, in this case her total amount of weekly exercise.

Figure 1 is a time-series graph, often used by behavioral scientists to depict behavior change in response to programmed treatments or interventions. The graph is called a time-series graph because time, in this case represented in successive days, is represented on the *x*-axis. The primary dependent variable, total duration of exercise per week, is plotted on the *y*-axis. The vertical lines drawn through the data paths represent changes in conditions, the first line depicting initial implementation of the monetary contingency (potential loss of money) and the second line depicting removal of this monetary contingency or return to baseline.

Figure 1 Maria’s exercise chart



The first week's data represent Maria's baseline level of exercise. Although Maria's exercise duration shows some degree of variability from day to day (a near guarantee for almost any behavior), there are no obvious trends upward or downward during this first week. Maria's disappointment in her exercise amount prompts her to enlist her friend and a simple behavioral contingency in an effort to enhance her exercise. The first vertical line, then, separates the first two weeks, or baseline period, from the initial treatment week, in which the potential monetary loss is in effect. Finally, after the intervention week, Maria decides to revert to the baseline condition, in which the potential monetary loss contingency is no longer in effect. This return to baseline is conceptualized as a replication of this nontreatment condition, and it serves as a comparison for the previous treatment condition.

Phase Changes

In single-case experimental designs, individual participants serve as their own controls, meaning their behavior is evaluated under both treatment and nontreatment conditions. This comparison is logically similar to comparing control and experimental groups in a more conventional large group experimental design. In the ABA design, visible change in the data path following a phase change (baseline to treatment or treatment to baseline) is suggestive of an effect of independent variable presentation (treatment) or removal (return to baseline).

The logic of the design actually allows for multiple alternations between baseline and treatment, so that although the design is called ABA, the baseline and treatment, so that although the design is called ABA, there are no formal limits on how many phase changes can be produced. Two presentations and subsequent removals of treatment, for example, would result in an ABABA design. Although this many phase changes may not be common, multiple phase changes are advantageous because each phase change represents a replication of an earlier condition change. Single-case experimental designs derive their inferential power from such replications, not through null hypothesis testing and statistical inference.

Identifying Treatment Effects

The hypothetical real-time data provided by Maria in Figure 1 would be subjected to visual analysis and possibly quantitative effect size measures in order to identify treatment effects. When analyzing such time-series data, researchers consider a number of characteristics of the data path. One possible comparison is to draw a horizontal line through the mean of a baseline phase and an adjacent treatment phase. The difference between these lines is interpreted as a change in level. In addition, a data path can be evaluated for trend, which is visible as clear movement up or down in the data path.

As stated earlier, it is important to ensure that no obvious trend is present in baseline data, as this would render any comparison of these data with initial treatment data problematic. For instance, if Maria’s exercise duration demonstrated a clear increase near the end of the first phase (A) in Figure 1, additional increases during the first intervention (B) could not be easily interpreted as a treatment effect. However, if baseline data remain stable and increase only in treatment phases, a stronger argument can be made for a treatment effect. In addition to visual inspection of data paths, a number of quantitative analyses have been developed for evaluating effect sizes in single-case experimental designs.

In the example presented earlier, Maria transitioned from a baseline condition to a treatment condition and finished with a return to baseline. The basic logic of the ABA design, however, allows for changing this sequence if necessary. In many instances, the behavior being targeted may be especially aversive, even dangerous, for the client or others in the client’s environment. When this is the case, and a substantial baseline period of observation is considered unethical (recall that baseline means that there is no treatment being delivered), a treatment condition can actually be instituted first, followed by a brief baseline period and a final treatment period. This effectively produces a BAB design rather than an ABA design.

Limitations of ABA Designs

ABA or reversal designs are powerful and flexible for identifying treatment effects in education and behavior analysis, but they are not without limitations. Many behaviors, especially those acquired in academic settings, are not easily reversed when interventions are removed. Imagine, for example, delivering a new reading readiness program to preschoolers. One would not expect the skills acquired during this program to disappear during a subsequent return to baseline. And, as mentioned previously, when the behavior of interest is potentially dangerous, it would be unethical to withdraw what appeared to be an effective treatment. For this reason, frequent reversals to baseline are not always feasible in applied settings. When this is the case, ABA designs are not advisable. Other single-case experimental designs, however, such as multiple-baseline designs, changing criterion designs, and alternating treatment designs, can be relied on to assess treatment effectiveness. These designs also utilize the principal tactic of replication to demonstrate treatment effects but do so in procedurally different ways and both within and across individual participants.

David Morgan

See also Applied Behavior Analysis; Behaviorism; Replication; Single-Case Research

Further Readings

Bailey, J. S., & Burch, M. R. (2002). Research methods in applied behavior analysis. Thousand Oaks, CA: Sage.

Blampied, N. M. (2012). Single-case research designs and the scientist-practitioner ideal in applied psychology. In G. Madden (Ed.), APA handbook of behavior analysis. Washington, DC: American Psychological Association.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Applied behavior analysis (2nd ed.). Upper Saddle River, NJ: Pearson.

Morgan, D. L., & Morgan, R. K. (2009). Single-case research methods for the behavioral and health sciences. Thousand Oaks, CA: Sage.

Richards, S. B., Taylor, R. L., & Ramasamy, R. (2014). Single subject research: Applications in educational and clinical settings. Belmont, CA: Wadsworth.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. Psychological Methods, 17(4), 510–550.

Mark S. Teachout Mark S. Teachout Teachout, Mark S.

Malcolm James Ree Malcolm James Ree Ree, Malcolm James

Thomas R. Carretta Thomas R. Carretta Carretta, Thomas R.

Ability Tests

Ability tests

6

10

Ability Tests

An ability test is an objective and standardized measure of a sample of behaviors at a specific point in time. Broadly defined, ability tests assess the innate and acquired capacity to perform mental or motor functions. This entry discusses the history of ability tests, their classification and standardization, the criteria for evaluating ability tests, and controversies over their use.

The modern scientific study of human cognitive abilities is often attributed to French psychologist Alfred Binet, who developed the Binet-Simon intelligence test, and to the World War I Army Alpha and Beta tests in the United States. The Army Alpha and Beta tests were used to assess cognitive ability for U.S. military recruits during World War I. The Army Beta test is noteworthy in that it was used to evaluate recruits who were illiterate, unschooled, or non-English speaking. This is considered an early example of cognitive tests that do not rely on verbal skills or learned content. An important consequence of the Army Alpha and Beta tests was the popularization of group-administered aptitude tests.

Debate on the structure of abilities began in the early 1900s and continues to the present. An important milestone in this debate was the development and application of factor analysis to determine the extent to which aptitudes were distinct from one another. Applying factor analyses, Louis Leon Thurstone proposed seven primary mental abilities. Others later reanalyzed Thurstone’s data and demonstrated a single, general ability factor that influenced the seven primary abilities. Despite this finding, the development of multiple aptitude tests thrived. Examples of these tests are the Differential Aptitude Tests, Multiple Aptitude Tests, and Comprehensive Ability Battery.

In the last part of the 20th century, the advent of fast and inexpensive computers played a major role in both test construction and test administration. Previously, test statistics were computationally burdensome, prone to errors, and time-consuming. Computers enabled test construction to be done more quickly with fewer errors while also making results available much sooner. It also has allowed for computer adaptive testing. In computer adaptive testing, a question is asked, the response is scored immediately, and the next item is selected to best suit the ability level of the examinee. This continues until an accurate measure of ability is obtained.

Computer adaptive testing has been implemented by governments and commercial endeavors. Three examples are the National Council Licensure Examination of the National Council of State Boards of Nursing, the Armed Services Vocational Aptitude Battery, and the GRE General Test. In 2011, the GRE General Test became adaptive only for groups of questions.

Classification of Ability Tests

A general distinction is often made between ability tests, sometimes referred to as intelligence or aptitude tests, and achievement tests. Although ability, intelligence, and aptitude are sometimes used synonymously, there are subtle distinctions. Ability and intelligence tests are usually considered as tapping more into fundamental abilities, while aptitude tests may include more of an accumulation of cognitive and motor abilities. In addition, intelligence tests are often defined in broad categories such as verbal and quantitative abilities, whereas aptitude tests are usually defined in more specific ways combining ability and accumulated knowledge (e.g., mechanical, musical, and spatial).

The following are two important factors that differentiate aptitude from achievement tests: (1) the prior experience of the examinee that is considered by the test developer and (2) the purpose for which the test scores are used. Early conceptions of these tests reflected a simplistic distinction based on heredity versus the environment. Aptitude tests were based on innate capacity or traits, independent of learning, while achievement tests were based on more specific learning.

Rather than considering ability and achievement as independent concepts, a useful approach is to consider ability tests on a continuum. All are developed abilities, differentiated by the type of prior experience that is considered in constructing the test items. At one end of the continuum, aptitude is acquired over years of education, experience, and life activities. Therefore, prior experience of the examinee is defined quite broadly and over a longer term. Some consider aptitude tests long-term achievement tests. At the other end of the continuum, achievement tests measure specialized knowledge or skills acquired through formal or informal education or training. Hence, prior experience is typically considered narrower and shorter term.

The second distinction is the purpose for which the test scores are used. While both are a snapshot of an attribute at a specific time, aptitude tests are designed for predictive purposes while achievement tests are designed for descriptive purposes. A typical aptitude or intelligence test is designed to assess the capacity of the examinee to learn both cognitive and motor skills in order to predict the potential to learn and use those skills in future situations, such as an educational, training, or work setting. An achievement test may examine mastery of knowledge or a motor skill to determine the level of competency of the examinee. For example, knowledge of regulations and driving ability are tested to obtain a driver's license, a language test is administered to select an interpreter, and state licensing exams are used for many professions such as the medical and legal professions.

Both aptitude and achievement are developed abilities. Aptitude tests describe knowledge and skills and measure attributes intended to predict future learning. Achievement tests measure the mastery of more specific subject matter. In practice, there may be confusion when there is overlap between the content and purpose of a test. That is, some tests may contain elements of both aptitude and achievement. When learned contents are used in aptitude tests, the blend becomes evident. Achievement tests are sometimes misused for predictive purposes.

Aptitude Tests

Some aptitude tests purport to measure a single aptitude, while others purport to measure multiple aptitudes. Typically, most aptitude tests measure a relatively standard set of constructs reflecting cognitive and motor skills. Cognitive skills are unobservable and represent different capacities for mental activity, information processing, understanding, and problem solving. Content may include verbal, numerical, spatial, abstract reasoning, and comprehension. Motor skills are observable and represent the ability to perform physical tasks that do not require a cognitive skill to understand. These are motor coordination, finger dexterity, and manual dexterity.

Some motor skill tests require the use of cognitive skills to make a physical response. For example, a block design test requires the examinee to view a picture of how blocks should look when assembled and then to assemble them as quickly as possible replicating the design in the picture. Some aptitude tests use composite scores obtained by combining two or more subtests. Scores can be interpreted as a unique test score and as part of a composite. For example, intelligence test scores are verbal intelligence (using verbal, numerical, and spatial aptitude) and performance intelligence (using abstract reasoning and object manipulation tasks). Other aptitude composites are verbal comprehension, perceptual organization, processing speed, and working memory.

Examples of aptitude tests are the Stanford–Binet Intelligence Scales, Wechsler Adult Intelligence Scale, Wechsler Intelligence Scale for Children, Wechsler Preschool and Primary Scale of Intelligence, Otis-Lennon School Ability Test, Differential Ability Scales, and the Woodcock-Johnson Tests of Cognitive Abilities.

There are ways of measuring aptitude without learned content such as the Raven’s Progressive Matrices and a procedure measuring the speed of neural processing. Raven’s Progressive Matrices is a test of abstract reasoning based on a series of geometric figures and is frequently regarded as “culture free.” An individual’s speed of neural processing can be assessed by having the individual look at a computer screen while a light is flashed and the speed with which a single nerve conducts the impulse is measured.

Admissions tests are used in the application process at private elementary and secondary schools, as well as at most colleges and universities. These are used to predict the probability of student success in these academic settings. Examples for secondary school include the High School Placement Test. Tests for undergraduate admission are the SAT and ACT. In addition, there are numerous admission exams for graduate and professional school, including the Graduate Management Admission Test (GMAT) used by business schools, GRE General Test, Law School Admission Test, and Pharmacy College Admission Test

Test, Law School Admission Test, and Pharmacy College Admission Test.

Achievement Tests

Tests of specific knowledge, professional certification, and licensing are not aptitude tests but rather achievement tests. Achievement tests measure specialized knowledge or skills acquired through formal or informal education or training. Thousands of achievement tests are developed and used nationally by states and local entities. Examples are the Wechsler Individual Achievement Test, Kaufman Test of Educational Achievement, Woodcock-Johnson Tests of Achievement, and Peabody Individual Achievement Test. Many states use specifically designed achievement tests for certification of teachers and principals as well as for medical and legal professions.

For public schools, the National Assessment of Educational Progress is used, whereas state achievement tests may be required for schools that receive federal funding. There also may be tests required for high school graduation such as the New York State Regents Examination. Other achievement tests include the GED test, which is taken to certify academic skills in lieu of a high school diploma. Tests created by private institutions are often used to monitor progress in K–12 classrooms.

Praxis certification exams for teacher certification measure academic skills in reading, writing, and mathematics. The Praxis Subject Assessments measure subject-specific content knowledge and the Praxis Content Knowledge for Teaching Assessments assess specialized content knowledge for K–12 teaching.

Finally, language proficiency exams such as the Test of English as a Foreign Language (TOEFL) are used to assess international students for admission to colleges and universities where English is used as the primary language.

Standardization of Ability Tests

Most ability tests are standardized and designed to provide an objective assessment of an individual’s abilities relative to data collected on a relevant reference group (i.e., normative group). This is known as a norm-referenced test. Comparing the examinee’s score to that of the normative group permits the interpretation of the score relative to the normed population, whether a raw score, standard score, or percentile. The normed group must be meaningful as a basis for comparison. For example, it would be misleading to compare high school student test scores to scores from a normative group of elementary school students because the high school student results would appear to be much higher than if compared to a more relevant norm group of their own age.

Tests often have more than one normative group. For example, a test designed for elementary school children may have a normative group of students for each grade. Subsequent examinee scores can then be compared to see whether they've scored at, above, or below their grade level. Norms are also often developed for gender and ethnicity.

Developing normative information is time-consuming and costly. Often an appropriate normative group is not available for comparison, so users must choose the most relevant norm group available. Another issue is that normative information may be out-of-date. The accuracy and usefulness of normative interpretations may decline as the age of the normative data increases.

Achievement tests may also be standardized, but unlike ability tests where test scores are compared to a normative population (e.g., high school graduates), achievement test scores are usually compared to a minimum acceptable score. Tests that are interpreted by comparing the examinee's score to a predetermined standard or cutoff score are called criterion-referenced tests. For example, to obtain a driver's license in most states, one must receive a minimum score on a test of traffic and safety rules.

Criteria for Evaluating Ability Tests

Three points of evaluation of ability tests are reliability, validity, and applicability for the examinees. Reliability creates consistency of measurement. Although there are several methods for determining reliability, the most obvious is test–retest, repeating the same test on two occasions with the same people and calculating the correlation between the scores.

The validity of a test is concerned with what the test measures and how well it measures it. There are several aspects of validity that are particularly relevant to ability tests, including predictive validity, content-based validity, and construct-based validity. Predictive validity determines whether the test predicts some important outcome such as success in a school course. Content-based validity arguments focus on the items or tasks that make up the test itself and the degree to which they appropriately sample from the universe of all possible items. Construct-based validity argues that the test scores represent the construct of interest within a theoretical framework.

Evaluation of reliability and validity depends on the use of the test scores. Rough rules of thumb for reliability are based on coefficient α, a common index of internal consistency: .5 is acceptable for research, .7 for group decisions, and .8 for decisions about individuals. Validity can be evaluated in several ways, but often the strongest evidence comes from correlations with other measures. Does the ability test show a statistically significant correlation with an external criterion test? Evaluation of reliability and validity is complex, however, and is not driven by one or two criteria.

Controversies in Ability Testing

Controversies in ability testing have a long history. Among the longest controversies is whether ability is unitary or made up of many separate parts (multiple abilities), and the way theories organize these parts. The controversy is sometimes cast in the model of general ability (*g*) versus specific abilities (*s*). This has implications for construction and use of ability measures. Early 20th century models of ability emphasized *g* but did not exclude the concept of *s*. Mid-century models focused on *s* which is reflected in the development of multiple abilities theories. This lead to the development of multiple ability test batteries that found their way into public schools as early as1935. The existence of a general ability factor is widely accepted, but the utility of *g* versus *s* is still debated.

Other controversies include the differences in mean test scores among groups and the impact of race, ethnicity, gender, and culture on ability test scores. These controversies have been widely studied and discussed in the professional literature. While group mean differences exist, there is no evidence that the predictive power of tests differs among and between groups. Additionally, there is empirical evidence that ability and achievement tests measure the same constructs for the various identifiable groups in the population.

More recently, the use of so-called high-stakes testing has become controversial, especially with the nationwide implementation of the No Child Left Behind Act of 2001, which was replaced in 2015 by the Every Student Succeeds Act. It is of note that New York state instituted high-stakes testing in secondary school with the Regents Examinations in the 19th century and that even before the No Child Left Behind Act of 2001 some states had school accountability systems that included public reporting of standardized test scores and sanctions for low performance.

Ability test misuse is frequently related to four deficiencies. The first is poor training and practice by the examiner. Some examiners will have been trained in past decades by experts whose training was completed decades before that. Therefore, they will not be knowledgeable about changes in scores and their meanings. Second, proper interpretation of scores depends on the knowledge of the situation in which the constructs measured will be relevant. Third is failure to evaluate the construct validity of the test. The claim of the test developer is not a sufficient substitute for examining the articles and reports that support the claims that an ability test actually measures the targeted ability. Finally, continuing education on changes in statistical methods in testing needs to take place.

Mark S. Teachout, Malcolm James Ree, and Thomas R. Carretta

See also Achievement Tests; Aptitude Tests; Computerized Adaptive Testing; Intelligence Tests; Reliability; Stanford–Binet Intelligence Scales; Validity

Further Readings

Anastasi, A., & Urbina, S. (1997). Essentials of psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Dubois, P. (1970). A history of psychological testing. Boston, MA: Allyn and Bacon.

Fleishman, E. A., & Quaintance, M. K. (1984). Taxonomies of human performance: The description of human tasks. Orlando, FL: Academic Press.

Jensen, A. R. (1980). Bias in mental testing. New York, NY: The Free Press.

Jensen, A. R. (1998). The g factor: The science of mental ability. Westport, CN: Praeger.

Zhang, H. (1988). Psychological measurement in China. International Journal of Psychology, 23, 101–117.

Frank R. Vellutino Frank R. Vellutino Vellutino, Frank R.

Ability–Achievement Discrepancy Ability–achievement discrepancy

10

14

Ability–Achievement Discrepancy

The ability–achievement discrepancy is defined as a statistically significant difference between a child's score on a measure of achievement in one or another academic domain such as reading or math and the child's score on a measure of intellectual ability, typically in the form of IQ. For a considerable period of time, the IQ-achievement discrepancy was the central criterion used by educators, school psychologists, and educational researchers to define specific learning disabilities in otherwise normal children. This entry discusses the history of the ability–achievement discrepancy, the origin of its use, and problems with its use to identify individuals with specific reading disability.

The use of the ability–achievement discrepancy criterion has a long history that dates back to Samuel Kirk and Barbara Bateman's suggestion that learning disabilities can be defined as a collection of developmental disorders of neurological origin that affect various types of school-based learning in children who are not mentally challenged or impaired by extraneous impediments to learning such as sensory deficits, emotional disorders, or socioeconomic disadvantage. The Education for All Handicapped Children Act (U.S. Public Law 94–142, later renamed the Individuals with Disabilities Education Act), adopted in 1975, mandated that learning disabilities be defined as the occurrence of achievement deficits in otherwise normal children who have at least average intelligence.

The Education for All Handicapped Children Act had significant impact because it led to the use of intelligence as a defining criterion in most state definitions of learning disability, typically in the form of an IQ-achievement discrepancy in one or another academic domain. The IQ-achievement discrepancy eventually became widely adopted as a basic prerequisite for diagnosing learning disabilities in schools and other institutions and for defining learning disabilities

disabilities in schools and other institutions and for defining learning disabilities in empirical research studying the etiology and nature of hypothesized impediments to success in school learning.

Origin of the IQ-Achievement Discrepancy Definition of Learning Disability

The definition of learning disability specified in U.S. Public Law 94–142 and the widespread use of the IQ-achievement discrepancy as the central criterion for diagnosing learning disabilities were in large measure influenced by the work of Michael Rutter, William Yule, and their associates. These investigators conducted a large-scale epidemiological study evaluating the etiology of reading disability—the most common form of learning disability—and found that the percentage of children whose scores on measures of reading ability were two standard errors or more below the scores that were predicted by their ages and IQs was significantly higher than the percentage of children who were expected to fall in this range on the assumption of normality (i.e., 2.3%), thereby creating a “hump” in the tail end of the distribution of residual scores.

Rutter and Yule hypothesized that there were two types of impaired readers: one said to be afflicted by “specific reading retardation,” as defined by a significant discrepancy between observed reading achievement and expected or IQ-based reading achievement, along with the absence of general learning difficulties, and a second said to be afflicted by “general reading backwardness,” as defined by general learning difficulties along with the absence of any significant discrepancy between observed and expected reading achievement. Rutter and Yule’s distinction between specific reading retardation and general reading backwardness was in keeping with Kirk and Bateman’s seminal definition of learning disability, and it subsequently became the basis for what can be called “exclusionary definitions” of reading and other learning disabilities having the IQ-achievement discrepancy as their central defining criterion.

Problems With Discrepancy Definitions of Reading Disability

The use of the IQ-achievement discrepancy as the basis for defining specific reading disability, also referred to as dyslexia, has qualified empirical justification at best. For example, Eve Malmquist reviewed results from a large number of studies that evaluated the relationship between intelligence and reading achievement and found that correlations between these two variables were modest, ranging only from .40 to .60. Malmquist obtained correlations of comparable magnitudes in a large multivariate study of first-grade, poor, and normal readers. Guy Bond and Robert Dykstra obtained similar results with a randomly selected sample of first graders.

Although several studies have obtained higher correlations between intelligence and reading achievement with older participants, the intelligence tests used in these studies consisted of items with high verbal content and/or depended on skill in reading. Thus, observed correlations between the measures of intelligence and the measures of reading achievement used in these studies, in many cases, may have been an artifact of shared variance contributed by language-based abilities underlying performance on both sets of measures. Additional support for this possibility is provided by a study conducted by Frank Vellutino and his associates with elementary and middle school-age children in which it was found that correlations between measures of reading subskills and a commonly used test of intelligence were significantly higher when the intelligence test evaluated verbal abilities than when it evaluated nonverbal abilities. Thus, discrepancy definitions of the types motivated by Rutter and Yule’s work may have been based on inaccurate conceptualization of the relationship between intelligence and reading achievement resulting from faulty analysis of the cognitive abilities underlying performance on both intelligence tests and reading tests.

Even more compelling evidence against using the IQ-achievement discrepancy to define reading and other learning disabilities is provided by results from several other studies that have addressed the question. First, a study conducted by Rodgers failed to replicate the type of bimodal distributions obtained by Rutter and Yule. This finding was replicated by David Share and his associates, and it was suggested that Rutter and Yule’s findings may have been an artifact of floor and ceiling effects on the reading measures used in their studies. This possibility was later given some credibility in independent studies conducted by A. van der Wissel and F. E. Zegers and by Share and his associates, in which it was found that bimodality in each of several distributions of IQ-reading residual scores could be artificially induced by creating false ceilings on the reading scores. These findings question the reliability of the results obtained by Rutter and Yule and, thereby, question the validity of using both IQ scores to estimate expected reading achievement and the validity of using an IQ-achievement discrepancy to define reading and other learning disabilities. More definitive evidence against these practices comes from several other studies that have appeared in the literature.

For example, in a study conducted by Jack Fletcher and his associates, it was found that impaired readers who had no significant IQ-achievement discrepancies performed no differently than impaired readers who did have significant IQ-achievement discrepancies, either on measures of reading achievement or on measures of cognitive abilities believed to underlie reading achievement (e.g., phonological awareness, verbal memory, word retrieval, and visual analysis). Not surprisingly, each of these groups performed significantly below a group of nondiscrepant typically developing readers on both sets of measures. However, of special interest is Fletcher and colleagues' finding that a group of children who would have been classified as "disabled readers" by virtue of significant IQ-achievement discrepancies (i.e., above average IQs coupled with at least average reading achievement) performed as well as nondiscrepant typical readers not only on tests of reading achievement but also on tests of reading-related cognitive abilities. Fletcher and colleagues were doubtful that the children in the former group had a reading disability and suggested that the IQ-achievement discrepancy risks either overidentifying or underidentifying children as disabled.

Results discussed thus far quite naturally raise two important questions: (1) To what degree can an individual's IQ set upper limits on or predict the individual's ability to learn to read? (2) To what degree can an individual's IQ predict response to remedial intervention? The first of these questions was addressed in independent studies conducted by Linda Siegel and by Share and his associates. In the study conducted by Siegel, children with and without reading disability across a broad age range (7–16 years) were administered a large battery of tests evaluating reading achievement and reading-related language and language-based skills, in addition to measures of verbal and nonverbal intelligence. The children in both groups were then stratified in one of four IQ subgroups (IQ < 80, IQ = 80–90, IQ = 91–109, and IQ > 110) and thereafter compared on the reading and cognitive measures.

Siegel found that within each IQ stratification the children with reading disability performed significantly below the children without reading disability on all cognitive measures. This finding is important because it indicates that readers with and without reading disability can be found within different IQ ranges, including those falling below the average range. This is contrary to the view that intelligence is highly correlated with reading ability and therefore sets upper limits on reading achievement.

Share and his associates later conducted a longitudinal study that replicated Siegel’s findings and also addressed the question of whether IQ can predict rate of growth in reading. The investigators tracked an unselected group of aged children from 3–13 years and periodically evaluated their reading achievement at ages 7, 9, 11, and 13. Intelligence in these children was assessed at ages 3 and 5. A composite measure based on these estimates was used to group the children into six IQ ranges and the children in each range were assessed at age 13 years on a measure of word recognition. In accord with results obtained by Siegel, Share found that IQ and reading ability were not highly correlated insofar as the full range of reading ability was represented within each IQ range. In addition, no strong or consistent differences were found among the different groups in rate of growth in reading.

The question of whether intelligence test scores can predict response to remedial intervention was initially addressed in a large-scale intervention study conducted by Vellutino and his associates that was published in 1996. In this study, reading growth in children identified as struggling readers in mid-first grade was tracked from the beginning of kindergarten until the end of fourth grade—that is, before and after they were identified as struggling readers. A randomly selected group of these children were provided with daily individual tutoring and the rest were provided with whatever remedial services were available at their home schools. Intervention was initiated in mid-first grade and was terminated at the end of first grade for children who were found to be readily remediated and in mid-second grade for children who were found to be more difficult to remediate. After one semester of project-based intervention, children who received this intervention were rank ordered on the basis of measures of growth in reading during that semester and thereafter separated into four groups designated as follows: “very good growth,” “good growth,” “limited growth,” and “very limited growth.”

For purposes of comparison, two groups of typically developing readers were also identified in mid-first grade: one group consisting of children with average intelligence and a second group consisting of children with above average intelligence. Reading growth in these children was also tracked from the beginning of kindergarten through the end of fourth grade. In addition, all groups were compared on measures of intelligence and reading related cognitive abilities in kindergarten, first, and third grade. Vellutino and colleagues found abilities in kindergarten, first, and third grade. Vellutino and colleagues found that the children in the four tutored groups did not differ on the measures of intelligence, especially those evaluating nonverbal intelligence, nor did they differ from the typical readers with average intelligence on these measures. In contrast, the children who were found to be difficult to remediate differed significantly from the children who were found to be readily remediated on measures of language-based skills, especially phonological skills that are important for learning to read (e.g., knowledge of letter names and sounds, phoneme awareness, letter–sound decoding, verbal memory, and name retrieval). In addition, the typical readers with average and above average intelligence did not differ on measures of basic word-level skills (i.e., word identification and word attack).

These findings were essentially replicated in a second major intervention study conducted by Vellutino and his associates, and the combined results from these two studies provide strong and compelling evidence that responsiveness to intervention may be a more valid means of identifying specific learning disability as compared with the IQ-achievement discrepancy. Research conducted more recently provides considerable support for this suggestion.

Frank R. Vellutino

See also Ability Tests; Achievement Tests; Aptitude Tests; Evidence-Based Interventions; Flynn Effect; Intelligence Quotient; Intelligence Tests; Learning Disabilities; Stanford–Binet Intelligence Scales; Wechsler Intelligence Scales; Woodcock-Johnson Tests of Achievement

Further Readings

Bond, G. L., & Dykstra, R. (1967). The co-operative research program in first grade reading instruction. Reading Research Quarterly, 2, 5–142.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Steubing, K. K., & Shaywitz, B. A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. Journal of Educational Psychology, 86(1), 6–23.

Kirk, S. A., & Bateman, B. (1962–1963). Diagnosis and remediation of learning disabilities. Exceptional Children, 29, 73–78.

Malmquist, E. (1960). Factors related to reading disabilities in the first grade of the elementary school. Stockholm, Sweden: Almqvist Wiksell.

Rodgers, B. (1983). The identification and prevalence of specific reading retardation. British Journal of Educational Psychology, 53, 369–373.

Rutter, M., & Yule, W. (1975). The concept of specific reading retardation. Journal of Child Psychology and Psychiatry, 16, 181–197.

Share, D. L., McGee, R., McKenzie, D., Williams, S., & Silva, P. A. (1987). Further evidence relating to the distinction between specific reading retardation and general reading backwardness. British Journal of Developmental Psychology, 5, 35–44.

Siegel, L. S. (1988). Evidence that IQ scores are irrelevant to the definition and analysis of reading disability. Canadian Journal of Psychology, 42, 201–215.

U. S. Office of Education. (1977). Assistance for states for education for handicapped children: Procedures for evaluating specific learning disabilities. Federal Register, 42, G1082–G1085.

van der Wissel, A., & Zegers, F. E. (1985). Reading retardation revisited. British Journal of Developmental Psychology, 3, 3–9.

Vellutino, F., Scanlon, D., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-achievement discrepancy definition of reading disability. Journal of Learning Disabilities, 33, 223–238.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R.,

& Denckla, M. B. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Toward distinguishing between constitutionally and experientially based causes of reading disability. Journal of Educational Psychology, 88(4), 601–638.

Vellutino, F. R, Scanlon, D. M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify child at risk for long-term reading difficulties. Reading and Writing: An Interdisciplinary Journal, 21, 437–480.

Mark Pedretti Mark Pedretti Pedretti, Mark

Abstracts

Abstracts

14

15

Abstracts

An abstract is a brief summary of a text—a journal article, conference paper, or dissertation—that highlights its most important claims and findings. Since first appearing in medical journals in the 1960s, they have become common in every field of study except the humanities, where they are nonetheless not altogether absent.

Function

Abstracts serve different functions for different readerships:

- For ordinary readers, they summarize the text, allowing readers to decide whether to read the entire piece and organizing their comprehension by providing a "road map."
- For journal editors and reviewers, they offer a ready-to-hand reference for evaluating a text for publication.
- For indexers, professional abstract writers, and information management professionals, they offer guidance for classifying and sorting a text.
- For conference organizers, editorial boards, and funding agencies, they "advertise" and "sell" a research project or paper.

Most abstracts can be described as *informative, indicative*, or *critical*. An informative abstract presents research findings directly; an indicative abstract describes the text's discussion of a topic. Whereas an informative abstract might say, "we conclude that peer support networks can improve teachers' motivation," and an indicative abstract would simply say, "implications for teachers are discussed.” Critical abstracts function like executive summaries, addressing strengths, and weaknesses of a text.

Length

The form of an abstract varies from field to field and even from journal to journal (guidelines are often included in a journal’s instructions to authors). At variance is often the length: While traditional abstracts are typically about 150 words long, structured abstracts can be anywhere from 250 to 400 words, and abstracts for short communications, such as conference proceedings or technical notes, can be as short as 50 words.

Abstracts are typically written retrospectively, toward the end of the composing process, to represent completed work. But abstracts can also be prospective; when scholars apply for conference presentations or grant funding, they often submit abstracts for work yet to be done. In these cases, the abstract functions as a proposal or research trajectory. Such prospective abstracts can be 500 words or longer, depending upon guidelines provided.

Structure

Despite their different lengths, most abstracts attempt to make five rhetorical “moves”:

1. introduce the topic, its context, and its importance;
2. present the research question or purpose;
3. describe methods and materials used;
4. present key results and findings; and
5. discuss the significance of the findings for relevant audiences.

Abstracts for short communications will emphasize moves 3 through 5. A structured abstract, by contrast, will separate each move into its own paragraph with a subheading. Because structured abstracts tend to be longer, they provide more information to readers and are considered to be more useful. Structured abstracts have only been in use since the 1980s but are becoming increasingly common.

Mark Pedretti

See also Dissertations; Journal Articles; Literature Review; Methods Section; Results Section; Significance

Further Readings

Cremmins, E. T. (1996). The art of abstracting (2nd ed.). Arlington, TX: Information Resources Press.

Hofmann, A. H. (2013). Scientific writing and communication: Papers, proposals, and presentations (2nd ed., Chap. 14). Oxford, UK: Oxford University Press.

Huckin, T. (2001). Abstracting from abstracts. In M. Hewings (Ed.), Academic writing in context: Implications and applications (pp. 93–103). Birmingham, UK: University of Birmingham Press.

Swales, J. M., & Feak, C. B. (2009). Abstracts and the writing of abstracts. Ann Arbor: University of Michigan Press.

Qianqian Pan Qianqian Pan Pan, Qianqian

Meagan Karvonen Meagan Karvonen Karvonen, Meagan

Accessibility of Assessment Accessibility of assessment

15

19

Accessibility of Assessment

Accessibility originates from the field of architecture that aims to make buildings and the physical environment accessible to the whole population, including people with and without physical disabilities. For example, curb cuts help people in wheelchairs cross a street, but the same curb cuts also benefit people who are not in wheelchairs. Similarly, educational assessments that are accessible provide students with the opportunity to demonstrate their knowledge and skills on the construct being assessed, regardless of personal characteristics unrelating to what is being assessed. In contrast, an assessment with poor accessibility introduces barriers based on personal characteristics. There can be negative consequences when assessment results reflect those personal characteristics and not just the construct being assessed. This entry further defines *accessibility* as the term is used in assessment and discusses the related concepts of accommodation and universal design for assessment (UDA). It then discusses the accessibility of computer-based assessments, the relationship of accessibility to validity, and ways to maximize the accessibility of assessment.

Accessibility is a characteristic of the assessment, but it requires consideration of the test taker's interaction with items or tasks. Each test taker might experience individual barriers during the assessment process. For example, a student with low vision may have difficulty reading items printed in a standard font size on a paper and pencil test, while a student with very little computer experience might have difficulty navigating and choosing answers in a computer-based test. Accessibility is also a consideration in other types of assessments, not just tests. For example, consider a performance assessment that consists of giving an oral presentation, where grading criteria include speed and fluency of oral communication. This assessment could introduce barriers to students with communication. This assessment could introduce barriers to students with speech disorders and nonnative language speakers who are still working toward language proficiency.

When individual characteristics present barriers during the assessment, these characteristics limit the test taker’s ability to demonstrate what they know. As a result, examinees’ true knowledge and skills tend to be underestimated. This in turn limits the validity of inferences that can be made based on the scores. Therefore, it is important to ensure the assessment is as accessible as possible to everyone regardless of language proficiency, disability status, or other unique characteristics that might detract from the assessment. Steps can be taken to promote accessibility at the assessment design, administration, and scoring and reporting phases.

Related Concepts

Accessibility is related to concepts such as accommodation and UDA, but it is still a distinct concept. All three of these concepts have the same objective, which is to maximize test takers’ opportunities to demonstrate their knowledge or skills of the construct being tested—removing the influence of knowledge or characteristics unrelated to the construct while also not changing the construct being assessed. However, test accommodations and UDA are different from accessibility of assessment in some respects.

Accommodation refers to a change in the assessment administration methods or environment for test takers with disabilities or limited language proficiency. There are a variety of test accommodations that change the student’s experience with the assessment. Accommodations may be made in timing and scheduling, setting, presentation, and/or response. However, accommodations are designed to alter the student’s experience after the assessment is designed and are only granted by exception. Test accommodations are only applied to examinees who are eligible under relevant laws (e.g., in the United States, the Individuals with Disabilities Education Act, and Section 504 of the Rehabilitation Act cover accommodations for individuals with disabilities). Decisions about which accommodations should be provided are made on a case-by-case basis.

UDA is a framework for promoting accessibility through all stages of assessment design and delivery. UDA often starts by considering inclusiveness of the population that is intended to be eligible for the assessment. The ultimate goal of UDA is to enhance the accessibility of the test, thereby improving the goal of UDA is to enhance the accessibility of the test, thereby improving the fairness of the assessment for all students and the validity of inferences made from test scores for all test takers. Both UDA and accessibility of assessment apply to the whole population, including people with and without disabilities or limited language proficiency. Accessibility of assessment can be regarded as one facet of UDA, focusing on the interaction of the test taker and the assessment content. Features of the items themselves and supports made available to test takers both facilitate accessibility.

Accessibility of Computer-Based Assessments

Computer-based assessments are becoming more prevalent in large-scale assessment programs and in classrooms. In addition to benefits such as cost and time savings, computer-based assessments also provide new opportunities to support accessibility. For example, it is much easier to customize item delivery methods and response formats via computer-based assessment than paper and pencil assessment. Additionally, computers increase the availability and flexibility of the use of accessibility supports. Some computer-based assessments treat accessibility features as tools that test takers may enable or disable on demand. For example, a student may turn on a magnification tool when viewing a diagram and turn it back off when reading a paragraph. Other common tools include on-screen highlighters, text-to-speech functions that read text aloud, changes to color configurations, and overlays that mask parts of the information on screen. Some systems are also compatible with assistive technology devices that allow individuals with disabilities to interact independently with the computer even without a standard keyboard and mouse.

While computer-based assessments can improve accessibility, there is also a risk of measuring knowledge and skills that are unrelated to the assessment. Test takers must be familiar with the testing platform and with the accessibility supports that are used. A test taker who lacks basic computer skills or access to computers in daily life may not be able to use an online system without some practice or coaching in advance. And when it comes to accessibility tools, more is not necessarily better. The test taker should be familiar with the tool, and the tool should not introduce distraction or confusion. Test developers can provide guidance to test administrators and test takers on the accessibility tools in order to promote good decisions about the use of tools to minimize barriers. Organizations such as the iMS Global Learning Consortium promote the use of common standards for accessibility across technology platforms in various sectors including education.

Relationship of Assessment Accessibility to Validity

Evidence of an assessment’s accessibility influences the validity of inferences that may be made about its results. An accessible assessment measures the intended construct and avoids measuring unrelated characteristics. Sources of construct-irrelevant variance that occur in assessments with poor accessibility include item bias, unclear presentation of information in instructions, and lack of or inappropriate use of supports.

It is important to avoid construct-irrelevant variance because poor estimation of a student’s true knowledge and skills can lead to negative consequences. For example, if results of a high school mathematics exam are used to determine whether a student has met graduation requirements and the items contain complex vocabulary unrelated to mathematics, a student with limited language proficiency may fail to pass the exam and meet the graduation requirements even while possessing the relevant mathematics knowledge and skills. An assessment is biased when it presents barriers to a subgroup of students and their results are negatively impacted despite their having the same construct-relevant knowledge and skills as other test takers.

Maximizing the Accessibility of Assessment

The accessibility of an assessment is maximized when the issue is considered at all phases of the assessment’s life span: during assessment design and development, during administration, and after results are available. Accessibility may be planned or evaluated at each of these steps.

During the Assessment Design and Development Phase

During the test development stage, developers can promote accessibility using an evidence-centered design framework. Using evidence-centered design, test developers carefully describe the construct being assessed and detail the behaviors that are required to demonstrate understanding of the construct. These descriptions can be reviewed before they are used to guide item writing in order to ensure there are no unintended barriers to demonstrating knowledge of the content for subgroups of individuals. Once items are developed, they can also be reviewed for evidence of barriers related to construct-irrelevant factors (e.g., unique cultural knowledge that would be required to answer the question, use of complex vocabulary unrelated to the measured content).

In addition to considering accessibility during item development, the test as a whole may be designed to minimize barriers. For example, test developers should make sure directions are clear and that the items are displayed in ways that minimize confusion. Assessment developers also define the supports that may be made available to increase accessibility without advantaging any subgroups. Items should not be easier or harder for examinees who choose to use extra supports.

In the prototype or early test design phrases, cognitive labs are a common method to evaluate whether items or tasks elicit the intended cognitive process. In a cognitive lab, test takers report their thoughts (i.e., think aloud) when they are reading, interpreting, and responding to an item. This method helps test developers evaluate whether item features are performing as intended and check for problems with clarity of the item content. Cognitive labs and observation methods can also be used to evaluate whether accessibility supports have the intended effect.

During the Assessment Administration Phase

During the assessment administration phase, educators and students play roles in ensuring accessibility is maximized. Educators must understand the supports that could be provided to each student. When they also choose supports for individual students, their choices should match the students’ current needs and preferences. Supports selected to meet a student’s need should not introduce unintended barriers. For example, enlarged font may be helpful for test takers with low vision; however, if the item cannot be shown in one page, the need for scrolling or reading an item across pages may introduce unintended barriers. When test takers themselves have the opportunity to choose the supports they use, they need enough information about the support to make informed decisions. Selected supports should be familiar to the student, either through use during instruction or through opportunities to practice similar activities prior to testing. Finally, educators may be responsible for providing some supports directly during an assessment. For example, an educator may read items aloud or help a student enter answers for a computer-based test. In these cases, educators need to administer the supports with fidelity and follow instructions for standardized administration.

During the test administration phase, teacher surveys, teacher interviews, and observations may be used to evaluate accessibility. Observations can determine whether an educator can implement accessibility supports with fidelity and whether the student was able to use the supports as intended. Through interviewing or surveys, teachers can provide their suggestions on usability, effectiveness, and suggestions for supports that tend to improve the accessibility of assessment.

After Administration

Once an assessment is administered to enough students, it is possible to evaluate accessibility even further. For example, there are statistical techniques that can be applied to determine whether items may be biased for subgroups of students. Differential item functioning analysis is commonly used to detect whether different groups of test takers with the same true ability have performed differently on items. Differential item functioning analysis can be used to check for potential item bias—for example, for groups of students with limited language proficiency, lower socioeconomic status, or disability. There are different statistical procedures to detect differential item functioning items, such as logistic regression and Mantel–Haenszel statistics. Items that are found to function differently for different subgroups of students should be investigated further. For example, a panel of educators could convene to review items and identify features that may be disadvantaging students. These techniques require data based on large samples, so they are most appropriate for large-scale assessments.

Although assessment developers may have accessibility as a goal during the design phase, planning for accessibility does not guarantee that the goal was met. Collecting evidence related to item and test features, and student experiences with accessibility supports, allows for ongoing evaluation and improvement of assessments for all students.

Qianqian Pan and Meagan Karvonen

See also Accommodations; Differential Item Functioning; English Language

Proficiency Assessment; Intellectual Disability and Postsecondary Education; Second Language Learners, Assessment of; Test Bias; Universal Design in Education

Further Readings

Beddow, P. A., Kurz, A., & Frey, J. R. (2011). Accessibility theory: Guiding the science and practice of test item design with the test-taker in mind. In S. N. Elliott (Ed.), Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy (pp. 163–182). London, UK: Springer Science … Business Media.

Iwarsson, S., & Ståhl, A. (2003). Accessibility, usability and universal design: Positioning and definition of concepts describing person-environment relationships. Disability and Rehabilitation, 25(2), 57–66.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. The Journal of Technology, Learning and Assessment, 4(2). Retrieved from http://www.jtla.org

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. Educational Measurement: Issues and Practice, 27(3), 3–16.

Russell, M. (2011). Accessible test design. In M. K. Russell (Ed.), Assessing students in the margin: Challenges, strategies, and techniques (pp. 407–424). Charlotte, NC: Information Age Publishing.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/nceo/onlinepubs/Synthesis44.html

Thurlow, M. L., & Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. Review of Research in Education, 39, 331–369.

Websites

iMS Global Learning Consortium https://www.imsglobal.org/

Judith R. Harrison Judith R. Harrison Harrison, Judith R.

Jeanette Joyce Jeanette Joyce Joyce, Jeanette

Accommodations

Accommodations

19

23

Accommodations

Accommodations are defined as adjustments for variances. As such, educational accommodations, the topic of this entry, are strategies utilized to remove content irrelevant variance from an assignment or test, allowing students to demonstrate what they have learned in relation to the specific academic content standard being targeted, without noise associated with any impairment related to a disability. For example, students with delayed processing may be given extended time on a standardized assessment in order to more accurately assess what they have learned rather than how rapidly the student can process the questions and retrieve the answers.

This entry further defines educational accommodations and looks at some of the issues surrounding the selection and use of accommodations. It then discusses the use of accommodations in classrooms and in testing, strategies for selecting accommodations, and selection of alternative interventions to teach students skills needed to address impairment for which accommodations are typically provided.

Accommodations are often confused with modifications given to students with disabilities. Although these terms are sometimes used interchangeably, they are not the same. Modifications represent a difference in what the student is expected to learn, as in having a student learn multiplication facts up to 5× while the class learns multiplication facts up to 10×. Conversely, the use of accommodations does not lower standards or change expectations. The use of accommodations provides a differential boost between those with and without disabilities. Thus, when students with and without disabilities utilize an accommodation, a greater increase in performance should be evident for students with the disability than for students without the disability. For example, if a student is deaf and communicates with American Sign Language, it is highly probable that the student's performance will increase with an American Sign Language interpreter interpreting instruction and interactions in the classroom, but it is not likely that the performance of students in the classroom who are not deaf will be affected.

Accommodations typically are changes in how the task or test is presented (e.g., questions read aloud), how the student is expected to respond (e.g., dictating answers), and the time allowed for the task (e.g., time and a half). For testing situations, an accommodation may be granted for students to take the test in a smaller group setting or in a distraction-free room. They may also be situation-specific. That is, the student may be provided extra time on writing tasks but not for math problem sheets.

An increased focus on accountability and use of high-stakes testing in schools, along with greater inclusion of students with disabilities in general education settings and the need to assure equal access to the general education curriculum and grade-level content standards, all heighten the need to understand the process of accommodating for impairment associated with disabilities. Teachers are expected to teach the same content to all students, those with and without disabilities, and students are expected to demonstrate proficiency on all standards. At the same time, assessment scores must represent what students have learned specific to the content being evaluated and not be affected by extraneous variance associated with a disability, such as the effects of a reading disability when interpreting a math word problem assessment. Thus, for students with disabilities served by special education, individualized education program (IEP) teams are charged with selecting appropriate accommodations.

However, issues exist with the selection and use of accommodations. Although federal law mandates the use of accommodations, research is far behind and provides minimal support to educators for the selection of specific accommodations for specific areas of impairment. In addition, an overreliance on accommodations exists and IEPs lack interventions specifically to teach strategic skills needed to decrease the negative effects of the disability on academic performance. A third concern about accommodations is a philosophical one. Changing or individualizing an assessment for some students conflicts with the goals of universal design that prefers a single assessment that is “accessible” to all students. Accommodations and universal design share an underlying principle, however, which is that the scores from assessments should be equally valid for all.

History of Accommodations

The history of educational accommodations is intertwined with the history of the inclusive education movement in the United States. Until the first federal special education law was adopted in 1975, schools in the United States first denied students with disabilities an education and then marginalized them to institutional or classroom “holding pens.” The federal Education for All Handicapped Children Act referred to the “least restrictive environment,” which indicated that students should be educated, whenever educational benefit could be achieved, in general education classrooms with their age-equivalent peers.

This language was further strengthened in the reauthorizations of the act in 1990 (the Individuals with Disabilities Education Act) and in 2004 (the Individuals with Disabilities Education Improvement Act). With this encouragement to educate students with disabilities in general education classrooms with support, the need to reduce the cognitive load of tasks and assessments to the most salient elements led to further development of accommodations. These accommodations are student-specific and are codified in an IEP or Section 504 plan.

Over the years, courts have established that school staff members are responsible for implementing the accommodations described in the IEP. The teacher responsible for teaching the child-specific content for which accommodations are specified on the IEP is legally bound to provide the accommodation. If an IEP team selects an accommodation and includes it in the student’s IEP and the accommodation is not implemented, then the child did not receive a free appropriate public education. If school districts do not inform teachers of the student’s IEP, then the district may be liable; and if the teacher elects not to implement the accommodations, then the teacher may be liable. If a teacher believes that an accommodation on an IEP is not in the best interest of a student, then the teacher should request an IEP meeting to discuss the accommodation rather than changing or disregarding the accommodations as written. Two methods of selecting accommodations are discussed in the next section.

Selecting Accommodations

Selecting Accommodations

Following best practice procedures, accommodations are selected based on the strengths and needs of the child identified on the child’s IEP under the category of Present Levels of Academic and Functional Performance, the grade-level standards, and the instructional tasks being used to assess proficiency on the standards. The Present Levels of Academic and Functional Performance is a section in the IEP that identifies “how the child’s disability affects the child’s involvement and progress in the general education curriculum” [IDEA Sec. 614 (d) (1) (A) (i) (I)].

However, reviews of IEPs published by Craig Spiel and colleagues (2014) and Connie Schnoes and colleagues (2006) suggest little rhyme or reason to the accommodations included on most IEPs. It appears, from these studies and anecdotal evidence, that educators select many accommodations from a laundry list without considering the impairment that is interfering with academic progress or the feasibility of all of the accommodations being implemented by teachers in classrooms. In fact, as of 2016, there are IEP writing software packages that include pull-down menus, allowing practitioners to “shop” for accommodations. Although these are editable, many practitioners report defaulting to the options provided in the software.

One of the accommodations most frequently listed in IEPs is the provision of extra time. Many students, regardless of their strengths and weaknesses, receive extra time on assignments and assessments, although empirical evidence is lacking as to whether this accommodation improves performance across the spectrum of disabilities.

One means of addressing this issue is for educators to follow procedures outlined in the accommodations manual published by the Council of Chief State School Officers to select, administer, and evaluate the outcomes associated with accommodations. The accommodation manual was written to provide guidelines to states for the selection and administration of accommodations and is written in such a way that state-level rules can be added as needed.

The accommodation guide provides four steps for selecting and implementing accommodations supplemented by 11 tools. The steps are to (a) expect students with disabilities to achieve grade-level academic content standards, (b) learn about accommodations for instruction and assessment, (c) select accommodations for instruction and assessment for individual students, and (d) administer accommodations during instruction and assessment. Potentially, there is room for a fifth step in which the benefits of the administered accommodations are considered and the selection revised as indicated.

The third step, select accommodations, is the focus of this section and suggests that IEP teams should consider seven factors when selecting accommodations: (1) student characteristics identified in the Present Levels of Academic and Functional Performances, (2) inclusion characteristics that need accommodation, (3) strategies to include the student in the process, (4) effectiveness of prior accommodations, (5) accommodations for instruction and assessment, (6) individual test characteristics, and (7) state accommodation policies. Specifically, IEP teams first consider student strengths and weaknesses and then the impact of the impairment associated with the disability on the student’s learning. Additionally, the team must determine the type of specialized instruction needed to master the grade-level content standards, drawing from the results of prior tasks when the specific accommodation was used.

Potential problems with the accommodations, and the perceptions of teachers and the student regarding the need and effectiveness of the accommodation, help to determine whether the accommodation should be included in the IEP. Finally, teams should consider the willingness of the student to utilize the accommodation, the need for the accommodation across educational settings, and the acceptability of the accommodation on high-stakes assessment. Accommodations used on high-stakes testing should be the same as those used in the classroom. Introducing students and teachers to a new accommodation on a high-stakes test is not effective.

Along the same lines, another model for selecting strategies for inclusion in IEPs, including accommodations, was developed by Judith Harrison and her colleagues. This model targets the use of accommodations that are being provided to students with the potential to learn the missing skill for which accommodations are being provided. One example would be teaching a student to take notes, as opposed to accommodating inattention to class discussion by providing a student with a copy of teacher notes, eliminating the need to take notes or as a supplement to notes taken by the student. It is assumed that these students can, in fact, improve their note-taking skills with scaffolding. Alternatively, this approach is not appropriate when accommodations are needed for a skill that cannot be taught, such as when a student is deaf and needs a sign language interpreter.

language interpreter.

The model developed by Harrison and colleagues is founded on the life-course model for mental health treatment selection developed by Steven Evans and colleagues. The premise behind the LCM is that services are provided to increase the acquisition of skills needed across the life span, focusing on both short-and long-term concerns. Similarly, Harrison and colleagues' model for IEP strategy selection is designed to assist IEP teams in selecting strategies designed to teach skills and provide accommodations only when the student does not have the capacity to learn the skills, or expectations need to be adjusted at the start of an intervention and faded with mastery of the skill.

An example of the use of Harrison and colleagues' model would be if a student is in an inclusive history class and does not attend and take notes effectively, the teacher could teach the student to self-monitor attention to task and note-taking skills. However, while she is teaching note-taking skills, she continues to teach content associated with grade-level content standards. In order for the student to keep up with the instruction while he is learning note-taking strategies, the teacher could give him a copy of her notes (an accommodation).

In this model, students are provided the accommodation until they have mastered the skill needed, after such time the accommodation is no longer needed. The length of this time period is dependent on the individual student. Accommodations are selected that directly address targeted areas of impairment and progress is monitored and accommodations changed if sufficient progress is not documented. For example, students with attention-deficit/hyperactivity disorder frequently struggle with completing and submitting homework due to organizational skill deficits. To address this issue, the child might meet with a counselor each morning before school, who would help her organize her binder. The counselor would continue to help the child with organization using a structured checklist to guide the process. However, the counselor would scaffold assistance, withdrawing support as the student learns to organize the binder without assistance.

Potential Accommodations

As previously mentioned, multiple accommodations are often recommended without any empirical evidence to support their effectiveness or usefulness in accommodating for impairment. In the following section, examples of potential accommodations (potential, as research is needed to determine whether the accommodations (potential, as research is needed to determine whether the strategies truly provide a differential boost) are described for the four areas of accommodations: presentation, response, timing/scheduling, and setting. However, accommodations must be selected based on the criteria described earlier, with a strong focus on student need. In addition, emphasis is placed on progress monitoring to determine that the student is benefiting from the use of the accommodation and that the accommodation does not cause harm, such as a reduction in effort or motivation.

Presentation accommodations are changes in the manner in which instruction, assignments, and/or assessments are presented to the student. For example, test or assignment questions might be read to a student, removing the need for the student to read the question by him-or herself. It is a frequent accommodation for students with reading disabilities to have math word problems read to them. In addition, recent evidence suggests that reading tests aloud to students with attention-deficit/hyperactivity disorder helps the student maintain attention to task even when the student does not have a reading disability. Students who have visual or hearing impairments frequently receive presentation accommodations, such as enlarged or magnified content or audio amplification, in order to access the information being taught.

Response accommodations are changes in the manner in which students respond to instruction via assignments, assessments, or organizational devices used by the student to determine and write a response. For example, a student who struggles with formulating written responses might be allowed to answer questions verbally instead of writing them on a test. This form of response might or might not include a scribe who writes the answers for the student or voice to text software to formulate a response to a question or to write a paper. Several potential response accommodations are considered methods of increasing active engagement for an entire class of students, such as choral response using whiteboards to respond to teacher questions or clickers to be used with an interactive whiteboard. Additionally, there are many technology-based strategies that can be used as response accommodations or as simply good teaching strategies to increase engagement classwide.

Timing and scheduling accommodations, those that change the amount or organization of time for an assignment or a test, are the most frequent accommodations. In fact, the one most frequently found accommodation on IEPs, extended time, falls within this category. Extended time is the provision of extra time for a task. For example, students who process information or read more slowly than others might be given 90 minutes instead of 60 minutes to complete a timed assessment. Recent research suggests that extended time is not an effective accommodation for students with behavioral disorders such as attention-deficit/hyperactivity disorder. Other timing accommodations include frequent breaks and giving tests at specific times of the day, such as in the morning.

Setting accommodations are those that change the location in which the assignment or test is completed. For example, students who are easily distracted frequently have accommodations of testing in small groups or in a distraction-free environment.

Alternative Interventions

Following the model of strategy selection for IEPs developed by Harrison and colleagues, interventions to teach skills are included prior to or in conjunction with accommodations. For example, students who struggle with initiating and maintaining attention to task are frequently given extra time to complete tasks with the rationale that more time is needed to compensate for the time spent off task. However, this does not teach the student the skill of attending; it merely reduces the expectation to complete tasks in the same amount of time as typically developing peers. Self-management is an intervention that can be tailored to teach students to attend to task. Specifically, students can be taught to self-monitor and document whether they were paying attention at a given interval over a set number of intervals. Students are taught not only to self-monitor but to set a goal for the number of intervals in which they will attend and then reward themselves when they meet their goal.

Accommodations are intended to facilitate demonstration of academic mastery by a student with disabilities, minimizing the impact of the specific disability on the student’s performance. These accommodations relate to how the task is presented to the student as well as how the students make their response and differ from modifications that change what is asked of the student. When paired with strategic interventions such as described earlier, accommodations have the potential to be beneficial to teachers and learners. However, there is some indication that this is not always being carried out as intended and much more research and diligence in practice is needed.

Judith R. Harrison and Jeanette Joyce

See also Attention-Deficit/Hyperactivity Disorder; Individualized Education Program; Individuals with Disabilities Education Act; Least Restrictive Environment; Progress Monitoring; Universal Design in Education; Universal Design of Assessment

Further Readings

Christensen, L., Carver, W., VanDeZande, J., & Lazarus, S. (2011). Accommodations manual: How to select, administer, and evaluate use of accommodations for instruction and assessment of students with disabilities (3rd ed.). Washington, DC: Assessing Special Education Students State Collaborative on Assessment and Student Standards, Council of Chief State School Officers.

Evans, S. W., Owens, J. S., Mautone, J. A., DuPaul, G. J., & Power, T. J. (2014). Toward a comprehensive life-course model of care for youth with attention-deficit/hyperactivity disorder. In Handbook of School Mental Health (pp. 413–426). Springer.

Harrison, J. R., Bunford, N., Evans, S. W., & Owens, J. (2013). Educational accommodations for students with behavioral challenges: A systematic review of the literature. Review of Educational Research, 83(4), 551–597. doi:10.3102/0034654313497517

Schnoes, C., Reid, R., Wagner, M., & Marder, C. (2006). ADHD among students receiving special education services: A national survey. Exceptional Children, 72, 483–496. doi:10.1177/001440290607200406

Spiel, C. F., Evans, S. W., & Langberg, J. M. (2014). Evaluating the content of individualized education program and 504 Plans of young adolescents with attention deficit/hyperactivity disorder. School Psychology Quarterly, 29, 452–468. doi:10.1037/spq0000101

Spiel, C. F., Mixon, C. S., Holdaway, A. S., Evans, S. W., Harrison, J. R., Zoromski, A. K., & Sadler, J. M. (2016). The effect of reading tests aloud on the performance of youth with and without ADHD. Remedial and Special Education, 37(2), 101–112. doi:10.1177/0741932515619929

Kenneth K. Wong Kenneth K. Wong Wong, Kenneth K.

Accountability

Accountability

23

25

Accountability

Accountability is a theory of action on raising student performance by applying pressure on, and providing support for, schools and districts that do not meet academic standards. Annual or periodic reporting on school performance forms the basis of actions to address academic needs. Simply put, what gets measured and reported receives attention from stakeholders in the public arena. This entry further defines accountability in the context of public education before describing the federal role in school accountability in the United States and how it has changed since the mid-20th century.

In the United States, accountability is defined in the context of a decentralized public education system. With a federal system of governance, states assume a leading role in primary and secondary education. The constitution in each of the 50 states affirms state responsibility in this policy domain. States and their localities continued to provide about 90% of the funding in public education. States exercise control over their academic content standards, educator preparation and recruitment, and the scope of intervention in low academic performance.

State dominance notwithstanding, accountability in public education has become a shared state–federal function. The 1960s marked the beginning of an active federal role to address educational inequity and the achievement gap between students from low-and high-income families. The U.S. Congress has established a grants-in-aid system to target federal support for students with particular needs, such as low-income students, English-language learners, Native Americans, and students with learning disabilities.

Grants from the federal government account for about 10% of total public school spending. In return for federal dollars, states and school districts are required to comply with federal standards on assessing students. Federal involvement in accountability intensified in 2001 when Congress passed the No Child Left Behind Act (NCLB). With the 2015 passage of the Every Student Succeeds Act (ESSA), states have regained some control over accountability policy.

NCLB expanded the federal role in educational accountability. The federal law required annual testing of students at the elementary grades in core subject areas, mandated the hiring of “highly qualified teachers” in classrooms, and granted states and districts substantial authority in taking “corrective actions” to turn around low-performing schools. Further, the law provided school choice to parents to take their children out of failing schools. Equally significant was NCLB’s intent to close achievement gaps among racial and ethnic subgroups as well as subgroups based on income, limited English proficiency, and special education.

Under NCLB, to determine whether a school met adequate yearly progress (AYP), student achievement for each school was aggregated by grade and subject area. All students in Grades 3–8 and one additional grade in high school were tested annually in mathematics and in reading/English-language arts. In addition, students in select grades were tested in science. The school-level report included the percentage of students proficient in each of the core-content areas, student participation in standardized testing, attendance rates, and graduation rates.

Equally prominent is the equity focus on NCLB. Depending on their socioeconomic characteristics, schools were required to report the academic proficiency of students of the following subgroups: economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and limited English proficiency students. In this regard, for accountability purposes, the NCLB promoted transparency on student progress. Schools that persistently failed to meet AYP were subject to a gradation of intervention, including school closure or conversion to a charter school.

The federal accountability agenda, as articulated in NCLB, encountered implementation problems. Tension occurred between the theory of accountability based on the federal intent and the practice of accountability at the state and local level. More specifically, the decentralized education system allows for varying degrees of policy specification and academic rigor across allows for varying degrees of policy specification and academic rigor across states. Within each state, the decision-making process allowed for multiple stakeholders to weigh in on the rigor, scope, timing, and cost of student academic assessment. Consequently, state assessments vary widely in terms of the level of rigor, as indicated by the substantial gap in many states between student proficiency on state tests and their performance on the National Assessment of Educational Progress, a set of standardized tests used throughout the United States.

Further, the NCLB accountability agenda encountered social constraints. The extent to which a district or a school met AYP was affected by the presence of student subgroups, including low-income students, English-language learners, students with disabilities, and racial and ethnic minorities. One study of AYP data in California and Virginia found that schools with more student subgroups experienced more difficulty in meeting AYP.

Recognizing the implementation problems with NCLB, in 2011, the Obama administration began granting waivers to states that exempted them from some provisions of the law, including meeting the target that 100% of students demonstrate proficiency on state tests by the end of the 2013–2014 school year. Over 40 states received the waivers.

States that received waivers still had to test students in certain core subjects annually, especially in Grades 3–8 and in one high school grade, and hold schools accountable for performance standards. In addition, states had to adopt reading/language arts and mathematics standards that were common to a number of states or had to show that their standards were certified by a state network of higher education institutions.

The simplest way to meet the waiver requirement on academic standards was to adopt the Common Core State Standards, which had already been adopted by many states. States were also encouraged to adopt the Common Core in exchange for federal funding as part the Race to the Top grant competition begun in 2009. With the Common Core, states can compare their academic progress to that in other states, improve economies of scale in terms of technical assistance, and streamline teacher recruitment and support. At the same time, there have been problems with the implementation of Common Core, including concerns that teachers were not adequately trained in the standards and the assessments aligned to them. In addition, many have criticized the use of financial and deregulatory incentives to encourage adoption of the Common Core as federal intrusion into state and local school affairs

Core as federal intrusion in state and local academic affairs.

In late 2015, the U.S. Congress replaced NCLB with the ESSA. To be sure, ESSA continues to build on the NCLB accountability policy. In particular, states will continue to conduct annual testing of core subjects in students of Grades 3–8 as well as in one grade in high school. States are required to issue annual report cards that show the performance of students from various subgroups, including students from low-income families, English-language learners, children with learning disabilities, and various minority groups.

Departing from NCLB, ESSA signals the return of some state control over accountability. ESSA places limits on federal prescriptions on intervening in low-performing schools. Under the law, states are required to adopt “challenging” standards, but the U.S. education secretary cannot use incentives encouraging states to adopt a particular set of standards. In addition, the law does not require states to set up teacher-evaluation systems that incorporate students’ test scores, as they were required to do to receive waivers from NCLB. ESSA also allows states to use multiple measures to assess student performance.

Under ESSA, states have gained control over several important aspects of education accountability. States can decide on academic standards, including developing their own standards and multiple measures of academic assessment. States can also establish criteria in identifying low-performing schools for direct intervention, although ESSA expects that states will focus on the bottom 5%. Evaluation of teachers will be determined by states, which do not have to use student test results as the basis for the evaluation.

Kenneth K. Wong

See also Accreditation; Achievement Tests; *Brown v. Board of Education*; Common Core State Standards; Every Student Succeeds Act; Great Society Programs; High-Stakes Tests; National Assessment of Educational Progress; No Child Left Behind Act; Partnership for Assessment of Readiness for College and Careers; Race to the Top; Smarter Balanced Assessment Consortium; Standards-Based Assessment

Further Readings

Cohen, D., & Moffitt, S. (2009). The ordeal of equality. Cambridge, MA: Harvard University Press.

Hess, F., & Petrilli, M. (2006). No child left behind primer. New York, NY: Peter Lang Publishing.

Kim, J., & Sunderman, G. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. Educational Researcher, 34 (8), 3–13.

Manna, P. (2010). Collision course. Washington DC: Congressional Quarterly Press.

Wong, K. K., & Nicotera, A. (2007). Successful schools and educational accountability. Boston, MA: Pearson Education.

Jacqueline Remondet Wall Jacqueline Remondet Wall Wall, Jacqueline Remondet

Accreditation

Accreditation

25

27

Accreditation

Designed to protect public health, safety, and interest, accreditation provides a system of quality assessment and improvement. In the United States, educational, human services, and health-care programs and institutions undergo accreditation review. Although each of these three sectors has unique accreditation processes, accreditation generally consists of a process of voluntary, external review that occurs and results in a decision based upon the institution’s consistency with accepted standards. This entry discusses the history of accreditation in the United States, the process of accrediting U.S. higher education institutions, criticisms of the accreditation system in higher education, and supporters’ responses to these criticisms.

Many accrediting practices provide recommendations to increase compliance and, therefore, offer opportunities for program improvement to those undergoing review. The accreditation process has evolved over the years in the United States and offers advantages and disadvantages for higher education and other professional entities. Accreditation started in higher education in the late 19th century as a way to verify student qualifications for entry into colleges and universities. This led to the formation of regional groups of higher education administrators to evaluate secondary education practices.

The federal government entered accreditation with the Servicemen’s Readjustment Act, commonly known as the GI Bill, in 1944 by providing educational funding for World War II servicemen. In 1952, this legislation was reauthorized and included a process of peer review to establish the legitimacy for institutions offering educational services. Since then, the role of the federal institutions offering educational services. Since then, the role of the federal government has continued through legislation enacted, establishing the U.S. Department of Health, Education, and Welfare in 1953. In 1979, when the U.S. Department of Education (ED) was created, the US Department of Health, Education, and Welfare was separated. The ED is designated to establish educational policy, to coordinate federal assistance to the educational enterprise, to enforce civil rights legislation in education, and to collect information on schools within the United States.

The emphasis placed on actions by the ED is to promote achievement of students and make schools accountable. The ED does not establish academic institutions or programs, nor does it perform accreditation. However, the Higher Education Act (1965) and subsequent amendments to the Higher Education Opportunity Act (2008) authorize the U.S. secretary of education to publish lists of recognized accrediting agencies. These recognized accrediting bodies not only provide ratings of educational quality, but many also allow students to access federal funding established by Title IV in the Higher Education Act. Specifically, Title IV authorizes programs to accept government monies to allow access to higher education; this funding requires state licensure of the institution and accreditation by an ED-recognized accreditor.

In addition to the ED, the Council for Higher Education Accreditation provides recognition of accrediting agencies. Council for Higher Education Accreditation, a voluntary membership organization with more than 3,000 institutions represented, establishes the quality of agencies that accredit programs and institutions that are regional, faith-based, career-focused, and specialty/programmatic in nature.

Accreditation in Higher Education

In the United States, higher education accreditation incorporates three separate pathways, termed *the triad*: the federal government, state governments, and accrediting organizations. Established to provide public protection through combined regulation processes set by governments and the development of peer-evaluation systems, these processes may be focused on the institution or on individual programs of study. States provide authorization and regulate educational institutions that operate within state boundaries. Accreditors evaluate educational system inputs as well as the effectiveness of education through examining student achievement and outcomes of the process.

Accrediting bodies may be regional associations that review entire institutions; national associations that primarily evaluate career, vocational, and trade schools; or specialty and programmatic accreditors that examine individual programs of study (e.g., medicine, dentistry, and teaching). Accreditation by some national and specialty accreditors and all regional accreditors provides access to Title IV funding.

The review processes typically begin with the institution or program being accredited, creating a self-evaluation. These self-evaluations are reports demonstrating compliance with standards or guidelines developed by the accrediting body. Next, it is common to have the self-evaluation reviewed and a visit to the site following this review. Volunteer peers not affiliated with the program or institution undergoing evaluation conduct these visits, reports from which are sent to the accrediting organization for review. The accrediting body then examines all information provided to measure compliance with a set of accreditation standards. Those programs and institutions meeting the standards are granted a limited time period of accreditation before they are reviewed. Most require information reported during this time and operate under a set of complex guidelines in the accreditation process.

In some countries, accreditation is a mandated government process, while in the United States, accreditation is founded on voluntary participation. Therefore, for the publics served by accredited programs and institutions, accreditation serves to indicate that the standards of quality are being met and the program or institution operates in accord with the agreed upon policies, procedures, and practices of the accrediting body. Irrespective of the authority and type of accreditation, the process represents a formal evaluation of an organization, a program, or a service against the best practice standards.

Accreditation standards are exemplars of quality standards that are developed by those within the area being assessed or subject matter experts. Subject matter experts develop these standards through an iterative process, reflecting that which is being evaluated in a manner that is acceptable to the broader group establishing and supporting the method of evaluation.

Although the system of accreditation provides a basis for ensuring that graduates of programs have the knowledge, skills, and abilities warranted for the type of education, the process is not without its critics. Some allege that the accreditation system is not understandable to the public; others complain that it remains difficult to compare institutions and that the process is not well received remains difficult to compare institutions and that the process is not well received by many who are subject to it. Some institutions see it as a burdensome and costly process, without substantive benefit. One challenge in evaluating educational quality is that student behavior, for example, motivation, in combination with inherent traits, ultimately influences learning and, therefore, achievement and outcomes. Even with such challenges, though, the present system maintains that it establishes an acceptable level of quality and accountability. Supporters state that for more than a century, the peer-review process effectively and collegially provided an approach to evaluate educational programs and institutions.

Jacqueline Remondet Wall

See also Certification; Program Evaluation

Further Readings

Association of Specialty and Programmatic Accreditors. (n.d.). About accreditation: Resources, documents and definitions. Retrieved from http://www.aspa-usa.org/about-accreditation/

U.S. Department of Education. (n.d.). Accreditation in the United States. Retrieved from http://www2.ed.gov/admins/finaid/accred/index.html?exp=2

U.S. Department of Education. (n.d.). The database of accredited postsecondary institutions and programs. Retrieved from http://ope.ed.gov/accreditation/Agencies.aspx

Websites

Council for Higher Education Accreditation http://www.chea.org

Yi-Fang Wu Yi-Fang Wu Wu, Yi-Fang

Achievement Tests

Achievement tests

27

33

Achievement Tests

The term *achievement tests* refers to tests designed to measure the knowledge, skills, and abilities attained by a test taker in a field, in a subject area, or in a content domain in which the test taker has received training or instruction. This entry first clarifies the difference between achievement tests and aptitude tests and introduces a brief history of achievement tests in the United States. It then describes the purposes of achievement tests, types of tests, and major steps in developing and administering achievement tests. The entry concludes with an overview of the benefits and limitations of achievement tests.

Achievement Tests Versus Aptitude Tests

Before the 21st century, achievement tests were not always distinguished from aptitude tests. William H. Angoff asserted that in educational assessments, there is neither a very clear distinction between achievement and aptitude nor a sharp difference between measures of achievement and measures of aptitude. The aim of aptitude tests is to indicate a test taker’s readiness to learn or to develop proficiency in some particular area if instruction or training is provided; achievement tests can serve the same purpose.

The main difference between the two constructs is that the achievement test is confined to a single subject area more completely than is the aptitude test. That is, while items and tasks on the achievement tests are based on specific content standards or are dependent on the materials in the curriculum that examinees are expected to learn in a subject area, those on aptitude tests may be based on skills not explicitly taught in school. Depending on the intended purposes for which the test was developed, the results from an achievement test can be used, for example, for assessing proficiency levels, diagnosing strengths and weaknesses, assigning grades, certification, licensure, course placement, college admission, curriculum evaluation, and school accountability.

Brief History of Achievement Tests in the United States

In 1845, the Boston School Committee led by Samuel G. Howe initiated a large-scale, group-administered written examination to facilitate comparisons across classrooms and to monitor schools' effectiveness. This test is probably the prototype of contemporary achievement tests in the United States, although the term *achievement tests* was not prevalent at the time. This test, which was intended to efficiently measure the knowledge and skills of a great number of students, carried many of the features relevant to the large-scale state and district tests in the late 20th and early 21st centuries. In addition to monitoring the effectiveness of schools, achievement tests of the mid-19th century were designed for selection purposes.

The publication of arithmetic and handwriting tests by Edward L. Thorndike and his students in 1908 symbolized the inception of the unceasing achievement testing movement. At the beginning of the 20th century, there had been some hundreds of achievement tests available for use in elementary and secondary schools; nearly 100 of them were standardized and a variety of content areas were measured: arithmetic, English, geography, handwriting, history, Latin, mathematics, modern languages such as French and Spanish, reading, science, and spelling. The development of achievement test batteries that were designed to inform the public about student learning and school effectiveness across multiple grade levels arose around that time.

The Stanford Achievement Tests developed by Truman L. Kelley, Giles M. Ruch, and Lewis M. Terman in 1923 was one of the first achievement test batteries for multiple grades. Another early test battery was the Iowa Every Pupil Examination for elementary and middle school students, which was developed in 1929 by Everett F. Lindquist and later became the Iowa Tests of Basic Skills. In 1945, Lindquist developed the Iowa Tests of Educational Development for high school learners. These tests were intended to assess students' achievement and to help teachers improve their quality of teaching.

The development of the National Assessment of Educational Progress, first launched in the early 1960s, was a landmark in the history of achievement tests. It is the largest nationally representative assessment designed to assess and monitor what American students know and can do in core subjects. The Elementary and Secondary Education Act of 1965 (ESEA) was enacted to offer equitable educational opportunities to disadvantaged students in the United States. Title I of the law, which provides assistance to school districts for the education of low-income students, has evolved over time and influenced education reforms and testing throughout K–12 education.

In the 1960s and 1970s, the use of standardized achievement tests to meet the ESEA assessment requirements developed incrementally. Standardized tests are administered under conditions that are consistent for all test takers and test scores are norm-referenced, reporting student performance in relation to others from the same population.

The desire to ensure that individual students reached an acceptable minimal level of proficiency resulted in the growth of state-mandated, minimum-competency testing programs throughout the 1970s and has continued to contribute to the spread of standardized achievement tests since then. State-mandated tests are tests and other assessments that the law requires to be administered to all students at designated grade level(s). This wave further triggered the use of criterion-referenced score interpretation for achievement tests.

Beginning in the 1980s, the education reform movement shifted the focus from minimum competency to the expectation of high levels of performance from all students. For achievement tests, this change resulted in a shift from an emphasis on knowledge of basic facts to a focus on more sophisticated reasoning and higher order thinking skills. It also led to changes in item format so that there was some movement away from a reliance on multiple-choice items toward increased use of performance assessments.

The reauthorization and renaming of the ESEA with the passage of the No Child Left Behind Act of 2001 forced state and local educational agencies to be accountable for student achievement and progress. The No Child Left Behind Act, like the original ESEA, aimed to improve the educational experience of disadvantaged populations, and it dramatically expanded the role of state-mandated, standardized achievement tests. Efforts to develop rigorous standards and assessments aligned to them were not new, but the movement toward standards-based testing accelerated after the passage of the No Child Left standards-based testing accelerated after the passage of the No Child Left Behind Act; high levels of student achievement and academic institution accountability were both expected.

Scores on achievement tests can be interpreted using norms, criteria, and/or standards. Standards-based testing has dominated state-level testing since the early 21st century and will likely continue its popularity and influence under new education reforms.

Essentials of Achievement Tests

Summative and Formative Purposes

Achievement tests may be incorporated into the learning process and instructional materials at different times. For summative purposes, testing is done at the end of the instructional process. The test results are viewed as the summation of all knowledge or skills acquired by test takers during a particular subject unit. Judgments about the quality or worth of test takers' achievement are made after the instructional process is completed.

For formative purposes, testing occurs constantly during the learning process so that teachers can evaluate the effectiveness of teaching methods and assess students' performance at the same time. The test results are used to improve teachers' teaching and to help guide students' next learning steps. Judgments about the quality of students' achievement are obtained while the students are still in the process of learning.

Types of Tests

Achievement tests can take either the form of a single subject assessment that focuses on achievement in a single area or the form of a survey battery that typically consists of a group of subject area tests designed for particular grade levels. In a classroom setting, teachers can use tests associated with textbooks as part of their formative and summative assessments to diagnose students' problems and to measure students' mastery. For admission purposes, achievement tests can offer a uniform measure of college readiness such that colleges can identify promising students who are deserving of admission. Furthermore, they provide admissions officers a means to distinguish between well-and poorly prepared applicants.

Survey or Test Battery

Achievement tests are not necessarily standardized tests. In the United States, however, achievement tests produced by test publishers in the form of a survey or a test battery, or the state-mandated tests that not only measure student achievement in K–12 for making instructional decisions but are also known to report public accountability, usually have a high degree of standardization.

Single-level standardized tests for one course or subject, sometimes called surveys, are developed for assessing achievement at only one education level or for one course (e.g., geometry). Usually they are stand-alone tests and are not associated with tests for other courses. For example, the California Standards Tests for Science, available for Grades 5, 8, and 10, are used to assess students' achievement against California's academic content standards in science.

Test batteries or survey batteries contain different tests that assess several curricular areas. There are often multiple levels, indicating that the test content spans several grade levels. For example, the Iowa Assessments are designed for students in kindergarten through the 12th grade. The tests are written for multiple grade levels, with each test level consisting of a series of content areas designed to measure specific skills. The TerraNova by CTB/McGraw-Hill and Data Recognition Corporation is a series of standardized achievement tests designed to assess K–12 student achievement in reading, language arts, mathematics, science, social studies, vocabulary, spelling, and other areas. The ACT Aspire, another example, can be modular or a battery, which provides a means to measure students' learning outcomes in English, mathematics, reading, science, and writing from Grade 3 through early high school.

Often, school districts use a standardized achievement battery to acquire supplementary information useful in curriculum and lesson planning. Achievement tests can serve diagnostic purposes—teachers may use the results of a single test or test battery to suggest areas for individual student development.

Classroom Use

Teacher-made tests and textbook or curricular accompaniments are also achievement tests. Teachers can craft tests to measure the specific learning goals the curriculum framework emphasizes, and the test content can be derived from the course syllabus, the class objectives, or textbook. Also, in teacher’s editions, there are usually tests at the end of textbook chapters, at the back of the book, built into instructional materials, or supplied separately with textbook series. These nonstandardized tests are designed to measure students’ mastery of a specific learning domain such that teachers can learn information about the skills of individual students that are most and least developed, followed by decision-making on the competency, placement, and/or advancement of the students; diagnostic purposes of achievement tests are fulfilled in this way.

Classroom achievement tests are often considered to be criterion-referenced since a student’s scores are compared against some standard, such as the learning objectives for a book chapter, rather than compared with the score of other students in the class. These tests are helpful for teachers to make timely instructional decisions because the turnaround time is controlled by the teacher.

Higher Education Admission

In addition to serving diagnostic purposes, achievement tests can also be prognostic tests used to predict achievement or future performance in a particular area or at a specific time. For example, the ACT and SAT are globally recognized college admission tests. The ACT consists of subject area tests designed to measure academic achievement in English, mathematics, reading, science, and writing. The SAT also tests students’ knowledge and skills learned in school, including reading, writing, and mathematics. For admissions officers, these two college admissions tests provide a predictive tool to distinguish between applicants who are likely to perform well or poorly in college.

College admissions tests are prevalent all over the world. Certainly, the college admissions tests from various countries differ in many ways and they evolve over time, but one constant characteristic of these tests remains: They are high stakes, competitive, and stressful.

Development and Administration

The use of achievement tests involves several major steps, including item development and test assembly, administration, scoring and score interpretation, and reporting. The methods used to design achievement tests must address constructs to be measured in terms of knowledge, skills, and cognitive processes constructs to be measured in terms of knowledge, skills, and cognitive processes. Item writers are content experts who usually begin with a list of content standards that specify what students are expected to know and learn in a given grade level. The number and type of test items can be determined by the grade-level content standards.

Item Types

Typically on a paper-and-pencil test, the item types used on achievement tests include multiple-choice items, true-false questions, short-answer open-ended items, and essay questions. Due to advancements in technology, test delivery can be done not only by paper but via computer and other electronic devices. For example, technology-enhanced items are computer-delivered items to which test takers respond based on interactions such as dragging and dropping, editing, highlighting, ordering, and sorting.

The choice among various item types is typically made on the assumption that some particular knowledge, skill, ability, or mental process can be measured by each of these item types. Whatever educational achievement that can be measured well by one type of test item can probably also be measured quite well by some other types. In practice, choosing among item types often takes into account development costs as well as testing and scoring time, which often works in favor of objective test items (e.g., multiple-choice items).

Administration

Most achievement tests, especially in education, are administered in group settings. For high-stakes or mandatory achievement tests, standardization is required such that the testing conditions are the same for all test takers; it is also common to establish norm-based score scales for interpreting test performance against a representative sample of individuals from the population with which the test is intended to be used. Individuals approved for test accommodations may be provided with a specialized version of a test, such as large type print, braille, audio, or Spanish language, or given an extended time to take a test.

Ideally, test takers should understand what the test requires them to do, attain an environment in which they can motivate themselves to do as well as they can, and have an equal opportunity to demonstrate their best efforts to achieve good performance. It is also important that achievement tests avoid being unduly speeded. Most test takers should have enough time to complete the test, which often enables the best performance on the tests and the most accurate predictions of subsequent achievement.

Scoring

In terms of scoring, each item type presents unique methods and problems for scoring. For teacher-crafted tests, answers to multiple-choice and true-false questions as well as other objective-item types can be marked directly on the test copy and later be scored by hand. For state-level or large-scale tests, scoring can be facilitated if the answers are provided by marking on a separate answer sheet such that they can be scored more quickly and accurately by electrical scoring machines. For open-ended items (e.g., short-answer questions) and essay questions, scoring rubrics are developed for the questions and used to train human raters or to program computer scoring algorithms. For short-answer questions, a scoring key that shows the kinds of answers eligible for full credit or partial credit is often recommended.

Essay scoring rubrics can be analytic or holistic. Irrelevant factors, such as the quality of handwriting, verbal fluency, and rater interests or biases should be avoided in the scoring process. In general, essay scoring takes considerably more time and is much more costly than the scoring of objective items.

Score Interpretation and Reporting

The meaning and interpretation of achievement test scores can be relative, absolute, or both. A norm-reference framework, which interprets test scores in a relative sense, indicates how the achievement of a particular student compares with the achievement of a well-defined group of other test takers (i.e., the norm group) who have taken the same test. Derived scores commonly used for the norm-referenced tests include percentile ranks; linear, normalized, or developmental standard scores; and grade equivalents.

A criterion-reference framework offers an absolute score interpretation by inferring the kinds of performance a student can do in a domain. The results of criterion-referenced testing can be presented by, for instance, the percentage of correct responses, the percentage of objectives mastered, the predefined quality level of student achievement (e.g., “excellent,” “mastery,” rating of “A” or “5”), or the precision of performance. For standards-based referenced scores and interpretation, performance-level descriptions are unique to the achievement test for which the defined levels of performance such as “basic,” “proficient,” and for which they define levels of performance such as “basic,” “proficient,” and “advanced.” A certain range of test scores is carefully associated with each of the achievement levels in a subject; the percentage of test takers at each level of proficiency is of the most interest.

Finally, score reporting typically contains at least three elements: types of scores provided on score reports, other information provided on or with the score reports (e.g., interpretive guides for students, parents, teachers, or principals), and other supporting information that may be available (e.g., technical reports). Accurate, efficient scoring and reporting makes the test score interpretation clearly communicated and strongly supportable and provides more directly useful information to guide instructional decisions and promote learning.

Benefits and Limitations

Achievement tests are standardized or nonstandardized tests used to measure acquired learning. A well-constructed test yields valid and reliable results, providing test takers with an opportunity to demonstrate what they have learned in school and to show to themselves and others the knowledge and skills that they have accumulated. The development of test specifications needs to be aligned with curriculum or professional content standards in a clear and coherent way. Following the specifications, items, questions, or tasks on the tests can then present the targeted procedural knowledge and cognitive processes.

Rigorous test administration helps with the interpretation of scores. The results of achievement tests serve as indicators of examinee progress. Score results can help examinees confirm their strengths and weaknesses. Also, a well-written, valid test equipped with efficient scoring and reporting will give teachers valuable information regarding students’ needs and abilities, offer teachers a useful measure of how well the students have achieved the course objectives, and assist teachers in evaluating teaching effectiveness. As outlined, achievement tests are also one of the many useful tools to predict college performance.

Concern has arisen over the increased use of standardized tests in schools, with some arguing that test takers may become anxious and frustrated from taking and preparing for the tests, which may subsequently lower their motivation to learn. Whether the content coverage and relevance as well as the construct representativeness of the tests are solid and sound is sometime a concern to the public. In addition, if items do not present a challenge to the test takers, the test may become meaningless since test users might question whether the test items adequately represent the content domain and the quality of performance such that we can assure successful performers on the tests actually meet the content standards.

However, even if the test is well written and content valid, the degree to which the achievement measure is authentic is sometimes doubted. Authenticity is limited due to the cost of test construction and the acceptable administration time. Items on achievement tests are often expressed in verbal or symbolic terms. Meanwhile, the knowledge obtained by direct perceptions of objects, events, feelings, or relationships, as well as mental and behavioral skills, such as leadership and friendship, are not assessed by most achievement tests.

Also, achievement tests may be able to measure what a person knows but not necessarily how effectively he or she uses that knowledge in practice. In addition, items do not necessarily reflect the actual, full picture of the learning outcomes from a classroom setting. It is not possible for an achievement measure to cover all knowledge and skills or to represent the whole of human achievement. A student with higher achievement scores is more likely to succeed than another student with low achievement scores, but high scores cannot guarantee future success.

Finally, one of the intended purposes for using K–12 standardized achievement tests is to provide information for public accountability. In some schools, test results are also used to evaluate teachers, although there is ongoing debate about the legitimacy of using standardized tests for this purpose. In any case, it is clear that standardized achievement test results should not become the single most important indicator of school performance or teacher evaluation. An overemphasis on test scores can result in pressure that can potentially lead to cheating by administers and teachers, which invalidates the whole idea of achievement testing.

Yi-Fang Wu

See also Ability Tests; ACT; Criterion-Referenced Interpretation; Formative Assessment; Norm-Referenced Interpretation; SAT; Standardized Tests; Standards-Based Assessment; Summative Assessment

Further Readings

Further Readings

AERA, APA, … NCME. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brookhart, S. M., & Nitko, A. J. (2014). Educational assessment of students (7th ed.). Upper Saddle River, NJ: Pearson.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 579–621). Westport, CT: American Council on Education/Praeger.

Haladyna, T. M., & Downing, S. M. (Eds.). (2006). Handbook of test development. Mahwah, NJ: Erlbaum.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Monroe, W. S., DeVoss, J. C., & Kelly, F. J. (1917). Educational tests and measurements. New York, NY: Houghton Mifflin.

Nitko, A. J. (1983). Educational tests and measurement: An introduction. New York, NY: Harcourt Brace Jovanovich.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R.

Garner (Eds.), Ability testing: Uses, consequences, and controversies, Part II (pp. 173–194). Washington, DC: National Academy Press.

The College Board. (2016). The SAT®. Retrieved from https://sat.collegeboard.org/home

Thomas, J. Y., & Brady, K. P. (2005). Chapter 3: The Elementary and Secondary Education Act at 40: Equity, accountability, and the evolving federal role in public education. Review of Research in Education, 29, 51–67. doi:10.3102/0091732X029001051

Whipple, G. M. (Ed.). (1918). The seventeenth yearbook of the National Society for the Study of Education, Part II: The measurement of educational products. Bloomington, IL: Public School Publishing Company.

Zhongmin Cui Zhongmin Cui Cui, Zhongmin

ACT

ACT

33

36

ACT

The American College Testing Program (ACT) is a curriculum-and standards-based educational and career planning tool assessing students’ academic readiness for college. The ACT comprises five tests, including four subject tests (English, mathematics, reading, and science) and an optional writing test. Depending on whether the optional writing test is taken, the total testing time is either 2 hours and 55 minutes or 3 hours and 35 minutes. Each of the five tests is scored on a scale from 1 to 36. A composite score of the four nonwriting subjects is also based on a scale from 1 to 36.

The ACT is created and administered by a nonprofit company, ACT, Inc. (formerly known as American College Testing). ACT scores are accepted by all 4-year colleges and universities in the United States. Students can also take the ACT overseas, and it is administered multiple times each year, both inside and outside of the United States. This entry discusses the history of the test, its components, methods of preparing for the test, and how the test is used.

History

On November 7, 1959, the first-ever ACT was taken by 75,460 high school students looking forward to joining college. Although another college admissions test (the SAT, then known as the Scholastic Aptitude Test) did exist at that time, the ACT was different because it was a test of achievement and did not purport to measure innate intelligence or intelligence quotient. Being unsatisfied with the existing system of admissions testing, E. F. Lindquist and Ted McCarrel cofounded the ACT (now ACT, Inc.) and created the first college admission test based on information taught in schools.

Since its inception, the ACT has grown rapidly. Since 1960, ACT has been taken in all 50 states. In 2012, for the first time, the number of students taking the ACT surpassed the number of students taking the SAT. In 2012, over half of the country’s high school graduates took the ACT. Part of the growth can be attributed to statewide administrations of the ACT. In 2001, Colorado and Illinois became the first states to adopt the ACT as part of their statewide assessment programs to measure students’ progress toward meeting state learning standards. Other states soon followed. For the 2014–2015 school year, the ACT was administered as part of a state assessment to students in 21 U.S. states.

In spring 2013, ACT announced enhancements to the ACT based on evidence from the ACT National Curriculum Survey and to reflect changes in the education market. These enhancements are as follows: (a) the online administration of the ACT, (b) the addition of questions on the reading test, addressing whether students can integrate knowledge and ideas across multiple texts, (c) the inclusion of additional statistics and probability items in the mathematics test for reporting of student achievement in this area, and (d) additional scores and indicators (STEM score, progress toward career readiness indicator, English language arts score, and understanding complex texts indicator). However, the 1–36 score scale remains. In fall 2015, enhancements were made to the design of the writing test and new writing scores were introduced. The score scale has been changed from 2–12 to 1–36. Instead of one holistic score, students receive four analytic scores (also known as domain scores) which are used to compute the writing score on the 1–36 scale.

Test Description

The ACT consists of four multiple-choice tests: English, mathematics, reading, and science. The ACT with writing includes the four multiple-choice tests and a writing test. All multiple-choice items have four choices except for those on the mathematics test which have five choices. Each item will have only one best answer (i.e., the correct answer). Students score one point by choosing the correct answer, with no penalty for incorrect answers.

The English test measures standard written English and rhetorical skills. The test consists of five essays or passages, each of which is accompanied by a sequence of multiple-choice test questions. It comprises 75 questions, and examinees have 45 minutes to finish them. Students' performances on the English test are reported on a scale from 1 to 36. In addition to the total test score, two subscores (usage/mechanics and rhetorical skills) are provided.

The mathematics test measures mathematical skills students have typically acquired in courses taken up to the beginning of Grade 12. It consists of 60 questions and examinees have 60 minutes to finish them. Students' performances on the mathematics test are reported on a scale from 1 to 36. In addition to the total test score, three subscores (pre-algebra/elementary algebra, intermediate algebra/coordinate geometry, and plane geometry/trigonometry) are provided.

The reading test measures reading comprehension. It consists of 40 questions, and examinees have 35 minutes to finish them. Students' performances on the reading test are reported on a scale from 1 to 36. In addition to the total test score, two subscores (social studies/natural sciences and arts/literature) are provided.

The science test measures the interpretation, analysis, evaluation, reasoning, and problem-solving skills required in the natural sciences. It consists of 40 questions, and examinees have 35 minutes to finish them. Students' performances on the science test are reported on a scale from 1–36. There are no subscores for the science test.

Starting from fall 2016, reporting category scores are also provided to students. The reporting categories for the English test include production of writing, knowledge of language, and conventions of standard English. The reporting categories for the mathematics test include preparing for higher math, number and quantity, algebra, functions, geometry, statistics and probability, integrating essential skills, and modeling. The reporting categories for the reading test include key ideas and details, craft and structure, and integration of knowledge and ideas. For the science test, the reporting categories include interpretation of data, scientific investigation, evaluation of models, and inferences and experimental results.

The optional writing test measures writing skills emphasized in high school English classes and entry-level college composition courses. It consists of one prompt that typically presents conversations around contemporary issues and offers three diverse perspectives that encourage critical engagement with the issue. Students have 40 minutes to develop and compose an argument that puts their perspective in dialogue with others. Student writing is evaluated in four domains: ideas and analysis, development and support, organization, and language use. The four domain scores are used to compute the subject-level writing score that, like other subject scores, is on a scale from 1 to 36.

Preparing for the ACT

To help students prepare for the ACT, ACT, Inc. publishes an official prep book, *The Real ACT Prep Guide*. In addition to information on how to register and prepare for the test day, this book includes five practice tests, each with an optional writing test, which were used in previous actual test administrations. For all multiple-choice items, this book explains why an answer choice is right or wrong. For the writing prompt, the book explains how it is scored. A review of important topics in English, mathematics, science, and writing is also included.

ACT, Inc. also publishes an electronic tool to help students prepare for the ACT, ACT Online Prep, available on both desktop computers and tablet computers. It is not an electronic copy of the official prep book, but an interactive tool to help students become familiar with the ACT, to know their strengths and weaknesses, and to improve. Using ACT Online Prep, students can take a short-form ACT to get a predicted score range. Based on students’ performance on the short-form ACT and their unique needs (e.g., the available preparation time before the test), the system can create a personalized learning path to guide the students through a library of learning content in the most efficient way possible. Students have different ways to learn, including flash cards, lessons, and practice questions. A dashboard is available for students to keep track of their progress and to get feedback on their strengths and weakness and how to improve. This tool also includes a full-length ACT, which was used in previous actual test administrations, to help students get familiar with the test and predict their performance.

Both *The Real ACT Prep Guide* and ACT Online Prep are available for purchase, though ACT provides low-income students with free access to its ACT Online Prep program. A free copy of *Preparing for the ACT* is available for download at the ACT website. This document includes test information, a complete practice test, and a sample writing prompt. It allows students to become familiar with the test before turning to the two more extensive test preparation tools.

There are many other test preparation companies that sell test preparation materials and offer preparation courses or training opportunities for students to prepare for the ACT (e.g., Kaplan, Inc.; Princeton Review). In addition to the ACT, these companies typically offer preparation services for various other tests. Although they typically claim coaching courses help increase their scores, there is no solid support for such claims from research.

ACT conducted several studies between the early 1990s and 2003 to examine ACT score increases attributable solely to short-term test preparation activities (e.g., commercial test preparation courses, commercial workbooks, test preparation computer software, test preparation workshops offered by local schools) using repeat test takers and cross-sectional samples of students who took the test at given time points. The results from these studies show that short-term test preparation activities have a relatively small positive impact on the ACT composite score when compared to long-term activities. The best preparation for the ACT is to take a rigorous core curriculum in high school.

Using the ACT

The ACT is designed to give students an indication of how likely they are to be ready for college-level work. ACT suggests that these scores or higher scores indicate readiness for college: English, 18; math, 22; reading, 21; and science, 24. Mean scores on the ACTs are about 20, with standard deviations of about 4½. Research has shown that students with higher ACT scores tend to be better prepared for college-level work as shown by higher first-year GPAs in college.

Like any other achievement test, the ACT neither measures everything students have learned in high school nor measures everything necessary for students to know to be successful in their next level learning. Such a test simply does not exist, so admissions decisions should not be made solely based on a single test. The reported scale score for an examinee is only an estimate of that examinee’s true score because of some measurement error. The ACT demonstrate very high reliability, however, and observed scores are considered very close to students’ “true” scores.

The ACT can be used for numerous and diverse purposes. Distinct validity evidence, however, is needed for each intended use, according to the *Standards for Educational and Psychological Testing* of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. The most common uses, according to the *Technical Manual* published by ACT, are to measure educational achievement in particular subjects, make college admission and college course placement decisions, evaluate the effectiveness of high schools in preparing students for college, and evaluate students' likelihood of success in the first year of college and beyond. For usage not covered by ACT's *Technical Manual*, it is advised that users support their usage by validity arguments.

Zhongmin Cui

See also Achievement Tests; College Success; SAT; *Standards for Educational and Psychological Testing*

Further Readings

ACT, Inc. (n.d.). News and FAQs. Retrieved from http://www.act.org/actnext/

ACT, Inc. (n.d.). Our products. Retrieved from http://www.act.org/content/act/en/products-and-services.html

ACT, Inc. (2016). Preparing for the ACT test. Retrieved from https://www.act.org/aap/pdf/Preparing-for-the-ACT.pdf

ACT, Inc. (n.d.). The Act Test for students. Retrieved from http://www.actstudent.org/

ACT, Inc. (2012). What kind of test preparation is best? Retrieved from https://www.act.org/research/policymakers/pdf/best_testprep.pdf

ACT, Inc. (2014). Technical manual. Retrieved from https://www.act.org/research/policymakers/pdf/best_testprep.pdf

ACT, Inc. (2015). Using your ACT results. Retrieved from http://www.act.org/aap/pdf/Using-Your-ACT-Results.pdf

American Educational Research Association, American Psychological Association, … National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC.

Strauss, V. (2012). Why ACT overtook SAT as top college entrance exam. Retrieved from https://www.washingtonpost.com/blogs/answer-sheet/post/how-act-overtook-sat-as-the-top-college-entrance-exam/2012/09/24/d56df11c-0674-11e2-afff-d6c7f20a83bf_blog.html

Rebecca H. Woodland Rebecca H. Woodland Woodland, Rebecca H.

Action Research

Action research

36

38

Action Research

Action research is a form of reflective inquiry with intended use and users; practitioners in their own educational settings conduct it in order to improve their own professional practice and outcomes for students. Through action research, educators identify pressing problems of professional practice and engage in data collection, analysis, and interpretation, with the intention of understanding gaps between desired and actual results and achieving genuine improvements in the quality of their instruction and student learning. The action research process and resultant findings can equip practitioners with the knowledge they need to make real-time, evidence-based decisions about schooling, teaching, and learning. This entry discusses the development of action research as a method of inquiry, how and why action research is used in PreK–16 schools and barriers to the action research use in these schools.

Educators engage in action research in order to continuously test their working theories as to what works and what doesn’t in schools and classrooms. It is the individuals’ immediate use of data to inform and/or improve their practice that most distinguishes “action” research from more traditional academic research where educators may have little input into study design, data collection, or interpretation. This method of inquiry is most common among professionals in PreK–16 educational settings, including regular and special education teachers, school principals, district superintendents, instructional coaches, school counselors, and school psychologists.

Action research is arguably the most valid, powerful, and important tool that professionals in PreK–16 settings have at their disposal to make meaningful, ongoing, and sustained positive changes to their practice intended to bring about ongoing, and sustained positive changes to their practice intended to bring about essential outcomes. Action researchers can acquire greater congruity between the values they espouse and the values they enact in practice in classrooms and across systems. Action research is ideal for those who hold participatory, democratic, and improvement-oriented worldviews concerned with the development of rich, valid, contextually useful information for the improvement of teaching, learning, and schooling.

Background

The term *action research* was first used in the early 20th century to characterize group research activities that resulted in changed community practices. A theoretical framework for action research emerged from Kurt Lewin and his studies of the workplace in the 1930s. He conceived of action research as an ongoing process of thinking and doing by organizational stakeholders, bringing about increases in employee morale and their work ownership. In time, the principles of action research began to be integrated into the examination of pedagogical and educational reform activities. Today, action research stands in marked contrast to traditional educational research (or “pure” research), in which an outside investigator examines an issue, disseminates findings (perhaps through publication in peer-reviewed journals), and then leaves it up to practitioners to locate, access, interpret, and implement the results.

Action research is increasingly recognized as a legitimate form of social science research. One such measure of legitimacy are peer-reviewed venues for publication on the topic. *Action Research*, an international, interdisciplinary, peer-reviewed journal that publishes articles on the theory and practice of action research, was launched in 2003. The journal publishes accounts of action research projects and articles that explore the philosophical underpinnings and methodology of action research. Action research, as a job-embedded process, stimulates the ingenuity of educators and cultivates their ability to creatively and collaboratively address immediate problems of practice. Educators who conduct action research experience intellectual and professional growth, develop more positive attitudes toward their colleagues, and improve their pedagogical skills.

Action Research for Informed Decision Making

Educators make hundreds of complex decisions every day, the most important of which are made in a quick and intuitive fashion during the act of instruction.

Given the high-stakes and multifaceted nature of learning environments, whose outcomes have immediate and long-standing ethical implications for individuals and society, educators cannot afford to practice their craft in an unexamined fashion. Decisions about what to teach and how to go about the business of teaching and learning are too often based on recollections of events, anecdotal information, ideas found through happenstance, and casual observations.

Through action research, school-based professionals can examine and interpret the learning environment in order to make informed decisions about curriculum, instruction, and assessment. It is the systematic collection and analysis of a variety of contextualized classroom-and school-based data, including observations, student artifacts, interviews, journal entries, formative assessment results, and video of teaching, considered in light of established theory that helps to transform typical and less rigorous forms of reflective practice into action research.

Cycle of Action Research

In a cycle of action research, an educator or team of educators will (a) identify a theory or argument about what is important and makes a difference in student learning; (b) formulate specific questions about teaching and learning related to their theory or argument; (c) identify and define the variables, terms, and concepts at the heart of their questions; (d) understand the already available key literature or studies that shed light on the questions of interest; (e) develop a hypothesis or supposition about what their studies’ findings might reveal; (f) collect and analyze data about the variables, terms, and concepts in their research questions; (g) interpret the results and make decisions about what or how to change or improve practice; (h) revisit, revise, and refine their original theory or argument; and (i) repeat a–h as an ongoing part of their regular professional work.

Action Research as Professional Development

Action researchers are working professionals who use applied social science data collection and analysis methods to explore and test new ideas, methods, and materials and assess the effectiveness of curricular approaches. Action research is most effective when it is conceived of as a regular and routine part of their professional practice, that is, when educators initiate and facilitate systematic inquiry as part of their teaching and administrative responsibilities. When educators undertake their own places of work, action research can become an exceptional vehicle for job-embedded professional development. School leaders can support action research activities by reserving space and time to enable educators to jointly carry out the cycle of action research. In addition, school leaders can incorporate the process and results of teacher action research into the more formal systems of supervision and evaluation. Teachers can use documentation of their individual or team-level cycle of action research as the mechanism for setting, monitoring, and reporting their instructional practice and student learning goals required by state and local educator evaluation systems.

Barriers to Action Research

If action research is one of the most effective means through which teachers and administrators can improve their practice, student achievement, and schooling, why is engagement in action research infrequent in schools? Although action research is an effective strategy for a continuous organizational and pedagogical improvement, there are few powerful federal or state-level policy proponents or legislative mandates that support it. As a result, educators do not typically have the resources or the impetus they need to carry out action research studies.

A significant resource in short supply is time. A significant amount of time is needed in order for educators to use action research to improve instruction and enhance student learning. Although there is no rule for how much time is needed for action research to be productive, studies suggest that any professional development endeavor in which teachers are engaged for less than an average of 8 hours *per month* will likely have little or no impact on instructional practice and student learning. In addition to a lack of resources such as time, educators may also lack the skills necessary for conducting high-quality, quantitative and qualitative data collection and analysis at the core of all social science research methods.

Rebecca H. Woodland

See also Applied Research; Conceptual Framework; Mixed Methods Research; Professional Development of Teachers; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Further Readings

Mills, G. (2014). Action research: A guide for the teacher researcher (4th ed.). Pearson.

Parson, R., & Brown, K. (2002). Teacher as reflective practitioner and action researcher. Belmont, CA: Wadsworth.

Sagor, R. (2000). Guiding school improvement with action research. Alexandria, VA: ASCD.

Wei, R. C., Darling-Hammond, L., & Adamson, F. (2010). Professional development in the United States: Trends and challenges. Dallas, TX: National Staff Development Council.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues … Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs

Lina Lee Lina Lee Lee, Lina

Active Learning Active learning

38

40

Active Learning

This entry describes active learning, addresses its benefits and challenges, and offers strategies for implementing active learning in classroom settings. Active learning shifts the focus of learning from passively receiving content information to diligently participating in learning activities. Student engagement in learning increases retention and understanding of course content and enhances the quality of learning outcomes. In active learning, with the guidance and assistance of the teacher, students learn and practice new concepts and use them meaningfully.

Although there are many definitions of active learning, it can be described as a student-centered approach to instruction. According to Charles Bonwell and James Eison, active learning is “anything that involves students in doing things and thinking about the things they are doing” (p. 2). The key elements of active learning are student involvement in the learning process and critical reflection on course material. Unlike the teacher-centered approach where students simply listen to lectures and take notes, in active learning, students engage with the course material, participate in the class, and collaborate with others. The process affords students the opportunity to explore and develop new concepts through meaningful discussions and problem-solving situations.

In active learning, teachers must shift their roles from “sage on the stage” to “guide on the side.” They are no longer information providers; rather, they are facilitators helping students understand a concept, demonstrate it, and apply it in the real-world situations. In active learning, students become autonomous and self-directed learners taking charge of their own learning by taking initiative, monitoring progress, and evaluating learning outcomes. Consequently, students not only develop knowledge and skills, they also show high motivation and good attitudes toward learning.

In today’s classrooms, there is increasing emphasis on equipping students with 21st-century skills, including critical thinking, creativity, communication, and collaboration. Critical thinking is often promoted through higher order thinking that requires students to use cognitive skills to understand, synthesize, evaluate, and make use of information to create content. Critical thinking helps students gain control of their own learning and make better informed decisions as to what, when, and how to learn. Furthermore, active learning promotes social interactions, allowing students to work collaboratively with their peers and teachers. Increased peer-to-peer and student-to-teacher interaction helps to build a learning community through which students develop, share, and exchange perspectives.

Challenges of Using Active Learning

Although active engagement empowers students to create their own learning experiences and is believed to enhance the quality of learning, both students and teachers perceive challenges. Some students may not be willing to abandon their passive roles of listening to lectures. Students may not have skills required, such as learning strategies and critical thinking, to participate in active learning. Class of large sizes can prevent teachers from implementing active learning due to limited class time. Teachers are preoccupied by not being able to cover the amount of course material or feeling a loss of control. They also fear that students may resist active learning. Other barriers include a lack of needed materials, equipment, or resources. The challenges of using active learning can be overcome by offering teachers effective strategies and techniques.

Strategies for Implementing Active Learning in Teaching

To effectively use active learning, teachers first need to openly communicate with students about their instructional goals and strategies. A common instructional strategy for active learning is to integrate student-centered activities into the traditional lecture. To maintain student attention span during the lecture, a combination of instructional techniques can be used, such as open-ended questions, small group discussions, and reflective responses. At the end of the lecture, students are asked to answer teacher-made questions called “minute paper,” allowing them to reflect on that day’s course material.

Several techniques are considered effective for active learning, including collaborative learning, problem-based learning, project-based learning, and technology-based learning. Collaborative learning activities allow students to work together with others to achieve a common goal, whereas problem-based learning, a student-centered approach, enables students to gain knowledge and skills through the experience of solving difficult and complex problems. Problem-based learning requires critical thinking, self-regulation, and self-motivation on the part of students.

Another way to embrace active learning is by using flipped learning. Due to the increasing availability of digital technology, teachers can easily prepare short video lectures for students to view and learn course material, at home before the class session. Flipped learning emphasizes students' learning responsibility. It allows teachers to free up class time to explore the challenging aspects of course content and engage students with the content, using various types of active learning activities such as open-ended discussions in pairs or small groups. In the flipped classroom, teachers provide personalized learning and meet individual student needs. Through active engagement with anytime and anywhere access to video lectures, students learn at their own pace to master the concept.

Lina Lee

See also Constructivist Approach; Instructional Theory; Mastery Learning; Social Learning

Further Readings

Barkley, E. (2010). Student engagement techniques. San Francisco, CA: Jossey-Bass.

Bonwell, C. C., & Eison, J. A. (1991). Active learning: Creating excitement in the classroom (ASHE-ERIC Higher Education Report No. 1). Washington, DC: The George Washington University, School of Education and Human Development.

Carr, R., Palmer, S., & Hagel, P. (2015). Active learning: The importance of developing a comprehensive measure. Active Learning in Higher Education,

16(3), 173–186.

Mayer, R. E. (2008). Learning and instruction (2nd ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Wenger, E. (1998). Communities of practice: Learning, meaning and identity. New York, NY: Cambridge University Press.

ADA

See Americans with Disabilities Act

Alan W. Brue Alan W. Brue Brue, Alan W.

Linda Wilmshurst Linda Wilmshurst Wilmshurst, Linda

Adaptive Behavior Assessments

Adaptive behavior assessments

40

44

Adaptive Behavior Assessments

Adaptive behavior refers to a group of basic skills that people must master in order to function and survive. These skills are conceptual, social, and practical skills used in daily life. Assessment of adaptive behavior skills is necessary as a component of the diagnosis or classification for having an intellectual disability.

People with an intellectual disability typically have significant deficits in their conceptual, social, and/or practical skills. These deficits can prevent them from being fully independent. Adaptive behavior measures can be used to help determine the level of impairment. This entry first looks at how the criteria for diagnosing intellectual disabilities have changed and now include deficits in adaptive functioning. It then describes the two main rating scales used to assess adaptive behavior skills, the Adaptive Behavior Assessment System (ABAS) and the Vineland Adaptive Behavior Scales.

Intellectual Disabilities and Adaptive Behavior Skills

A deficit in adaptive behavior skills has not always been a part of assessment for intellectual disabilities. When the American Psychiatric Association first published the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) in 1952, the classification category of *mental deficiency* was introduced to account for cases that were primarily a defect of intelligence present at birth with no known organic brain disease or known prenatal cause for the deficits. Cases were to include only individuals with familial or idiopathic (unknown origin) mental deficiencies, and severity was to be determined by IQ scores in the following three ranges: *mild* (an IQ of approximately 70–85), *moderate* (IQ 50–70), and *severe* (IQs below 50). Although IQ scores were necessary to determine the range and expectations, the *DSM* noted the importance of considering other factors.

When the second edition of *DSM* (*DSM-II*) was first published in 1968, the term *mental retardation* (MR) replaced *mentally deficient*. The *DSM-II* better aligned with what was then called the American Association on Mental Retardation (now the American Association on Intellectual and Developmental Disabilities) and supported five ranges of severity (borderline, mild, moderate, severe, and profound), with the borderline range for IQ scores in the 68–85 range. It listed clinical codes for 9 subcategories for the disorder, based on the circumstances of origin (e.g., following infection and intoxication; following trauma or physical agent).

In 1980, the *DSM-III* placed MR in a new section titled “Disorders Usually First Evident in Infancy, Childhood or Adolescence.” The three main criteria for a diagnosis of MR remained consistent with the previous version (i.e., impaired IQ, impaired adaptive behaviors, and onset during the developmental period); however, these criteria were further refined at this time. Subnormal intelligence was now set two standard deviations below the mean (IQ of 70), instead of one standard deviation (IQ of 85) with the addition of a five-point interval to be considered (IQ 65–75) to account for the standard error of measure. Onset during the developmental period was defined as occurring below 18 years of age. Impairments in adaptive functioning were required; however, the *DSM* noted that the then-current measures were not considered valid to be used in isolation to make this decision and recommended that clinical judgment should evaluate adaptive functioning in individuals relative to similar aged peers.

In the *DSM-IV*, first published in 1994, the three criteria for diagnosing MR were retained from the previous version. The criterion of adaptive functioning was further defined as requiring deficits in two of 10 possible areas: (1) functional academic skills, (2) social/interpersonal skills, (3) communication skills, (4) self-care, (5) home living, (6) use of community resources, (7) self-direction, (8) work, (9) leisure, and (10) health/safety. These deficits were determined by an individual’s score on an adaptive measure that was 2 standard deviations (*SD*s) below the norm. These criteria remained consistent in the subsequent text revision of the *DSM-IV-TR* in 2000.

In 2002, the American Association on Mental Retardation made the landmark decision to change the way it defined the severity of MR, moving away from classifying levels based on intellectual functioning to levels of supports needed (*intermittent, limited, extensive,* or *pervasive*) to close the gap between problems in adaptive functioning and enhancing an individual's capabilities. The Supports Intensity Scale was developed to measure the need for supports and includes 49 life activities grouped into six subscales: Home Living, Community Living, Lifelong Learning, Employment, Health and Safety, and Social Activities. In 2008, the American Association on Mental Retardation reported that SIS has a .87 inter-rater reliability coefficient, which the organization said put the scale in an "excellent range" of reliability in assessment instruments. Recent research suggests that proper training in the administration of the SIS increases the reliability of the instrument.

There were many changes in the way that disorders are conceptualized with the publication of the *DSM-5* in 2013. In an attempt to move away from a purely categorical classification system and to incorporate more of a dimensional approach to regarding disorders along a continuum, the *DSM-5* is organized using a developmental framework. The section labeled "Disorders Usually First Diagnosed in Infancy, Childhood, or Adolescence," in the *DSM-IV* was removed and in its place a new section called "Neurodevelopmental Disorders" was added. The term *MR* was replaced by *intellectual disabilities*, also known as *intellectual developmental disorders*, which include categories for global developmental delay (for children under 5 years who demonstrate delays and have not yet been assessed) and unspecified intellectual disability (for cases over 5 years of age where assessment cannot be conducted due to other factors such as severe behavior problems or sensory/motor impairments). The *DSM-5* continues to use specifiers (mild, moderate, severe, and profound) to identify the severity of the disorder; however, unlike previous versions of the *DSM*, the severity no longer is based on IQ scores but now refers to levels of adaptive functioning in the conceptual, social, and pragmatic domains.

Rating Scales

Adaptive behavior rating scales are used to obtain feedback from parents, caretakers, teachers, and employers. It is important to obtain feedback from multiple sources. If a child has two parents, it is customary to ask each parent to complete a rating scale. Professionals have found that parents can differ in their viewpoint of a child's abilities. Parents also sometimes yield scores higher than viewpoint of a child's abilities. Parents also sometimes yield scores higher than those from other sources because parents may overestimate their child’s ability or may not be able to compare their child to a child without adaptive behavior deficits as easily as a teacher can because teachers also spend time with students who do not have delays.

If adaptive scores from parents are inconsistent with a teacher scale and other information gathered, it is often wise to consider following up with an interview. Another parent adaptive rating scale may be used and administered in an interview format. By questioning a parent and providing examples of what the item is asking, feedback may be provided that more readily matches a child’s deficits.

There are two major rating scales used to assess adaptive behavior in children, adolescents, and adults: Adaptive Behavior Assessment System, Third Edition (ABAS-3) and Vineland Adaptive Behavior Scales, Third Edition (Vineland-3).

ABAS

The ABAS-3, published by WPS, is one of the leading adaptive behavior measures. The measure, authored by Patti Harrison and Thomas Oakland, was updated in 2015. The ABAS-3 includes five rating forms: Parent/Primary Caregiver Form (for ages 0–5 years), Teacher/Daycare Provider Form (for ages 2–5 years), Parent Form (for ages 5–21 years), Teacher Form (for ages 5–21 years), and an Adult Form (for ages 16–89 years).

Parents, close family members, teachers, day care staff, supervisors, or others who are familiar with the daily activities of the person being evaluated can complete any of these forms. Eleven adaptive skill areas are assessed by the ABAS-3; either nine or 10 skill areas are included on each form, depending on the age of the person being rated. The three adaptive domains that are addressed are conceptual, social, and practical. In addition, the test provides an overall General Adaptive Composite. All scores are categorized descriptively as extremely low, low, below average, average, above average, or high.

The conceptual composite consists of the following skill areas: communication, functional academics, and self-direction. The communication skill area assesses how well one speaks using appropriate grammar. It also looks at the ability one has in stating information about oneself and how well one converses with others. Functional academics assesses how well one performs the basics in academics in order to function daily at school, home, and the community. Self-direction assesses how well one acts responsibly. For example, this can include completing schoolwork and chores, controlling anger and frustrations appropriately, and making responsible choices in spending money.

The Social Composite Scale consists of information regarding the following skill areas: leisure and social. Leisure includes things one does when not in school or doing chores at home. Examples could include the following: reading a book or putting together a puzzle, playing games with friends, joining in sports activities, and/or joining some type of club. Social involves the ability to make friends and maintain friendships. It also assesses how well one is aware of other people’s feelings and appropriate actions taken in certain situations.

The Practical Composite Scale measures the following skill areas: community use, home living, health and safety, and self-care. Community use assesses how well one functions in the community. For example, this can include using the library and mailing letters at the post office. Home living evaluates how well one is able to do things at home for oneself. Making the bed, preparing food for oneself, and washing one’s dishes are all examples of this skill area. Health and safety is an important skill in that it looks at one’s ability to be healthy and safe in everyday situations. This may include following rules: using caution around a hot stove and seeking help when someone is hurt. Self-care assesses how well one functions independently in taking care of self. One must be able to do everyday things on one’s own, such as dressing oneself, bathing, and using the bathroom.

The ABAS-3 standardization was completed using 7,737 research forms completed by the respondents who rated the adaptive behavior of 4,500 individuals. Sample sizes were 1,420 for the Infant and Preschool Forms, 1,896 for the Parent and Teacher Forms, and 1,184 for the Adult Forms. The sample represented the 2012 United States population in terms of ethnicity, gender, and household education level, and all geographic regions were represented. Compared to the U.S. Census, there was an overrepresentation of White individuals and those with a higher level of education.

Internal consistency, which indicates the degree to which test items correlate with each other and is often treated as an estimate of reliability, is excellent. The α reliability coefficients range for the broad adaptive domains was 0.90–0.98 on the Teacher/Daycare Provider Form, 0.93–0.99 on the Teacher Form, 0.85–0.98 on the Parent/Primary Caregiver Form, 0.94–0.99 on the Parent Form, 0.94–0.99

on the Parent/Primary Caregiver Form, 0.94–0.99 on the Parent Form, 0.94–0.99 on the Adult self-report form, and 0.96–0.99 on the Adult rated by others form. For the adaptive skill areas, the reliability coefficient range was 0.72–0.97 on the Teacher/Daycare Provider Form, 0.82–0.99 on the Teacher Form, 0.76–0.97 on the Parent/Primary Caregiver Form, 0.81–0.99 on the Parent Form, 0.80–0.99 on the Adult self-report form, and 0.82–0.99 on the Adult rated by others form.

Test–retest reliability refers to the stability of test scores over a time period. The correlations on this measure are very good. The average corrected test–retest correlations on the Parent/Primary Caregiver Form are .70 for the adaptive skill area scaled scores, .76 for the adaptive domain standard scores, and .82 for the General Adaptive Composite. The average corrected test–retest correlations on the Parent Form are .77 for the adaptive skill area scaled scores, .80 for the adaptive domain standard scores, and .86 for the General Adaptive Composite. The average corrected test–retest correlations on the Teacher/Daycare Provider Form are .80 for the adaptive skill area scaled scores, .80 for the adaptive domain standard scores, and .86 for the General Adaptive Composite. The average corrected test–retest correlations on the Teacher Form are .80 for the adaptive skill area scaled scores, .81 for the adaptive domain standard scores, and .84 for the General Adaptive Composite. The average corrected test–retest correlations on the Adult Form (self-report) are .76 for the adaptive skill area scaled scores, .85 for the adaptive domain standard scores, and .87 for the General Adaptive Composite. The average corrected test–retest correlations on the Adult Form (rated by others) are: .75 for the adaptive skill area scaled scores, .85 for the adaptive domain standard scores, and .89 for the General Adaptive Composite.

Vineland Adaptive Behavior Scales

The Vineland-3, published by Pearson, is another leading adaptive behavior measure. The measure, authored by Sara Sparrow, Domenic Cicchetti, and Celine Saulnier, was updated in 2016. The Vineland-3 includes three rating forms: interview form (for ages 3 to adult), parent/caregiver form (for ages 3 to adult), and a teacher form (for ages 3–21). Parents, close family members, teachers, day care staff, supervisors, or others who are familiar with the daily activities of the person being evaluated can complete any of these forms. Test items may be read aloud to those with poor vision or poor reading skills. The Vineland-3 offers online and paper administration options for all forms and computerized or hand scoring for all forms. Administration time is approximately 20 minutes for the Interview Form and 10 minutes for the Teacher Form.

The remainder of this section provides information on the Vineland-2 because the Vineland-3 was not published at the time this entry was written. Nine adaptive skill areas are assessed by the Vineland-2. The three adaptive domains that are addressed include communication, daily living skills, and socialization; there are optional motor skills and maladaptive behavior domains. In addition to domain scores, the test provides an overall adaptive behavior composite. All scores are categorized descriptively as low, moderately low, adequate, moderately high, or high.

The communication domain score consists of the following subdomains: receptive, expressive, and written. The receptive subdomain assesses how an individual listens and pays attention and what the individual understands. The expressive subdomain assesses what an individual says and how the individual uses words and sentences to gather and provide information. The written subdomain assesses what an individual understands about how letters make words and what the individual reads and writes.

The daily living skills domain score consists of the following subdomains: personal, domestic, and community. The personal subdomain assesses how an individual eats, dresses, and practices personal hygiene. The domestic subdomain assesses what household tasks an individual performs. The community subdomain assesses how an individual uses time, money, the telephone, the computer, and job skills.

The socialization domain score consists of the following subdomains: interpersonal relationships, play and leisure time, and coping skills. The interpersonal relationships subdomain assesses how an individual interacts with others. The play and leisure time subdomain assesses how an individual plays and uses leisure time. The coping skills subdomain assesses how an individual demonstrates responsibility and sensitivity to others.

The Vineland-2 standardization was completed using 3,695 individual cases. The norm sample was stratified based on demographic variables such as sex, race/ethnicity, socioeconomic status, and geographic region. Recruitment was based on the 2001 U.S. population demographic data. All regions of the United States were represented. Sample sizes were 1,085 for ages 0:0–4:11; 2,290 for ages 5:0–21:11; and 320 for ages 22:0–90.

Internal consistency is good. The reliability coefficient range for the communication domain is .84 to .93; in the subdomains, it was .59 to .80 for receptive, .76 to .93 for expressive, and .73 to .85 for written. The reliability coefficient range for the daily living skills domain is .86 to .91; in the subdomains, it was .66 to .83 for personal, .72 to .85 for domestic, and .77 to .83 for community. The reliability coefficient range for the socialization domain is .84 to .93; in the subdomains, it is .76 to .87 for interpersonal relationships, .58 to .83 for play and leisure time, and .78 to .88 for coping skills.

Test–retest reliability correlations are very good. The average adjusted test–retest correlation across all forms is .88 for domains; it is .88 for the communication domain, .89 for the daily living skills domain, and .85 for the socialization domain. The average adjusted test–retest correlation across all forms is .85 for subdomains. Within the communication domain, it is .89 for receptive, .84 for receptive, and .87 for written. Within the daily living skills domain, it is .85 for personal, .89 for domestic, and .87 for community. Within the socialization domain, it is .82 for interpersonal relationships, .79 for play and leisure time, and .80 for coping skills.

Inter-interviewer reliability is good for the Survey Interview Form. The average correlation between interviewers is .73 for domains; it is .68 for the communication domain, .80 for the daily living skills domain, and .72 for the socialization domain. Across all forms, the mean correlation is .70 for subdomains. Within the communication domain, it is .69 for receptive, .77 for receptive, and .74 for written. Within the daily living skills domain, it is .77 for personal, .75 for domestic, and .67 for community. Within the socialization domain, it is .71 for interpersonal relationships, .53 for play and leisure time, and .63 for coping skills.

Inter-rater reliability (Parent/Caregiver Rating Form) is good for the Survey Interview Form. The average correlation between raters is .77 across domains, .77 for the communication domain, .71 for the daily living skills domain, and .78 for the socialization domain. Reliability across all forms is .77 for subdomains. Within the communication domain, it is .82 for receptive, .72 for receptive, and .81 for written. Within the daily living skills domain, it is .63 for personal, .78 for domestic, and .85 for community. Within the socialization domain, it is .73 for interpersonal relationships, .74 for play and leisure time, and .73 for coping skills.

Alan W. Brue and Linda Wilmshurst

See also *Diagnostic and Statistical Manual of Mental Disorders*; Diagnostic Tests; Rating Scales

Further Readings

American Association on Intellectual and Developmental Disabilities. (2008). Supports Intensity Scale™ Information. Retrieved from https://aaidd.org/docs/default-source/sis-docs/latestsispresentation.pdf?sfvrsn=2

Brue, A. W., & Wilmshurst, L. (2016). Essentials of intellectual disability assessment and identification. Hoboken, NJ: Wiley.

Wilmshurst, L., & Brue, A. W. (2010). The complete guide to special education: Expert advice on evaluations, IEPs, and helping kids succeed (2nd ed.). San Francisco, CA: Jossey-Bass.

Jennifer A. Brussow Jennifer A. Brussow Brussow, Jennifer A.

Adequate Yearly Progress

Adequate yearly progress

44

47

Adequate Yearly Progress

Adequate yearly progress (AYP) is a federal accountability measure established under the No Child Left Behind Act of 2001 (NCLB). Under the AYP system, states established timelines for improving the academic achievement level of over 12 years, at the end of which 100% of students in all subgroups should perform at the proficient level or better. This entry first discusses the regulations establishing the AYP measure and the reception of the AYP process by administrators, parents, and educators. It then looks at the impact of AYP on student achievement, how the percentage of schools failing to make AYP increased over the years, and the waiver system introduced in 2011.

Regulations

Each state's department of education sets the AYP targets for each state's public schools. Private schools were not required to participate in the AYP system. NCLB requires each state's targets to follow a timeline ensuring that by the end of the 2013–2014 school year, 100% of students, including 100% of students in identified subgroups, were meeting or exceeding the state-defined level of proficiency on academic achievement assessments. With the December 2015 authorization of the Every Student Succeeds Act, the AYP system was replaced by state-determined long-term goals.

According to the federally mandated schedule, states were required to align their tests with their chosen state academic standards and begin testing students. Students in Grades 3–8 were tested annually in reading and math, and those in Grades 10–12 were tested at least. Additionally, a sample of fourth and eighth graders in each state are expected to take the National Assessment of Educational Progress reading and mathematics tests every other year. The National Assessment of Educational Progress data were used to make cross-state comparisons and compile a national report card showing aggregate levels of student proficiency.

Annual AYP targets were set separately for reading and for math achievement. Overall state targets were set for the total population of students, and separate targets were set for the subgroups of economically disadvantaged students, students of identified major racial and ethnic groups, students with disabilities, and students with limited English proficiency. The process of improving student subgroups' attainment of academic proficiency at a rate faster than the overall improvement rate is sometimes referred to as "closing the gap."

In order for a school to be considered to be making AYP, three conditions had to be met. First, at least 95% of overall students as well as 95% of the students in each subgroup with 45 or more students must have been tested. Additionally, the overall population of students as well as each subgroup of students was required to meet or exceed the state-determined objectives or increase the percentage of students meeting or exceeding the target by at least 10%. Federal guidance indicated that students can be counted more than once when determining proficiency rates. Finally, the school also had to meet the minimum annual state target for attendance rate for elementary and middle schools and graduation rate for high schools.

Using these targets, state departments of education were responsible for determining the schools and districts considered to be making AYP. When schools failed to make AYP for multiple years in a row, they were subject to the following system of penalties outlined in NCLB. Title I schools that failed to make AYP for 2 consecutive years were enrolled for the program improvement process and were designated as schools in need of improvement. Parents of children in those schools were given the choice to transfer their children to other schools that were not identified for improvement and not identified as persistently dangerous. Priority in school choice was mandated to be given to low-achieving children from low-income families. If all schools in a district were classified as in need of improvement, districts were encouraged to cooperate with neighboring districts in order to provide school choice.

After 3 years of failing to make AYP, the school was required to provide tutoring and other supplemental services for low-income students in addition to tutoring and other supplemental services for low-income students in addition to providing parents with the option of school choice. After 4 years of failing to make AYP, schools were subject to additional corrective actions. These actions included replacement of specific school staff relevant to the failure, institution of a new curriculum, the appointment of outside experts to advise the school, extension of the school year and/or school day, and internal restructuring.

Failure to make AYP in the 5th year led to development of a plan to reopen the school as a charter school, replace most or all school staff, turn over school operations to the state or to a private company, or enact some other major restructuring. After 6 years of failing to make AYP, the school is expected to implement the plan designed in the previous year. In practice, most schools opted for the “other major restructuring” option rather than completely replacing the staff or surrendering operations to an external entity. A school was eligible to exit the program improvement process when it had met AYP for 2 out of the past 3 years.

Reception

The accountability measures under NCLB and the AYP system met with resistance from many school administrators, educators, and parents. Although many lauded the legislation’s goal of having 100% of students in all demographic subgroups score proficient or better by the end of 2014, this goal was quickly criticized as impossible to attain, especially for subpopulations such as students with disabilities and students with limited English proficiency.

Some evidence suggested that the AYP system was causing beneficial increases in schools’ attention to the alignment between curriculum and instruction. However, there was also concern that apart from imposing penalties, the schools failing to make AYP were systematically stripped of needed resources instead of providing them with assistance, thus setting up a cycle of failure. Additionally, some school buildings that had performed well on other measures of success struggled to meet the proficiency benchmarks required for AYP, which caused confusion surrounding the assessment practices driving the accountability process.

As increasing numbers of buildings failed to make AYP, some observers became convinced that the AYP system overidentified schools as being in need of improvement. Additionally, several reports indicated that states were employing a variety of strategies to slow or reverse the trend of increasing numbers of a variety of strategies to slow or reverse the trend of increasing numbers of schools failing to make AYP, which included changing state testing policies by lowering cut scores, adopting new tests, and revising test administration policies. By implementing these strategies, some states managed to successfully reduce their number of buildings failing to make AYP. However, such changes to the testing process subverted the intention of the accountability system to accurately measure and improve student achievement consistently through time.

The increased testing schedule mandated in NCLB was another cause for concern. Although only 19 states had annual reading and mathematics tests in place in 2002, all states had adopted this testing schedule by 2006. This increase in time spent on testing drew concern from some parents and educators. Moreover, many parents and educators felt that the need to meet the consistently rising achievement goals set through the AYP process pressured classroom instructors to narrow their curriculum in order to address test content, thus decreasing the attention paid to subjects other than reading and math and to content that did not appear on the test, thereby depriving children of a balanced education. Finally, many states reported that they lacked sufficient funds or staff needed to implement the requirements of the AYP system, especially the corrective actions mandated for schools in the program improvement process. As a result of these concerns, several articles and reports called for the abolishment of the accountability system and the AYP measure.

Impact

The percentage of students nationwide performing at or above the proficient level on state tests increased for many subgroups of students under NCLB. In addition, scores on the National Assessment of Educational Progress generally improved since the adoption of NCLB, although critics of the law argue that the trend of performance improvement had already been established before the NCLB took effect.

Despite state-level gains in the percentage of students performing at or above the proficient level, an increasing percentage of schools failed to make AYP as time went on and the proficiency targets grew closer to 100%. From 2010 to 2011, the percentage of U.S. public schools failing to make AYP increased from 39% to 48%, which was the highest percentage since NCLB took effect in 2002, and it represented an increase of 19 percentage points over the 2006 rate of 29% of schools failing to make AYP. In 2011, at the state level, 21 states and the District of Columbia had more than half of their schools failing to make AYP. In 2011 of Columbia had more than half of their schools failing to make AYP. In 2011, the percentage of schools failing to make AYP varied widely by state, from 7% in Wyoming to 91% in Florida.

Waivers

As a result of the increasing percentage of schools failing to make AYP, the U.S. Department of Education introduced a formal process for waiving accountability requirements in 2011. The waiver process allowed states flexibility in setting new annual measurable objectives to use in determining AYP and waived the penalties for schools failing to make AYP. Waivers initially granted states flexibility in meeting the provisions of NCLB through the end of the 2013–2014 school year, though states could apply for an extension through the 2014–2015 school year. At the end of the 2014–2015 school year, states had the option to request a 3-year renewal of flexibility. By 2014, a total of 43 states, the District of Columbia, Puerto Rico, and a group of California school districts received approval for the waivers.

Jennifer A. Brussow

See also Accountability; High-Stakes Tests; National Assessment of Educational Progress; No Child Left Behind Act; Standardized Tests; U.S. Department of Education

Further Readings

No Child Left Behind Act of 2001: Part A—Improving Basic Programs Operated by Local Educational Agencies, Pub. L. No. 107-110, 20 U.S.C. § 1111 (2005). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg2.html#sec1111

Riddle, R., & Kober, N. (2011). State policy differences greatly impact AYP numbers. Washington, DC: Center on Education Policy. Retrieved from http://www.cep-dc.org/displayDocument.cfm?DocumentID=414

U.S. Department of Education. (2015). ESEA flexibility. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Usher, A. (2012). AYP results for 2010–11—November 2012 update. Washington, DC: Center on Education Policy. Retrieved from http://www.cep-dc.org/displayDocument.cfm?DocumentID=414

Websites

The Nation's Report Card: https://www.nationsreportcard.gov/

ADHD

See Attention-Deficit/Hyperactivity Disorder

Magnus Wikström Magnus Wikström Wikström, Magnus

Christina Wikström Christina Wikström Wikström, Christina

Admissions Tests

Admissions tests

47

51

Admissions Tests

Admissions tests usually refer to tests designed to find candidates suitable for higher education. Such tests and other forms of entrance examinations can be made mandatory for applicants in a country or region to take or they may be specific to a university or a university program. This entry first discusses the roles and importance of admissions tests, the principles on which they are based, and their history. It then looks at how admissions tests can be characterized and issues in admissions testing.

In an admission decision, there are two fundamental roles that a test can fulfill: to identify candidates who have sufficient knowledge to be able to complete an education (eligibility) and to rank the candidates and to make a selection in cases where there are more eligible candidates than there are available slots (selection). A test can be designed to meet either of these two roles, although most admissions tests are used only for selection purposes. Admissions tests can be further categorized with respect to the construct or constructs they are assumed to measure. Standardized aptitude tests measure aptitude in general cognitive skills and are designed to determine a person’s ability to learn. Entrance examinations are generally achievement oriented and focus on what a candidate has learned.

Having a fair selection model is of greatest importance in a democratic society. Although modern higher education often can be regarded as education for large parts of the population, universities are still institutions educating those who will hold important positions and influence society. An important question is how the number of slots at these universities should be distributed and what constitutes a fair admissions system, as this is not an easy or uncontroversial question to answer.

Fairness and Meritocracy

Fairness is closely connected to distributive justice, which concerns a socially just allocation of goods. There are several types of distributive norms describing how goods can be allocated such as equity, equality, power, need, and responsibility. In selection to higher education, there are different views on what can be considered a fair system.

Although not uncontroversial, it is common to base admissions systems on the idea of meritocracy. Applicants are ranked by their merits, usually measured by test scores or previous grades. But there may also be influences from a utilitarian approach, where equality, need, and responsibility also become important. Universities may aim for a selection that makes the student body more representative of society as a whole or in some other way more balanced.

There may be practical and ethical problems with following principles other than the meritocratic because it may be regarded as unfair to give certain groups advantages even if they are underrepresented in higher education. Often, it is prohibited to set quotas for certain groups in selection situations, irrespective of whether the group in question is underrepresented. An exception can be when the selection is made between two individuals with equal merits. In order for the meritocratic principle to be maintained, the challenge is therefore to find or develop instruments for eligibility and selection that measure the relevant construct or constructs without any bias related to student background. This has proven to be very difficult, as all measurement instruments, including grades and tests, are known to have error and often work differently for different groups of individuals.

History of Admissions Tests

Just how long higher education entrance examinations have existed is a source of debate. It has been claimed that such tests were first introduced in France in the 18th century. This fits well with the historical situation at the time when principles of equality were stressed in connection to the French Enlightenment.

Meritocracy became a leading word. It would no longer be burden and privilege that decided who was to enter higher education. Rather, it was the best performance in terms of preparation or preknowledge that would determine selection.

Other sources indicate, however, that entrance examinations to higher education were introduced much earlier, for example, in Spain in the late 1500s. In England and Germany, admissions testing was introduced in the mid-19th century. But despite the early introduction of testing, the emergence of admissions tests was still modest at this time. Although there were meritocratic ambitions in training and selection, money and privilege were still the best entry tickets to higher education well into the 1900s, especially into the more prestigious schools.

Emergence of Standardized Testing in Admissions

Even if tests have existed for a long time and in many countries, the United States should be considered the country where admissions tests were first developed on a larger scale. There, admissions tests have been in use for a long time, with the Scholastic Aptitude Test (SAT) as the first and most important of a number of different tests.

The SAT was developed by the College Board in the United States during the 1920s in order to standardize the selection process for higher education. It was originally based on the Army Alpha test, which was used in the recruitment of soldiers during World War I. The Army tests were descendants of IQ tests. For that reason, the early versions of the SAT were quite similar to IQ tests.

World War II increased the need for testing for military purposes. The demand for skilled labor increased and a large number of war veterans aimed for higher education, which boosted the industry of testing for educational purposes. Up to that point, scoring was performed manually, but by 1939, scoring became automated, which simplified the procedures surrounding the SAT.

By 1941, the SAT was psychometrically advanced, with normed and standardized scores in order to enable comparisons over time, which should be considered unique for the time period. In 1947, the Educational Testing Service was established, and since then, it has been responsible for the development and administration of the SAT. As a consequence of increased demand for higher education, the educational sector in the United States expanded significantly. This had the consequence that the market for eligibility and selection tests flourished, and during this period, the organization the American College Testing Program was formed. The company, now ACT, Inc., developed ACT as a competitor to the SAT. Both of these tests have since received great recognition and been of great importance to education in the United States, when it comes to both higher education and various preparatory courses.

Use of Admissions Tests Today

The United States is not the only country that has developed entrance examinations to higher education. But no other country has done it in the same scientific, large-scale, and standardized way. Admission tests are available in many countries, but it is then common for universities to use smaller scale tests targeted to specific training. However, there are some entrance examinations similar to the American selection tests in design and purpose, including tests used in Sweden, Israel, and Georgia.

Admission tests are usually used in conjunction with other selection instruments, and there are a large number of different models applied throughout the world. As mentioned, in the United States, the SAT and ACT are common tests to take if one wants to enter university education. Schools use tests in combination with other materials, such as secondary school performance and recommendation letters. A few other countries use aptitude-like tests such as Sweden (SweSAT) and Israel (PET). These countries also rely on other material in the selection process. In Sweden, one third of the students are admitted by the SweSAT and the remaining students are admitted by upper secondary school grades. In Israel, the PET is combined with national secondary leaving exams.

A common model used in many countries is to require students to take entrance examinations. National exams are used in countries such as China and Georgia. Turkey and Spain use national exams in combination with high school performance, while Japan and Russia use multiple examinations where one part constitutes a national exam and other parts institutionally conducted examinations.

Finally, some countries do not rely on entrance examinations at all. In those instances, high school performance is the most important selection instrument. One example where there are no tests involved is Norway. In the case of entrance examinations, the contents vary by country and institution. In some institutions, the entrance examination covers a wide field of knowledge, whereas in other countries and institutions, the exams cover more specific areas connected to the university program that is being applied for.

Characterization of Admissions Tests

The distinctions between different types of tests have to do with what one wants to measure and how one wants to use the results. A test can be classified in several ways. Tests used in the admission context are what are called indirect measurements. This means that the test asks questions about knowledge or skills, which in turn are indicators of some kind of superior knowledge, skill, or ability.

A second distinction concerns test type. Admission tests are usually norm referenced. This means that the result of the test is interpreted purely in comparison to the results of other test takers. The score on a norm-referenced test does not tell whether an individual qualifies in terms of knowledge. It merely indicates whether the individual scored higher or lower than the individual’s peers. This is natural because the purpose of admission tests is usually to rank individuals or at least to separate them. The opposite of norm-referenced tests is called criterion-referenced tests. In this case, the purpose is to measure knowledge in relation to a predetermined criterion. Such tests give information about what a person can or cannot do. The latter types of tests are sometimes used in admissions, to establish whether a candidate meets eligibility conditions.

A third characterization of tests concerns what type of question is posed, referred to as item type. Most commonly, admission tests contain closed format items such as multiple-choice items, where a respondent answers the questions by selecting one out of several prespecified alternatives. However, there are several different item formats used, and often a multiple-choice test cannot provide information that is relevant for the selection at hand. One example of this is if a selection to an artistic school is being made. In this case, it is very difficult to use a multiple-choice test because such a test will not be able to test the skills required, for example, how well a person plays the violin. In such cases, standardized multiple-choice tests may still be used but only in conjunction with other measurement of skills such as an audition.

Multiple-choice tests may also give limited information in other respects. An alternative to prespecified answer alternatives is to use an open question format, where the test taker writes an answer. This is used in some of the major admission tests such as the SAT where a writing assignment is included in the test. From a general point of view, the choice of item formats can be seen as a trade-off, where multiple-choice items have the advantage of being simple to score and considered objective in the sense that there is no human involvement in scoring, whereas open questions may have greater realism at the expense of being subjective and harder to score.

A fourth characterization concerns the format of the test. When tests use a paper-and-pencil format, the respondent takes the test in a school or test center and a proctor collects the answer sheets at the end of the test. Computerized tests have some advantages over paper-and pencil tests: A test can be scored immediately after the testing session, it is easier to vary the order of which items are presented among test takers to prevent cheating, and adaptive testing can be implemented to shorten the number of items necessary to establish a final score. One major drawback with computerized tests is that they are considered expensive and difficult to implement on a large scale.

Issues in Admissions Testing

Despite the long use of admission tests in higher education selection, there is an ongoing debate about the usefulness and the drawbacks of tests compared to other ways of ranking students. Below, two areas where the debate is continuing are discussed.

Test Validity

The meritocratic principle means that individuals who have the best chances of completing a university education should be those who are admitted. The instruments used in the selection of candidates should therefore be capable of predicting what candidates will be successful in higher education. Predictive validity is used in psychometrics as a concept to reflect the extent to which a test score predicts an outcome, in this case, how successful students will be at their academic studies.

To operationalize predictive validity, researchers and other investigators measure whether the score on the instrument correlates with academic achievement. There are numerous correlation studies made with regard to achievement. There are numerous correlation studies made with regards to admission tests. In the American context, studies are usually performed by correlating the score on the SAT (or ACT) with achievement in higher education, usually measured as the grade point average after Year 1 in university. Most studies find that admission tests, to some extent, can predict academic achievement. However, there is no consensus of exactly how well admission tests serve in this respect.

Admission test scores can also be compared to other admission instruments, such as high school grades. Comparisons of grades and test scores in their ability to predict future performance usually show that grades predict academic achievement better than test scores, but both instruments have been shown to have predictive power, indicating that using both instruments to predict academic achievement is better than just using one of the instruments.

Coaching and Cheating

In comparison with previous academic performance, one potential drawback with using tests is that the examiner cannot observe the candidate for a longer time period and is therefore less certain that the candidate has the required skills. Although the modern tests are designed so as to limit coincidence or luck, it is nevertheless problematic that candidates vary with respect to things such as test anxiety. Some individuals do not perform well when they are exposed to high-stakes tests. To limit anxiety, a candidate may benefit from learning about the test format and content.

Receiving professional coaching on the contents of a test has become increasingly popular. Although it is beneficial for the candidate in the short run, it may be detrimental for learning as well as time-consuming. In those instances where admission tests are used as the only selection instrument, learning in high school may well be hampered by the focus on a single exam. To what extent coaching is harmful for learning has yet to be determined.

A general problem with tests is cheating, such as by obtaining test questions beforehand. New technology makes cheating easier. There are examples of spy-like technology having been used to obtain the test items during a session, which then are delivered to other test takers using earpieces. Preventing and detecting cheating is very important because the aim of a test is to rank individuals to find the most suitable candidates.

Magnus Wikström and Christina Wikström

See also Achievement Tests; ACT; Aptitude Tests; Predictive Validity; SAT; Standardized Tests

Further Readings

Code of Fair Testing Practices in Education. (2004). Washington, DC: Joint Committee on Testing Practices.

Helms, R. M. (2009). University admissions: Practices and procedures worldwide. International Higher Education, 54, 5–7.

Lemann, N. (1999). The big test: The secret history of the American meritocracy. New York, NY: Farrar, Straus and Giroux.

Linn, R. (2001). A century of standardized testing: Controversies and pendulum swings. Educational Assessment, 7, 29–38.

Zwick, R. (2002). Fair game? The use of standardized admissions tests in higher education. New York, NY: RoutledgeFalmer.

Meagan M. Patterson Meagan M. Patterson Patterson, Meagan M.

Adolescence

Adolescence

51

54

Adolescence

Adolescence is a transition period from childhood to adulthood, typically spanning approximately from 12 to 18 years of age. Development during adolescence involves attaining physical and sexual maturity, along with increased complexity of thought and social behavior. Understanding adolescent development is critical for the development of policy and practice related to secondary education. This entry discusses the history of the construct of adolescence, developmental contexts and tasks of adolescence, the stages of adolescence, and major domains of adolescent development.

History

The notion of a distinct developmental stage between childhood and adulthood is a relatively new concept. In the late 19th century, theorists such as G. Stanley Hall began promoting the idea of adolescence as a distinct life stage. Prior to this time, there was a sense of youth (roughly the period from one’s midteens through early 20s) as an important and impressionable period of development, but not the modern sense of adolescence as a time of identity exploration and relative lack of adult responsibilities. The increasing availability of public education and decline in child labor contributed to the view that a period of transition from childhood to the assumption of adult roles was necessary.

Throughout much of the 20th century, adolescence was viewed as a period of “storm and stress,” in which adolescents struggled to manage their emotions, had frequent conflicts with parents and other authority figures, and engaged in high-risk behaviors. Despite this popular view, researchers have argued that the notion of adolescence as a time of high drama is largely exaggerated. Although adolescents do report more negative moods and more frequent mood swings than either adults or younger children, the majority of adolescents report feeling happy and confident most of the time.

Researchers operating from the perspective of positive youth development argue that the traditional view of adolescence has been overly negative and focused on deficits (e.g., mental illness, alcohol and drug use). Positive youth development theorists argue that researchers and practitioners should instead focus on adolescents’ strengths and structure environments such that these strengths are tapped and promoted.

Developmental Tasks and Contextual Demands

Developmental tasks are fundamental abilities and achievements that must be acquired for optimal development at a given life stage and appropriate progress toward the next phase of life. Key developmental tasks of adolescence include development of realistic self-perceptions (awareness of one’s strengths and weaknesses in various domains), identity development (including development of a vocational identity in preparation for a career), establishing autonomy from parents, engaging in appropriate peer relationships (belonging to a peer group, forming and maintaining friendships), navigating sexuality and romantic relationships, and development of coping skills (such as conflict resolution and decision making).

As in other life stages, one key determinant of individual outcomes is the goodness of fit between the individual’s needs and capabilities and what is required of the individual by the environment. Due to their developmental needs, adolescents often desire greater autonomy and flexibility from their environments than children do. Parents and teachers can promote optimal development for adolescents by allowing autonomy and choice while still acting as a source of emotional support and monitoring adolescents’ behavior and emotional states.

Stages

Researchers typically divide adolescence into three stages: early, middle, and late adolescence. Early adolescence typically begins around 11–12 years of age and continues through approximately age 14. Early adolescence is a time of rapid physical changes associated with puberty. Along with these physical changes, there are important social changes, most often including greater emotional independence from parents and increasing reliance on friends as sources of social and emotional support. During middle adolescence (approximately ages 14–16), pubertal changes near completion. Consistent with their greater physical maturity, middle adolescents often show increasing interest in romantic and sexual relationships. By late adolescence (approximately ages 16–18), pubertal changes are complete and adult appearance is in place. Late adolescents have a firmer sense of identity than younger adolescents and are clearly moving toward assumption of adult roles and responsibilities.

The onset of puberty is a key marker of the transition from childhood to adolescence. Multiple factors, including heredity, nutrition, and overall physical health, influence when an individual goes through puberty. Pubertal changes include overall body growth (in both height and weight), the onset of menstruation for girls, and the development of secondary sexual characteristics (including development of body hair, facial hair growth for boys, and breast development for girls). Typical pubertal development takes approximately 4 years.

Girls tend to begin and complete puberty earlier than boys; thus, it is not unusual for girls to be taller, heavier, and more mature in appearance than boys of the same age during early adolescence. These biological changes, combined with media images emphasizing beauty and thinness as key determinants of women’s worth, may contribute to negative body image and decreased self-esteem among adolescent girls.

Puberty may impact other aspects of development, such as self-concept and peer relationships. The timing of puberty seems to be particularly important, with early puberty relative to peers having a negative impact for girls but a more positive impact for boys.

Brain Development

Shortly before the onset of puberty, the brain experiences a growth spurt of sorts, with rapid growth of synapses (the connections between neurons that allow for the transmission of signals across the brain). During adolescence, those synapses that are not used are pruned and disappear. Throughout adolescence, the amount of myelin (a fatty sheath that increases the speed of communication between neurons) in the brain also increases. These processes of synapse generation, synaptic pruning, and myelination occur throughout the brain during adolescence but particularly in the frontal cortex. The maturation of the frontal cortex contributes to increased executive function capabilities (including control of attention, inhibition of impulses, and improved decision making) over the course of adolescence. The frontal lobe of the brain is one of the last areas of the brain to mature fully, and the relative immaturity of this structure may contribute to risk taking among adolescents.

Neural connections between various brain regions increase in strength during adolescence. These strengthened connections contribute to the cognitive advances seen in adolescence, including improvements in attention, planning, problem-solving, and self-regulation.

The brain’s sensitivity to certain neurotransmitters also shifts during adolescence. These changes mean that adolescents respond to both stressful and pleasurable events more strongly than do younger children or adults. Increased sensitivity to neurotransmitters may contribute to certain risk-taking behaviors (such as drug use) and to psychological disorders such as depression.

Cognitive, Social, and Identity Development

Cognitive Development

Cognition in the adolescent stage shows many advances over childhood cognition. These include increases in abstract thinking, scientific reasoning, planning, hypothetical reasoning (including thinking about the future), perspective taking, and metacognitive skills. Executive function skills, including selective attention, inhibition, and cognitive self-regulation, also improve. Thus, thinking in adolescence is more abstract, logical, flexible, and well-organized than children’s thinking. The increasing ability to consider hypothetical outcomes may lead adolescents to be especially idealistic in their thinking, particularly regarding abstract concepts such as justice or discrimination.

Social Cognition

The social, environmental, and biological changes that occur in adolescence lead to a greater variety of relationships and social encounters. Along with these changes come greater awareness of and interest in other people. In some cases, this awareness may lead adolescents to become self-conscious and preoccupied with how others view their appearance and behavior. This increasing self-awareness, combined with the view that one is unique and particularly worthy of others' attention, comprises the phenomenon of adolescent egocentrism.

Identity Development

As individuals move through adolescence, their self-concepts become more accurate, detailed, and nuanced. Compared to children, adolescents have a better sense of their individual strengths, weaknesses, and capabilities across a variety of domains. Across the course of adolescence, individuals become increasingly aware of and are able to consider ways in which they may have a different self in different contexts (e.g., being outspoken with friends but reserved with family members). Adolescents who have warm, supportive relationships with their parents tend to have more positive views of themselves than do adolescents whose parents are harsh, critical, or uninvolved. Similarly, encouragement from teachers, coaches, or nonparental relatives can help to promote positive self-views for adolescents.

Identity development is one of the most widely studied aspects of development in the adolescent stage. In his theory of psychosocial development, Erik Erikson describes the central psychological conflict of adolescence as identity versus identity confusion. The major developmental task of this stage is for adolescents to explore various aspects of identity (such as vocational aspirations, political beliefs, and cultural or ethnic identity) and to ultimately commit to a personal identity that is coherent and well integrated. This process often involves questioning one's own previously held beliefs or the beliefs of family and community members. A well-established identity, in Erikson's view, provides a sense of who one is and where one is going in life. Adolescents whose families provide support while encouraging exploration and self-expression tend to have a positive sense of identity. Close, supportive relationships with friends can also facilitate the identity development process.

Sexuality

Negotiating sexual identity and sexual behavior is another key developmental task of adolescence. Adolescents often spend a great deal of time and energy thinking about romantic and sexual relationships. By age 18, approximately two thirds of adolescents report having had at least one sexual partner. Adolescents who are involved in serious or exclusive dating relationships tend to initiate sexual activity earlier than those who are not involved in such relationships.

For adolescents who identify as gay or lesbian, the average age of “coming out” (disclosing one’s sexual orientation to others) is 16–17 years. Most youth come out to friends before disclosing to parents or other relatives. Coming out now typically occurs several years earlier than in previous decades, largely due to greater visibility and acceptance of gay and lesbian individuals.

Family Relationships

Although adolescents generally rely less on parents for social and emotional support than younger children, parents are still an important source of support and guidance through the adolescent years. Adolescents who have warm, supportive relationships with their parents tend to have positive outcomes in the areas of peer relationships, identity development, and academic achievement.

Responding to the adolescent’s increasing desire for autonomy is often a challenge for parents. A cooperative parent–child relationship, open communication, and continued parental monitoring of adolescents’ behavior (e.g., knowing where the child is after school, enforcing curfews) are associated with positive outcomes for adolescents.

Peer Relationships

Peers become an increasingly important source of social and emotional support during adolescence. Over the course of adolescence, friendships become more focused on intimacy (such as being able to disclose thoughts and feelings) and loyalty. This emphasis on emotional closeness in friendships is especially strong for girls. Close, supportive relationships with friends can promote positive identity development and engagement with school. Friendships tend to be closer and more intimate than romantic relationships, particularly for early and middle adolescents.

Parents and teachers are often concerned about peer pressure among adolescents. Although adolescents (particularly early adolescents) are somewhat more likely to conform to peers than younger children or adults, this is not always detrimental. For example, adolescents tend to conform to their peers in domains such as academic engagement and participation in extracurricular activities. Adolescents whose parents use an authoritative parenting style (including a balance of warmth and appropriate limits on behavior) tend to be more resistant to antisocial peer influence.

Mental Health

Many mental illnesses, such as depression, schizophrenia, and eating disorders, tend to emerge for the first time during adolescence. Depression is the most common psychological problem seen among adolescents. Both physical and environmental factors can contribute to the development and emergence of mental illness; risk factors include a family history of the disorder, high levels of family conflict, harsh or uninvolved parenting, experiences with trauma (such as abuse, sexual assault, or death of a loved one), and peer rejection. Adolescents who are gay or lesbian may be at greater risk of psychological problems such as depression or substance abuse, particularly if their family or other environments are unsupportive.

Meagan M. Patterson

See also Adultism; Childhood; Erikson’s Stages of Psychosocial Development; Puberty

Further Readings

Casey, B. J., Jones, R. M., & Hare, T. A. (2008). The adolescent brain. Annals of the New York Academy of Sciences, 1124, 111–126.

Dubas, J. S., Miller, K., & Petersen, A. C. (2003). The study of adolescence during the 20th century. History of the Family, 8, 375–397.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. American Psychologist, 48, 90–101.

Larson, R. W. (2000). Toward a psychology of positive youth development. American Psychologist, 55, 170–183.

Steinberg, L. (2001). We know some things: Parent–adolescent relationships in retrospect and prospect. Journal of Research on Adolescence, 11, 1–19.

Steinberg, L. (2005). Cognitive and affective development in adolescence. Trends in Cognitive Sciences, 9, 69–74.

Tomasik, M. J., & Silbereisen, R. K. (2012). Social change and adolescent developmental tasks: The case of postcommunist Europe. Child Development Perspectives, 6, 326–334.

Zimmerman, M. A., Stoddard, S. A., Eisman, A. B., Caldwell, C. H., Aiyer, S. M., & Miller, A. (2013). Adolescent resilience: Promotive factors that inform prevention. Child Development Perspectives, 7, 215–220.

John Bell John Bell Bell, John

Adultism

Adultism

54

57

Adultism

Adultism refers to all attitudes and actions that flow from the idea that adults are superior to young people and have the right to control and punish them at will. These attitudes are embedded in institutions, customs, child rearing practices, and relationships between young people and adults. Psychologist Jack Flasher is generally credited with first using the term in this sense in a 1978 journal article. Although not widely accepted, the concept of adultism has received attention in the children’s rights movement and within critical psychology.

Adultism is pervasive, often unconscious, and deeply influences relationships between youth and adults. It is difficult to identify, challenge, and eliminate precisely because everyone has experienced it to some degree and because much adultist behavior is considered natural and normal by most people. This entry describes examples of adultism in society; its effects, including its emotional legacy and links to other forms of societal mistreatment; and how individuals and organizations, including schools, can assess their level of adultism and find ways to avoid it.

Understanding Adultism

Children are for the most part highly controlled by adults, who tell them what to eat, what to wear, when they can talk, that they will go to school, and which friends are OK. Even as they grow older, young people are punished freely by adults, their opinions are not valued, and their emotions are often considered immature. Adults reserve the right to threaten young people, take away their “privileges,” and ostracize young people as part of disciplining them; in some cultures, even beating children is considered acceptable as part of discipline.

The fact that adults genuinely have enormous importance in and responsibility for the lives of young people may make it difficult to understand adultism. Not everything adults do in relation to young people is adultist. Young people need love, guidance, rules, expectations, teaching, role modeling, nurturance, and protection. The attitude that defines adultist behavior is disrespect for the young person’s intelligence or autonomy; this attitude allows adults to treat young people in a way that they would never treat another adult.

Adults’ approaches to young people are based partly on culture, ethnicity, gender, class, and religion, complicating the identification of adultism. Different cultures accept or reject different behaviors from children and youth; different cultures accept different degrees of harshness by adults in the punishment of unacceptable childhood behavior. Virtually no culture has identified and accepted the concept of adultism as an oppressive set of attitudes and behaviors to be understood and rejected. Something can be considered adultist if it involves a consistent pattern of disrespect and mistreatment.

Examples of Adultism

In the extensive research literature on children and youth, there is very little stating that young people are an oppressed group, with parallels to other oppressed groups. Those who do see prevalent attitudes toward young people as comparable to racism and sexism point to common statements and occurrences as examples of adultism. Common statements that show disrespect are the following: “You’re so smart for 15!” “When are you going to grow up?” and “What do you know? You haven’t experienced anything!”

Physical and sexual abuse of children is all too common. Physical punishments such as hitting, beating, and constraining children for bad behavior are widely accepted, even when illegal. Nonphysical punishments that show disrespect include routinely criticizing or yelling at young people and arbitrarily grounding them or denying them privileges. Punishment often becomes more severe if young people protest against the mistreatment.

On the other end of the spectrum, far from punishment but disrespectful nonetheless, adults often pick up little children and kiss or tickle them without asking them or allowing for the treatment to be mutual. Adults often grab things out of children’s hands without asking. These actions are not ill intended but rather conditioned by the wider culture.

Adults often talk down to children, talk about them in their presence as though they were not there, and give young people orders or lay down rules with no explanation. Although adults expect young people to listen to them, they generally do not take young people’s concerns as seriously as those of adults. Adults typically do not respect the way young people think in the way they respect adult thinking.

Adultism in Schools

Schools use hall passes, detention, suspension, expulsion, and other penalties to control students. All communities need rules to live by, but the rules in most school communities are imposed on young people and enforced by adult staff, with no input from the children or youth. Teachers sometimes yell at students and are not disciplined, but students who yell at teachers generally are disciplined. In cases of a teacher’s word against a student’s, in many schools, the teacher’s version typically prevails. Students are graded and those grades can, over time, cause them to internalize a lifelong view of themselves as “smart” or “average” or “dumb”—with a profound impact on their lives. Students, however, do not assign grades to their teachers, and when a student gets a poor grade, it is typically assumed that the student and not the teacher is to blame.

Regardless of whether school is an effective learning environment for a particular young person, an American student must attend school until at least age 16 years (and in many states until 18 years), unless parents exercise the demanding option of homeschooling. Most elementary and secondary schools give students little to no voice, power, or decision-making avenues to make significant changes.

Youth Roles and the Youth Market

Throughout U.S. society, young people find few decision-making roles and no real opportunities for developing policy or holding political power. At the same time, however, the youth market is exploited for profit as the manufacturing and entertainment industries manipulate styles, fads, and other aspects of mass culture to appeal to young people.

Effects of Adultism

The main negative messages young people receive from the treatment described earlier are that they are not as important as adults, are not taken seriously, and have little or no power. The emotional legacy of this kind of treatment, depending on its intensity, may leave scarring including anger, feelings of powerlessness, insecurity, inferiority, depression, lack of self-confidence and self-respect, and hopelessness. These emotional states can lead to unhealthy behavior. Some young people respond to these feelings by bullying, being prone to violence, or rebelling against the norm. Some become self-destructive and commit suicide, abuse alcohol or drugs, become depressed, or engage in behaviors such as cutting. Some isolate themselves, feel lonely, don’t ask for help, don’t trust, and have few close relationships. Other factors, such as poverty, trauma, serious physical or mental abuse, disability, or poor health, may also produce these results. But systematic disrespect and mistreatment over years simply because of being young contributes to feelings of powerlessness and low self-esteem.

Adultism has links to other forms of prejudice, including that mistreatment can condition people to act out against others who are less powerful. In this way, adultism conditions young people to play their respective roles in the other structures of oppression, such as sexism and racism. All of these structures of oppression reinforce each other, and how young people are treated or mistreated is closely tied to their class, gender, and ethnicity. Yet the phenomenon of being disrespected simply because of being young holds true across diverse backgrounds and environments.

Assessing Adultism

It is useful to examine youth–adult interactions, program practices, policies, and power relationships through the lens of adultism. One might ask questions such as “Would I treat an adult this way?” “Would I use the same tone of voice?” “Would I grab this out of an adult’s hand?” “Would I listen to an adult friend’s problem in this same way?” The opposite of adultist behavior includes listening attentively to young people; asking them questions about what they think and implementing some of their ideas; curbing the inclination to take over; giving them freedom to make mistakes, within safety limits; and supporting their initiative. Parents and teachers can reexamine their approach to discipline to discern possible adultism.

discern possible adultism.

On an organizational basis (e.g., school, classroom, and youth program), the following questions can help assess the level of adultism: How are young people involved in decision making? What is the evidence that young people's capabilities and intelligence are being respected? How balanced are the power relationships between adults and young people? Have the discipline policies and practices been reexamined for adultism? Do young people have an appropriate engagement in policies? For example, in schools, are elementary students' assessments of their teachers systematically gathered or are high school students involved in staff hiring, curriculum assessment, or teacher evaluation? Are the opinions and ideas of young people valued in obvious ways?

John Bell

See also Corporal Punishment; Educational Psychology; Emotional Intelligence; Erikson's Stages of Psychosocial Development; Kohlberg's Stages of Moral Development

Further Readings

Bell, J. (1995). Understanding adultism: A key to developing positive adult-youth. relationships. Somerville, MA: YouthBuild.

Burman, E. (2008) Deconstructing developmental psychology (2nd ed.). London, England: Brunner-Routledge.

Flasher, J. (1978). Adultism. Adolescence, 13(51), 517–523.

Krey, K. (2015). Adults just don't understand: Checking out our everyday adultism. Retrieved from http://everydayfeminism.com/2015/02/everyday-adultism/

Sazama, J. (2004). Get the word out! Somerville, MA: Youth on Board. Retrieved from https://youthonboard.org/publications

Sazama, J., & Young, K. (2006). 15 points to successfully involving youth in decision-making. Somerville, MA: Youth On Board. Retrieved from https://youthonboard.org/publications

Wright, J. (2001). Treating children as equals. New Renaissance Magazine. Retrieved from http://www.ru.org/index.php/education/371-treating-children-as-equals

Jennifer C. Greene Jennifer C. Greene Greene, Jennifer C.

Advocacy in Evaluation Advocacy in evaluation

Advocacy in Evaluation

Advocacy in evaluation involves a set of inherent tensions in the commissioning and practice of policy and program evaluation. The two primary tensions are (1) advocacy for particular social goods, or what are the most legitimate *values* that can be advanced in evaluation studies and (2) advocacy for particular constituencies, or who comprises the most important *audiences* for evaluation studies. These tensions arise because evaluation is both a technical and a social practice that typically takes place in politicized contexts and because evaluations of programs, especially public programs, have multiple legitimate interested audiences. This entry concentrates on evaluations of public programs and policies wherein the issues of advocacy are most salient and consequential.

Evaluation as the Social Practice of Valuing

Evaluation is a technical activity that relies on various methodologies and tools of social science. Evaluators conduct experiments, surveys, and case studies, and they use assessments, questionnaires, interviews, and observations as primary data collection techniques. Yet, as distinct from most social science research, evaluation is *also* a social practice of valuing, as it explicitly involves making *judgments of quality* regarding the program being evaluated.

The core logic of evaluation, as articulated by evaluation expert Michael Scriven, involves the comparison of data gathered to established criteria or standards of goodness. These criteria define, for that evaluation study, a good or effective program. For example, criteria for judging the quality of a new high school biology curriculum could include strong student average test performance, favorable teacher ratings of the substantive and pedagogical attributes of the curriculum, or the documented success of the new curriculum in attracting students from groups traditionally underrepresented in the sciences to biology as a field of study and possible career.

Further, various criteria convey different values regarding a good or effective program. In an evaluation of an educational program, a criterion that specifies acceptable student performance, *on average*, advances the values of egalitarianism. A criterion that addresses the *reach of this program to all students* advances values of social equity. And so the selection of criteria for judging program quality inevitably involves the advancement, or the advocacy, of some values and not others.

Recognized Audiences and Purposes for Evaluation

Audiences for evaluation are also known as evaluation stakeholders, or individuals and groups who have a legitimate stake or vested interest in the program being evaluated. There are three recognized groups of legitimate evaluation stakeholders: (1) those responsible for authorizing and funding the program—policy and other decision makers; (2) those responsible for implementing the program—administrators, staff, and volunteers; and (3) the intended beneficiaries of the program, their families, and communities as well as the broader public.

Stakeholders characteristically have different interests in and thus different evaluative questions about the program. Different evaluation questions are further linked to different evaluation purposes. Decision makers usually want to know if the program "worked," that is, achieved its intended programmatic and policy *outcomes*. Program implementers typically want to know how they could *improve* the design and implementation of the program, for example, to reach more people or to adapt program materials for newly arrived immigrant families. And program participants are usually interested in how well the program's *promised benefits match their own particular needs*.

Evaluation audiences and purposes are linked, in part, to the developmental stage of an educational program. A new computer technology program that provides a laptop computer and related instruction for every child in Grades 3–5 in a particular urban school district would be evaluated for the purposes of program improvement and thus for audiences of program developers, administrators, and staff. An established technology-oriented after-school program that has been through several cycles of implementation and improvement would be most appropriately evaluated to assess its outcomes, both intended and unintended, for broad audiences of educational policy makers and community families alike.

The selection of audiences for an evaluation study is further contingent on additional contextual factors that include policy priorities, funding, and sociocultural factors. Even so, the evaluator has both the authority and the responsibility to contribute to the identification of key evaluation audiences and purposes. And so, beyond programmatic and contextual contingencies, the selection of audiences for an evaluation study involves the privileging, or advocacy, of some stakeholder priorities over others.

Evaluators’ Responsibilities for Advocacy

Social policy and program evaluations are nearly always initiated by those responsible for funding, implementing, or critically reviewing a given social, educational, or health program. Evaluators are then called in to conduct an evaluation that is already substantially framed and bounded by extant priorities and expectations. This framing usually includes expectations of the criteria to be used to make judgments of quality, the primary purposes and audiences for the evaluation, and the key questions to be addressed.

Often, all of the factors that frame an evaluation are not explicitly stated. Even so, it is the evaluator’s responsibility to ensure that the advocacy in that evaluation—which values are advanced and whose interests are addressed—is defensible, fair, and serves the broader public good.

Jennifer C. Greene

See also Democratic Evaluation; Ethical Issues in Evaluation; Evaluation Versus Research; Values

Further Readings

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., & Weiner, S. S. (1980). Toward reform of program evaluation. San Francisco, CA: Jossey-Bass.

Greene, J. C. (1997). Evaluation as advocacy. Evaluation Practice, 18, 25–35.

House, E., & Howe, K. (1999). Values in evaluation and social research. Thousand Oaks, CA: Sage.

Scriven, M. (1981). The logic of evaluation. EdgePress.

AEA

See American Evaluation Association

AERA

58

AERA

See American Educational Research Association

Wilfridah Mucherah Wilfridah Mucherah Mucherah, Wilfridah

African Americans and Testing African americans and testing

58

61

African Americans and Testing

This entry describes issues related to African Americans and testing and discusses possible reasons for the relatively low average performance of African Americans on standardized tests. The difference in test performance, or *achievement gap*, between African American students and European American students has been the subject of much research. Many factors may contribute to the achievement gap, including socioeconomic differences, the home environment, parent educational level, and teacher perceptions of students’ academic abilities. Understanding and addressing the reasons for lower test performance among African Americans is important because of the use of test results in diagnosing learning problems, determining a student’s academic level, and making other significant decisions about schools and students.

Importance of Testing

Various kinds of decision making are involved in teaching. It is necessary for teachers to know their students’ performance in the classroom. Testing is one way of doing this. Testing allows a teacher to compare one student’s performance on a particular task with a set standard or the performance of other students by gathering information about student learning. Many people can design a test, including a classroom teacher; local, state, or federal government agencies such as school districts; or commercial test development firms. When done properly, testing can provide unbiased data for decision making, giving it a large role in many classroom decisions.

Testing became more common in K–12 classrooms, with the No Child Left Behind Act of 2001, which used the results of standardized tests to measure school performance. Test results are also used to make many decisions about individual students. Achievement tests are common measures given to students and are meant to measure a student’s level of learning in specific content areas such as reading comprehension, language usage, computation, science, social studies, mathematics, and logical reasoning.

Admission to kindergarten; promotion from one grade to the next; high school honors, Advanced Placement classes and graduation; access to special programs; placement in special education classes; teacher licensure and tenure; and school funding may all be affected by test results. In making these decisions, it is of great importance to consider the quality of the test itself and the way the test is used. For example, many African Americans are inappropriately identified as needing remedial instruction, and a majority are placed in special education classes.

It is important to know the consequences of choosing one test over another for a particular purpose with a given group. Of equal significance is understanding how the test scores of minority-group students will be interpreted and the effect of testing on each of the students. In addition, we need to be cognizant of what we mean by intelligence, competence, and scholastic aptitude and what implications come with each of these. Do our views agree with those implied by the tests we use to measure these constructs? How will other information about the individual be integrated with test results to make major decisions or judgments? Responding to these questions requires choices based on values as well as accurate information about what tests can and cannot tell us.

African American Students and Test Performance

Research evidence shows African American students underperform on academic tests due to a variety of reasons including socioeconomic differences, the home environment, parent education, teaching practices, teacher perceptions, and anxiety. The increase in child poverty and a greater emphasis on low-level skills in high-poverty schools may be worsening this gap.

When compared to students from low-socioeconomic families, students from higher socioeconomic families are rated as having higher academic ability. African American students tend to be overrepresented in the lower socioeconomic status. One possible explanation as to why there is a significant difference by socioeconomic status in teacher perceptions of students’ academic ability is the mismatch between what the test measures and what the teacher is evaluating. For instance, research shows African American students and students from low-socioeconomic families possess a smaller vocabulary than European American students and students from middle-/high-socioeconomic families. Socioeconomic status may be a proxy for vocabulary differences that teachers may be incorporating into their perceptions of students' academic ability, which are not captured by the tests and teacher perceptions.

Depending on the teacher–student interaction, the skills and habits that a student demonstrates may be rewarded differently. Parents and students with higher socioeconomic status tend to possess higher levels of cultural capital that is valued in the school and testing settings. The cultural capital that a teacher may reward in the classroom includes social and behavioral skills.

Some researchers point to institutional-based racism toward African American students in educational testing as another reason for the achievement gap, with some of these researchers using the mathematics teaching in schools as an example. They argue that approaches to mathematics are mainly driven by what works for European American students, and the poor achievement test scores in predominantly African American schools are a reflection of this institutional racism. One way to address this problem is to intentionally include components of the cultural history of African Americans that could be applied to mathematics.

Research examining various factors that contribute to the achievement gap between African American and European American students has shown that teachers' perceptions of students' academic ability is a significant factor in the observed gap. Research evidence reveals that in addition to academic skills, teachers' valuations of students include their work habits, motivation, effort, and behavior, commonly referred to as social and behavioral skills.

African American students tend to receive lower ratings on measures of social and behavioral skills than European American students even when controlling for other characteristics such as students' socioeconomic status, gender, age, family structure, test scores, and prior social and behavioral skills. Teacher reports of social and behavioral skills are seemingly more important for teacher perceptions of student academic ability for African American students than for European American students. In other words, classroom behavior has a larger influence on how teachers perceive the academic ability of African American students than it does for European American students. According to research in this area, teacher perceptions of students' academic ability are sustained over this area, teacher perceptions of students' academic ability are sustained over time.

Anxiety appears to play a significant role in the achievement gap between African American and European American students. A study examining the achievement gap between minority and majority racial groups in schools sought to measure academic test results and anxiety related to those tests among diverse high school students. The study results showed that European American students performed better on the tests, and race accounted for 9–23% of the variance, even after controlling for educational opportunities and socioeconomic status.

African American students are often anxious about negative consequences of failing, such as not receiving a regular diploma and having restricted access to college or trade school. Another aspect of anxiety that may be experienced by African American students during testing is stereotype threat, which refers to the risk of confirming a negative stereotype of one’s group. A seminal 1995 study by Claude Steele and Joshua Aronson found that African American students did better on a test composed of difficult verbal questions from the GRE General Test when they were told it was for research on psychological factors in problem-solving than when they were told it measured their verbal abilities. Numerous studies have since been published on stereotype threat among a range of groups.

High-Stakes Testing and African American Students

High-stakes testing may have further widened the gap between the scores of African American and European American students. Decisions affected by test scores are so critical that many educators call this process high-stakes testing. High-stakes testing refers to standardized tests whose results have powerful influences when used by school administrators, other officials, or employers to make decisions. For example, high-stakes test results can be used to hold teachers, schools, and administrators accountable for student performance.

It is reasonable to expect that a test measures what has been taught due to the weight of what rides on test results. However, research shows that fewer than 10% of the items in students’ curricula overlapped with both the textbooks and the standardized tests students were given. In response to this mismatch, some teachers have resorted to “teaching to the test.”

Because of the average difference between the test scores of African American

Because of the average difference between the test scores of African American and European American students, teachers of predominantly African American students are most likely to teach to the test in an effort to narrow the achievement gap. This emphasis on test performance may, however, lead to increased dropout rates among African American students if they feel they are going to fail the exam that many states require for high school graduation. Without the expectation of graduating, they see no point in continuing to attend school.

Finally, African American students may underperform on achievement tests because teaching strategies and academic content often do not align with their lived experiences. African American students may perform better on tests if teaching strategies align closely with the specific types of problems that the students will encounter on the test and are embedded in a culturally responsive pedagogy.

Wilfridah Mucherah

See also Achievement Tests; Alignment; Anxiety; Asian Americans and Testing; Cultural Competence; Culturally Responsive Evaluation; Gender and Testing; High-Stakes Tests

Further Readings

Chavous, T. M., Bernat, H., Schmeelk-Cone, K., Caldwell C. H., Kohn-Wood, L., & Zimmerman, M. A. (2003). Racial identity and academic attainment among African American adolescents. Child Development, 74(4), 1076–1090.

Davis, J., & Martin, D. B. (2008). Racism, assessment, and instructional practices: Implications for mathematics teachers of African American students. Journal of Urban Mathematics Education, 1(1), 10–34.

Herman, M. R. (2009). The black-white-other achievement gap: Testing theories of academic performance among multiracial and mono-racial adolescents. Sociology of Education, 82(1), 20–46.

Minor, E. C. (2014). Racial differences in teacher perception of student ability. Teachers College Record, 116(10), 1–22.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? Contemporary Educational Psychology, 26(3), 291–310.

Zachary Conrad Zachary Conrad Conrad, Zachary

Age Equivalent Scores

Age equivalent scores

61

62

Age Equivalent Scores

An age equivalent (AE) score is a type of norming that provides an estimate of the chronological age (CA) at which a typically developing child demonstrates the skills displayed by the child being assessed. Scores are intended to convey the meaning of test performance in terms of what is typical of a child at a given age based on the mean raw score on a test obtained by the group of children in the normative sample at a specific age. This entry describes how AE scores are calculated and discusses their limitations.

AE scores are often expressed in years and months (e.g., 5;0 for 5 years, 0 months). In simple terms, if on average children at 36 months of age obtain a score of 10 correct responses on a particular test, then any child obtaining a score of 10 correct will receive an AE of 36 months.

Limitations

Despite the wide use of AE scores, there are several well-documented limitations associated with these scores. First, in contrast to standard scores and percentile ranks, AE scores do not take into consideration the range of normal performance for individuals whose scores fall within the average range. Rather, these scores represent the age at which a given raw score is average. It would be expected that half of the examinees on a test will achieve a higher AE score than their corresponding CA. Similarly, half of the examinees should receive a lower than average AE score.

The lack of consideration for a range of normal performance results in AE scores implying a false standard of performance. For example, one might expect a 4-year-old child to earn an AE score of 4;0. However, due to the nature of AE scores, half of the 4-year-old examinees will earn an AE score that is below their CA. A child who receives a standard score or percentile rank that is below the mean for a given age-group may be performing well within the range of normal or within 1 standard deviation away from the mean. This same examinee might earn an AE score that is significantly below the examinee's CA. Therefore, AE scores make no attempt to describe a normal range of performance, therefore these types of scores are ineffective in case management decisions.

A second reported limitation of AE scores is that these scores promote typological thinking. AE scores compare children to the "average *x*-year-old." However, the average *x*-year-old does not exist. Rather, the term *average* represents a range of performance for a particular age-group.

A third serious limitation of AE scores is the lack of information about a test taker's performance on a given test. When two children earn the same AE score, the examiner cannot assume that the children responded the same way to the stimulus items on the test. Earning the same AE score simply means that these two children answered correctly the same number of questions. Although a 5-year-old and a 10-year-old may earn the same AE score, these two children may have approached the stimulus items differently. That is, they may have demonstrated varying performance patterns. It is likely that the younger child performed lower level work with greater consistency, reaching a ceiling early on. The older child likely attempted more problems but performed at a lower accuracy level. Consequently, AE scores would be ineffective in making inferences about what can be expected from these children regarding their language abilities. AE scores may also be ineffective in assessing children with severe developmental delays or mental retardation. AE scores are not valid when evaluating children with Down syndrome because these children may use different underlying processes when approaching stimulus items. If the development of these two groups of children is not comparable, AE scores are no longer valid for children with Down syndrome. In other words, the fact that a child with Down syndrome has an AE score that is similar to that of a much younger typically developing child does not mean that the child with Down syndrome is using the same underlying processes as the younger child.

A fourth commonly reported limitation of AE scores is the derivation method of these scores. AE scores are derived through interpolation and extrapolation. For example, when deriving AE scores for a test, the test developers may examine a example, when deriving AE scores for a test, the test developers may examine a group of children in the normative sample whose CAs fall between 5;0 and 5;5. These children’s scores are plotted and smoothed into a graph. Using this graph, the AE scores for children at each month interval are estimated or extrapolated. The AE score for each age represents the average raw score for that age-group of children. Thus, when AE scores are calculated, they represent a mean score of a group of children who were not actually tested.

A fifth problem with AE scores is that these scores falsely imply that abilities increase at a constant rate from year to year. Unlike standard scores, which follow an equal-interval scale, AE scales are ordinal, with a flattening of the curve as the age increases. That is, as age increases, similar differences in AE scores are due to the smaller and smaller differences in raw scores. For example, a difference in AE scores of 3 months for a 4-year-old is more significant than a difference of 3 months for a 14-year-old. Therefore, AE scores cannot be used to demonstrate change in a child’s skills over time.

Zachary Conrad

See also Grade Equivalent Scores; Norming; Scales

Further Readings

Elliott, C. D., Smith, P., & McCulloch, K. (1996). British Ability Scales Second Edition (BAS II). Administration and scoring manual. London, UK: Nelson.

Maloney, E., & Larrivee, L. (2007). Limitations of age-equivalent scores in reporting the results of norn-referenced tests. Contemporary Issues in Communication Science and Disorders, 34, 86–93.

McCauley, R. J., & Swisher, L. (1996). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. Journal of Speech and Hearing Disorders, 49, 338–348.

Salvia, J., Ysseldyke, J., & Bolt, S. (2006). Assessment: In special and inclusive education. Boston, MA: Houghton Mifflin.

Andrew T. Roach Andrew T. Roach Roach, Andrew T.

Jacquelyn A. Bialo Jacquelyn A. Bialo Bialo, Jacquelyn A.

Alignment

Alignment

62

64

Alignment

In educational assessment, alignment refers to how well assessments measure what is taught or intended to be taught. In 2002, Norman Webb described alignment as the “extent to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (p. 1). This entry discusses the models used to measure and understand alignment and the reasons why alignment is important.

Most alignment models consider the match or overlap between curriculum (in the form of content standards or curriculum guides), tests or other assessment tools, and classroom instruction. Because of this, measures of alignment are best thought of as a form of content-related evidence of validity. According to Stephen Haynes, David Richard, and Edward Kubany, content-related validity is understood to be how well an assessment instrument reflects the particular construct that is being measured by the instrument. Although the concept of alignment can be applied in a variety of contexts (e.g. credentialing, employment tests), its most frequent application has been in the realm of K–12 standards-based accountability.

Among the many provisions in the No Child Left Behind Act of 2001, perhaps none received as much attention as the requirement that states develop and administer annual statewide standardized tests in Grades 3–8 and at least once in high school. These tests were intended to measure both students’ knowledge and their progress toward meeting state-defined performance standards. The idea behind the testing requirement was that combining student achievement data with strong accountability consequences for schools, districts, and state education agencies would result in improved academic outcomes. To achieve this objective, educational systems needed to ensure (and were federally mandated to demonstrate) the alignment between their content standards and standardized assessments.

The Council of Chief State School Officers issued a monograph that reviewed the three frameworks most commonly used by states and test developers for evaluating alignment: (1) the Webb model, (2) the surveys of enacted curriculum model, and (3) the Achieve model. Each of these frameworks involves expert review of standards and assessments that results in a series of indices characterizing the extent of match or overlap in state standards, assessments, and (in the case of the surveys of enacted curriculum) classroom instruction.

It is important to note that alignment is not a dichotomous variable (i.e., aligned *vs*. not aligned). Rather, the information produced by alignment studies can be used by policy makers, test developers, and educators to make adjustments to test content or instructional practices to improve the extent of alignment with the curricular expectations outlined in content standards.

Clearly, an insufficient degree of alignment (i.e., a significant mismatch between content standards and test content) can result in fragmentation and confusion for educators and students. For example, in the absence of alignment, how are educators to determine the skills and knowledge most important to teach? Moreover, if test content does not match what was taught to students in class, they may experience frustration and failure on required assessments. A lack of alignment between these elements also calls into question any inferences drawn from assessments. Without demonstrating adequate alignment between tests and content standards, it is impossible to determine whether a school’s success or failure in demonstrating adequate yearly progress can be attributed (at least in part) to the quality and content of classroom instruction.

Although other models have been proposed for understanding alignment, the intended curriculum model developed by Alexander Kurz and Stephen Elliott is the most recent and comprehensive one. The model demonstrates curricular expectations expressed at different levels in the educational system: system-wide, classroom, and individual student.

According to the intended curriculum model, two types of curriculum exist at the system level: the *intended curriculum* and the *assessed curriculum*. Intended curriculum refers to subject-and grade-specific content and skills that are outlined in content standards, teacher's manuals, or curriculum guides. The second system-wide curriculum is the assessed curriculum, representing the content actually measured during testing. Both of these system-level curricula could be viewed as policy tools sending messages to educators and students about the skills and concepts that are valued and important. For example, teachers may make decisions about the topics to be emphasized based on what they know or believe will be on the subsequent high-stakes test.

Moving from the system level, Kurz and Elliott define a series of curricula at the teacher and student level. The *planned curriculum* represents the teacher's actually teaching plans based on the content outlined in the standards, whereas the *enacted curriculum* represents the content the teacher subsequently delivers during instruction.

The planned and enacted curricula can introduce substantial variation in alignment across the system. For example, a number of factors (e.g., teacher expertise, student skill level) may result in an individual teacher's decision to emphasize some aspects of the intended curriculum while simultaneously de-emphasizing or skipping others. The most widely used alignment evaluation frameworks (the Webb model, the surveys of enacted curriculum model, and the Achieve model) focus on alignment at the intended curriculum model's system level and (sometimes) the teacher level.

Kurz, however, indicates that variation at the individual student level may influence alignment as well. The *engaged curriculum* consists of instructional content on which a student is productively engaged. Student engagement might be influenced by a variety of factors including difficulty level of the task, classroom behavior, or the quality of teachers' classroom management. Only instruction that is provided in a manner and context that facilitates productive engagement is likely to become part of students' *learned curriculum* and subsequently part of their *demonstrated curriculum*.

The demonstrated curriculum represents the skills and understanding students are able to produce as part of the standardized test or other assessment strategies. Factors beyond quality of instruction and student engagement also influence students' ability to demonstrate their learning. If assessment items are poorly designed or the needed testing accommodations are not made available, students designed or the needed testing accommodations are not made available, students may be unable to produce responses that represent the true scope of their learning.

Andrew T. Roach and Jacquelyn A. Bialo

See also Adequate Yearly Progress; Content-Related Validity Evidence; No Child Left Behind Act

Further Readings

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment, 7, 238–247.

Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), Handbook of accessible achievement tests for all students (pp. 99–129). New York, NY: Springer.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2009). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. The Journal of Special Education, 44, 131–145.

Kurz, A., Talapatra, D., & Roach, A. T. (2012). Meeting the curricular challenges of inclusive assessment: The role of alignment, opportunity to learn, and student engagement. International Journal of Disability: Development and Education, 59, 37–52.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. Review of Educational Research, 79, 1332–1361.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. Psychology in the Schools, 45, 158–176.

U.S. Department of Education. (2004). NCLB: Title I—Improving the academic achievement of the disadvantaged. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/pg1.html#sec1001

Vockley, M. (2009). Alignment and the states: Three approaches to aligning the national assessment of educational progress with state assessments, other assessments, and standards. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2002). Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states. Washington, DC: Council of Chief State School Officers.

Hyung Won Kim Hyung Won Kim Kim, Hyung Won

Alpha Level

Alpha level

64

66

Alpha Level

In statistical hypothesis testing (or *tests of significance*), one assumes that the null hypothesis is true about a reference population and attempts to reject it by seeking evidence for the alternative hypothesis. This is done by taking a sample and evaluating whether the sample provides evidence to support the alternative hypothesis. To do so, it is customary to compute the *p* value. The rejection of the null hypothesis depends on the comparison of the *p* value with a threshold probability value (chosen by the experimenter), which is referred to as the α *level* (or *level of significance*) of the test and is symbolized as the Greek letter α. This entry discusses the calculation and interpretation of the α level, the history of its use in statistics, statistical hypothesis testing using the rejection region, and misconceptions surrounding the α level.

Comparing a *p* value with a chosen α level allows one to make a conclusion about the statistical significance of results. Suppose that the *p* value associated with a sample is very small. This means that the sample outcome (a statistic of the sample) is very unlikely under the assumption that the parameter is the value stated in the null hypothesis, and it serves as evidence favorable to the alternative hypothesis. Suppose contrarily that the *p* value associated with a sample is not small. This means that the sample outcome is not unlikely under the same assumption and that the data fail to serve as evidence for the alternative hypothesis.

To determine how small a *p* value has to be to reject the null hypothesis, one needs a threshold value: α. That is, if the *p* value is less than the α, one is able to reject the null hypothesis; but if the *p* value is greater than the α, one cannot reject the null hypothesis. While customary α levels are .001, .01, .05, or .1, in most applications .05 or .01 is specified. If, for example, α = .05, then the confidence level that the test would lead one to the correct conclusion that the null hypothesis is true when it is in fact true is .95 (=1 − α). It is important that a researcher specify the α level prior to setting up the statistical test. This is because it is ethically problematic to choose an α level after identifying the *p* value, which would allow a researcher to manipulate the conclusion.

Underlying Meaning and Interpretation of α Level

An α level of .05 means that we allow a 5% risk of rejecting the null hypothesis even if it is true, and the difference between the obtained outcome statistic and the parameter specified in the null hypothesis is due to sampling error. The α level of .05 defines what results are improbable enough to allow an experimenter to take the risk of rejecting the null hypothesis when it is true. That is, if the *p* value is less than .05, one would conclude that the observed effect actually reflects the characteristics of the reference population rather than just sampling error. Contrarily, if the *p* value is greater than .05, one would fail to make this conclusion. Other α levels, such as .1 or .01, may be adopted, depending on the field, the nature, and the circumstances of the study. Compared to the α level of .05, the α value of .01 is more cautious, while the α value of .1 is less cautious.

The process of making conclusions entails the possibility of two types of errors: (1) concluding that the observed effect of a statistical outcome (an observed value) occurred due to actual changes in the reference population when the effect is actually due to sampling error alone and (2) concluding that the observed effect of a statistical outcome occurred due to sampling error alone when the effect is actually due to a change in the parameter. These are referred to as Type I and Type II errors, respectively.

Charles Henry Brase and Corrinne Pellillo Brase have stated that the α level of a test (the probability of rejecting the null hypothesis given that it is actually true) may be defined, in terms of *risk* and *error*, as the probability at which one is willing to risk a Type I error. While a Type I error depends solely on the choice of α level, a Type II error depends on a Type I error (the α level selected before the test), the initial estimation of changes in the parameter, and the sample size. The probability of committing a Type II error is denoted by the Greek letter β.

Hypothesis testing methods require controlling α and β values to keep them as small as possible. Depending on the nature and context of the test, controlling one type of error may be more important and viable than controlling the other type. Setting the α level at .05 means setting the probability of making a Type I error (or the conditional probability of rejecting the null hypothesis given that the null hypothesis is true) at 5%. Represented graphically in terms of area, an α level of .05 means that the area in which the evidence leads to a rejection if the sample statistic falls into it is 5% of the total area of the sampling distribution.

History

The idea of significance testing in statistics was initiated and outlined by a British statistician, Ronald Fisher, in the early 20th century. In 1925, Fisher published *Statistical Methods for Research Workers*, where he suggested the probability of .05 (1-in-20 chance) as a cutoff level to reject the null hypothesis. Later, Fisher changed this recommendation, suggesting that the cutoff level should be chosen by the experimenter depending on the specific circumstances of the experiment and the field. Two other early to mid-20th-century statisticians, Jerzy Neyman and Egon Pearson, collaboratively contributed to the development of hypothesis testing theory and laid the foundation of modern statistical hypothesis testing. In particular, they noted the importance of setting the α level prior to any data collection.

Statistical Hypothesis Testing Using Rejection Region

A null hypothesis can be considered as the default statement that indicates no change in the parameter of interest. As shown earlier, one way to determine whether a null hypothesis should be rejected (or retained) is to compare a *p* value with an α level. Another way involves considering the *rejection region*.

The graphical representation of the α level is as part of the total area under the probability curve of the test distribution; the part corresponding to the α level is the rejection region. A *critical value* is the point where the rejection region is cut off from the nonrejection region. In order to decide whether a null hypothesis should be rejected, one can compare the test statistic (calculated from the statistic of the sample) with the critical value. If the test statistic falls in the rejection region, it leads to the conclusion that the null hypothesis should be rejected. In a one-tailed test, the rejection region for an α level of .05 would be allocated to one side (or one tail) of the test distribution and take up 5% of the area under the density curve. In a two-tailed test, the rejection region for an α level of .05 still takes up 5% of the area under the density curve but is divided between the two ends (the tails) of the test distribution, each with 2.5% of the area.

Misconceptions

The α level of a hypothesis test should be interpreted as the probability of rejecting the null hypothesis when the null hypothesis is true. Common misinterpretations of the α level include that it instead indicates the level at which the null hypothesis is proven improbable or false, or, conversely, true; the level of the probability of accepting the null hypothesis when it is true; or the level of confidence in the probability of the null hypothesis being false.

Hyung Won Kim

See also Hypothesis Testing; Inferential Statistics; *p* Value; Significance; Type I Error

Further Readings

Brase, C. H., & Brase, C. P. (2003). Understandable statistics: Concepts and methods (7th ed.). Boston, MA: Houghton Mifflin.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? Journal of Statistics Education, 17(2). Retrieved from www.amstat.org/publications/jse/v17n2/castrosotos.html

Everitt, B. S. (2006). The Cambridge dictionary of statistics. New York, NY: Cambridge University Press.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, Scotland: Oliver and Boyd.

Johnson, B., & Christensen, L. (2008). Educational research: Quantitative, qualitative, and mixed approaches. Thousand Oaks, CA: Sage.

Leon-Guerrero, A., & Frankfort-Nachmias, C. (2014). Essentials of social statistics for a diverse society. Los Angeles, CA: Sage.

Moore D. S. (2003). The basic practice of statistics (Vol. 3). New York, NY: W. H. Freeman.

Martha L. Thurlow Martha L. Thurlow Thurlow, Martha L.

Alternate Assessments Alternate assessments

66

70

Alternate Assessments

Alternate assessments are measures of academic content or English proficiency intended for students with disabilities. Alternate assessments are different in one or more ways from general assessments intended for the majority of students in schools, including students with disabilities. This entry describes the three types of alternate assessments for students with disabilities that have been identified in federal laws or regulations. It examines each type in terms of the students for whom it was intended, its use in states, and the technical qualities it was expected to meet. The entry concludes with evidence of the consequences that have been attributed to one of the types of alternate assessments, the alternate assessment based on alternate achievement standards, which by 2015 was the only alternate assessment recognized in federal law.

In the early 1990s, alternate assessments were initially used by some states as a way to include all students with disabilities in large-scale assessments designed to measure the academic achievement of students on state-defined content standards. Students with disabilities are a diverse group of students, with varying disability characteristics. Among the most prevalent of the disabilities that qualify students for special education under federal law are specific learning disabilities, speech and language impairments, autism, emotional disabilities, other health impairments, and intellectual disabilities. Among the least prevalent of these disabilities are visual impairments, hearing impairments, and orthopedic impairments.

Depending on the disability a student has, the student may need to be assessed with accessibility supports or accommodations that may be different from those needed by another student with a disability. Small numbers of students with disabilities may need to be assessed with an assessment that is different in some way from the general assessment that most students with disabilities take either with or without accessibility supports and accommodations. These other assessments may be in a different format or require the student to meet different expectations from the general assessment.

In the 1997 reauthorization of the Individuals with Disabilities Education Act, alternate assessments were first introduced in federal law. The requirements for alternate assessments have been defined and refined because they were first introduced as assessments for those students with disabilities unable to participate in general assessments even with accommodations. Accommodations are changes in testing materials or procedures that provide access to the assessment without changing what the test is intended to measure.

Because the Individuals with Disabilities Education Act required the development of state alternate assessments for students with disabilities, three types of alternate assessments have been identified in federal regulations for the Elementary and Secondary Education Act (ESEA), which was reauthorized in 2002 as the No Child Left Behind Act and in 2015 as the Every Student Succeeds Act (ESSA). These are alternate assessments based on alternate achievement standards (AA-AAS, also referred to as alternate assessments based on alternate academic achievement standards), alternate assessments based on grade-level achievement standards (AA-GLAS, also referred to as alternate assessments based on grade-level academic achievement standards), and alternate assessments based on modified achievement standards (AA-MAS, also referred to as alternate assessments based on modified academic achievement standards). These alternate assessments varied in terms of the performance standards (called achievement standards) on which they were to be based for judging student performance. All of the alternate assessments have been required to be based on the same grade-level content that is assessed by general assessments. The requirements for the achievement standards were defined through regulations tied to the ESEA.

AA-AAS

Alternate achievement standards are standards for performance that are different in complexity from the GLAS held for students taking general assessments. For example, when a student in the general assessment is asked to provide an extended written response to a prompt, the student participating in the AA-AAS might be asked to enter picture cards into a graphic organizer to convey a response.

Students Intended to Take the AA-AAS

The AA-AAS is intended for a small group of students who have significant cognitive disabilities. This group of students does not comprise a disability category but is generally recognized to include primarily students with intellectual disabilities, autism, and multiple disabilities but not all students in those categories. Students with significant cognitive disabilities have disabilities that affect their intellectual functioning and adaptive behavior. Adaptive behavior refers to the knowledge and skills needed for living independently and functioning safely in daily life.

Under No Child Left Behind Act accountability rules setting targets for the percentage of students testing proficient in English and math, no more than 1% of tested students could be considered proficient based on their performance on an AA-AAS. Because of this, it was sometimes referred to as the 1% assessment. In 2015, when Congress replaced the No Child Left Behind Act with ESSA, the AA-AAS was incorporated as an expected part of assessments used for school accountability. Participation in the AA-AAS was limited to 1% of the total student population at each grade to ensure that only students with the most significant cognitive disabilities were included in the assessment. In addition, the AA-AAS was the only alternate assessment recognized by ESSA for inclusion in school accountability systems. Neither the AA-GLAS nor the AA-MAS could be used for school accountability purposes under the requirements of ESSA.

Use of AA-AAS in States

All U.S. states, along with the District of Columbia, Puerto Rico, Virgin Islands, Guam, and other jurisdictions that receive federal special education funding, have developed AA-AAS for their students with significant cognitive disabilities. These assessments are primarily item-based assessments, similar to states' general assessments, although some are body-of-evidence portfolios and some are rating scales. All states have AA-AAS for English language arts, mathematics, and science, as required by federal law. Some states also have AA-AAS for other content areas, such as social studies.

Clarification of the Qualities of AA-AAS

Clarification of the Qualities of AA-AAS

AA-AAS are to meet the same technical quality requirements as other assessments in which students with disabilities participate. These technical quality requirements include validity, reliability, fairness, and accessibility, and others that address alignment to content standards, test design and item development, scoring, and test security. Addressing these technical quality requirements evolved over time as the understanding of students with significant cognitive disabilities improved and expectations for their performance increased.

Alternate Assessments Based on Grade-Level Achievement Standards (AA-GLAS)

GLAS are standards for performance that are the same as the GLAS held for students taking general assessments. This type of assessment addressed the need for different procedures for demonstrating grade-level performance, such as completing a portfolio or participating in a performance assessment to demonstrate the same level of proficiency as could be demonstrated on traditional general assessments that included multiple-choice and extended response items.

Few states ever developed an AA-GLAS. Instead, they relied on accommodations to ensure that assessments were appropriate for their students with disabilities working on GLAS. In 2015, ESSA eliminated this alternate assessment as an optional assessment for English language arts, mathematics, and science.

Although AA-GLAS were eliminated for content area assessments, federal guidance released in 2014 confirmed that states needed to have alternate assessments for some English learners with disabilities taking the state English language proficiency assessment. The guidance required that these alternate assessments be based on the same criteria for proficient performance as English language proficiency assessments for English learners without disabilities.

Students Intended to Take the AA-GLAS

The AA-GLAS was intended for students with disabilities who needed a different way to demonstrate their grade-level performance. Among the students identified as needing a different way to demonstrate the same level of identified as needing a different way to demonstrate the same level of performance as demonstrated by other students were students with significant motor disabilities, disabilities that hypothetically would prevent them from responding to a paper and pencil or computer-based test.

Use of AA-GLAS in States

Only one state developed and implemented an AA-GLAS that was considered by the U.S. Department of Education to meet the requirements for a technically adequate assessment of GLAS. That state used a portfolio approach to hold students to the same achievement standards as students taking the general assessment.

Clarification of the Qualities of AA-GLAS

Defining the qualities necessary for an AA-GLAS is more difficult than defining the technical qualities of a more traditional assessment. Further, with the push in ESSA to include in the general assessment items that are delivered in the form of projects, portfolios, and extended performance tasks, the distinction of an AA-GLAS disappeared because all students were considered to have access to these different procedures for assessing students’ knowledge and skills.

AA-MAS

In 2007, modified achievement standards were introduced through federal regulation. They were defined as reduced, less difficult expectations for students with disabilities on challenging assessments aligned to grade-level content. AA-MAS were presented as an optional assessment that states could elect to develop for a small group of students with disabilities. With this option, states could count up to 2% of all students as proficient who met the AA-MAS proficiency standards. Because of this, it was sometimes called the 2% assessment. The allowance for the AA-MAS was rescinded through federal regulation in August 2015, just months before the reauthorization of ESEA eliminated it as an optional assessment for states to develop for students with disabilities.

Students Intended to Take the AA-MAS

Defining the students for whom an AA-MAS was appropriate was challenging

Defining the students for whom an AA-MAS was appropriate was challenging for states. The AA-MAS regulation indicated that students with disabilities who participate in an AA-MAS could be from any disability category. Further, the students were described as ones who had access to quality grade-level instruction but who were unlikely to achieve grade-level proficiency within the time period covered by their IEPs.

Use of AA-MAS in States

The number of states that developed and implemented AA-MAS varied over the years when it was allowed for ESEA accountability. In 2012, 5 years after states were first allowed to develop this optional assessment, there were 12 states using it for some of their students with disabilities. The participation policies in these states differed, but most of them included previous poor performance on state assessments, or on state assessments and other measures, for defining which students should participate in the AA-MAS.

Clarification of the Qualities of AA-MAS

AA-MAS were to meet the same quality requirements as other assessments in which students with disabilities participate. These requirements were difficult for states to meet because of the difficulty identifying less difficult but challenging performance for students with disabilities. Research confirmed this difficulty. Considerable evidence was accumulated that indicated the lowest performing students with disabilities often were assigned to the AA-MAS in 1 year, then to the AA-AAS in another year, and sometimes to the general assessment in another year.

Evidence of the Consequences of AA-AAS for Students With Significant Cognitive Disabilities

The AA-AAS, which as of 2015 was the only alternate assessment to continue to be allowed for purposes of federal ESEA accountability, has resulted in significant changes in understanding the characteristics of students with significant cognitive disabilities and in providing grade-appropriate academic instruction to these students. These changes, in turn, have affected understanding of how to best assess grade-level academic content for these students.

With the development of new, more rigorous AA-AAS and the collection of data on the students participating in the assessment, the characteristics of students with significant cognitive disabilities were examined. These examinations revealed that most students with significant cognitive disabilities communicate with symbolic-level skills, both receptively and expressively, and also responded to social interactions. Most students with significant cognitive disabilities also were able to read text with basic understanding and compute numbers, either with or without a calculator. Further, most had normal vision, hearing, and motor function, with or without correction. Fewer than 10% of the students who participated in the AA-AAS were viewed as communicating primarily through cries, facial expressions, or changes in muscle tone or as having no social interactions with others or even not being aware of them.

Unlike assessments developed in the late 1990s that focused primarily on functional skills, by 2015 states' AA-AAS focused on grade-level English language arts, mathematics, and science content, with standards set to reflect alternate achievement of the content. Students with significant cognitive disabilities were being held to much more rigorous expectations and in general tended to be on the path to meeting those expectations.

Martha L. Thurlow

See also Accommodations; Every Student Succeeds Act; Individuals with Disabilities Education Act; No Child Left Behind Act

Further Readings

Kleinert, H., & Kearns, J. Alternate assessment for students with significant cognitive disabilities: An educator's guide. Baltimore, MD: Paul H. Brookes.

Lazarus, S. S., Thurlow, M. L., Ysseldyke, J. E., & Edwards, L. M. (2015). An analysis of the rise and fall of the AA-MAS policy. Journal of Special Education, 48(4), 231–242.

Perie, M. (Ed.). (2010). Alternate assessments based on modified achievement standards (AA-MAS): Research, best practices and recommendations for their design and development. Baltimore, MD: Paul H. Brookes.

Thurlow, M. L., Lazarus, S. S., & Bechard, S. (Eds.). (2013). Lessons learned in federally funded projects that can improve the instruction and assessment of low performing students with disabilities. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., & Quenemoen, R. F. (2016). Alternate assessments for students with disabilities: Lessons learned from the National Center and State Collaborative. In C. Wells & M. Faulkner-Bond (Eds.), Educational measurement: From foundations to future. New York, NY: Guilford.

Wiener, D. (2006). Alternate assessments measured against grade-level achievement standards: The Massachusetts “competency portfolio” (Synthesis Report 59). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Gabriel J. Merrin Gabriel J. Merrin Merrin, Gabriel J.

American Educational Research Association

American educational research association

70

72

American Educational Research Association

The American Educational Research Association (AERA) is the nation's largest professional organization dedicated to education research. Founded in 1916, the primary focus of AERA is to facilitate the creation of rigorous education research for the improvement of educational practices, experiences, and outcomes. Education research is examining the process of education and learning throughout the life span while considering individual and contextual differences. This entry further describes AERA, including its structure and function, and then discusses its Graduate Student Council (GSC) and annual meeting.

AERA has a total membership of approximately 25,000, including researchers, students, and practitioners from around the world. Although the majority of the association is made up of education researchers (approximately 74%), some members conduct research in other disciplines, including psychology, history, philosophy, statistics, anthropology, sociology, and political science. Although AERA is more than a century removed from its inception, the central mission of supporting, advancing, and disseminating education research to improve educational processes and influence public policy has remained the fundamental focus for over a century.

Structure and Function

The governance structure of AERA includes four main units: the council, executive board, standing committees, and annual committees. There are 12 divisions that represent various areas of educational research:

Division A: Administration, Organization, and Leadership
Division B: Curriculum Studies
Division C: Learning and Instruction
Division D: Measurement and Research Methodology
Division E: Counseling and Human Development
Division F: History and Historiography
Division G: Social Context of Education
Division H: Research, Evaluation, and Assessment in Schools
Division I: Education in the Professions
Division J: Postsecondary Education
Division K: Teaching and Teacher Education
Division L: Educational Policy and Politics

In addition, AERA has over 160 special interest groups that represent subfields within education research, with groups focusing on measurement and assessment in higher education and on leadership for social justice. Generally, members belong to one or two of the 12 divisions and a few special interest groups of their preference.

The association publishes six peer-reviewed journals including *Educational Researcher* and the *Journal of Educational and Behavioral Statistics*. AERA also issues a free monthly online newsletter named *AERA Highlights*, which keeps readers current with AERA-related news and initiatives related to education. Additional to the six journals and monthly newsletter, AERA publishes books on timely and prominent educational topics. Books published by AERA include *Standards for Educational and Psychological Testing* and *Prevention of Bullying in Schools, Colleges, and Universities*.

To encourage scholarship and stimulate change in specific areas, AERA organizes targeted programs. These include the government relations program, which engages with federal agencies encouraging funding of education research, and the social justice program, which supports and disseminates scholarship on issues of social justice in education. AERA also takes a stand on social issues by releasing position statements. Some notable examples include position statements on the 2015 Charleston, South Carolina, church shootings and racism in the United States and on the use of value-added models to evaluate educators and educator preparation programs. Furthermore, the annual *Brown Lecture*, started in 2004 to commemorate the *Brown v. Board of Education* decision, highlights the importance of research in the pursuit of equality in education and demonstrates the organization’s effort to take firm positions on various education-related issues.

Graduate Student Council

More than 28% of AERA’s membership consists of students including 6,500 graduate students and 600 undergraduate students. As such, several resources devoted to student growth and development as well as numerous opportunities for student involvement. The GSC is a student-run council that facilitates and supports the development and transition from graduate student to professor or practitioner. By helping students navigate many obstacles and challenges of the academy, the GSC advocates for and serves the needs of students.

The GSC is made up of nine council members and 24 division representatives. The council’s role is to support all students across the entire association, whereas the division representatives’ role is to support the students within each of their respective divisions. The council has four elected positions including chair elect, secretary/historian, newsletter editor, and web secretary and two appointed positions including program chair and community leader. Each of the 12 divisions have two representatives, a junior and senior representative.

The GSC has five major responsibilities that consist of planning the annual meeting for students, advocating for student needs, disseminating information, community building, and self-governance. The GSC is governed entirely by students, who host 23 sessions at each annual meeting created for students. In addition, the GSC holds an annual community service project with an organization in the local community where the conference is held.

Annual Meeting

The annual AERA conference is a 5-day meeting held in either the United States or Canada. Approximately 14,000 researchers travel to the conference each year, attending hundreds of sessions dedicated to presenting the latest education research across various education disciplines. Invited presidential sessions on a current and major issue in education are typically given by prominent scholars or public figures. Respective divisions host several sessions in various formats that include paper symposiums, roundtable discussions, fireside chats, and poster sessions.

Prior to the annual meeting, various divisions host a preconference for graduate students and early career scholars, focusing on topics such as the job search, grant writing, and tenure review. For individuals interested in measurement and statistics, the annual meeting is held in accordance with National Council on Measurement in Education conference, and AERA members frequently attend both the conferences. The large variety and volume of topics and sessions offered at the annual conference provide several options for all AERA attendees.

Gabriel J. Merrin

See also American Evaluation Association; American Psychological Association; *Brown v. Board of Education*; Educational Researchers, Training of; National Council on Measurement in Education; Value-Added Models

Further Readings

American Educational Research Association. (n.d.). Graduate student council. Retrieved from www.aera.net/About-AERA/Member-Constituents/Graduate-Student-Council

American Educational Research Association. (n.d.). Featured education research jobs. Retrieved from careers.aera.net

Mershon, S., & Schlossman, S. (2008). Education, science, and the politics of knowledge: The American Educational Research Association, 1915–1940. American Journal of Education, 114(3), 307–340.

Websites

American Educational Research Association: www.aera.net

Jennifer A. Brussow Jennifer A. Brussow Brussow, Jennifer A.

American Evaluation Association American evaluation association

American Evaluation Association

The American Evaluation Association (AEA) is a nonprofit international professional association for evaluators. AEA’s publications, conferences, and topical interest groups (TIGs) deal with program evaluation, personnel evaluation, and other forms of evaluation designed to assess the strengths and weaknesses of programs, policies, personnel, and organizations. As of January 2016, AEA comprised approximately 7,000 members from all 50 U.S. states and over 60 other countries. Members include evaluators, researchers, educators, students, and stakeholders. This entry provides an overview of AEA’s creation and mission; organization; establishment of professional guidelines for evaluators; and professional development opportunities, collaboration with other organizations, and awards for members.

Creation and Mission

AEA was formed in 1986 as a result of the merger between the Evaluation Research Society and Evaluation Network. Its mission is “to improve evaluation practices and methods, increase evaluation use, promote evaluation as a profession, and support the contribution of evaluation to the generation of theory and knowledge about effective human action” (AEA, n.d.).

AEA values high-quality, ethical, culturally responsive evaluations that are intended to improve the evaluated entities’ effectiveness. It seeks to develop an international, diverse, and inclusive evaluation community in order to provide continual development opportunities for evaluation professionals to deepen their understanding of evaluation practices and methodologies. To these ends, AEA’s goals are to ensure that evaluators have the skills necessary to be effective, culturally competent, contextually sensitive, and ethical; to provide a sense of professional affiliation between evaluators; to increase evaluation’s visibility and perceived value as a field; to create informed policy so that communities and organizations can participate in and learn from evaluation; and to ensure that members value their membership.

Organization

AEA is led by a 13-member board of directors responsible for programmatic decisions for the association. Included on the board are four principal officers, namely, a president, a president-elect, a past president/secretary, and a treasurer, who are nominated from and elected by the membership. Additionally, over 50 TIGs provide networking opportunities and conference programming surrounding their particular interests.

As of January 2016, the five TIGs with the most members were nonprofit and foundations evaluation; independent consulting; organizational learning and evaluation capacity building; collaborative, participatory, and empowerment evaluation; and qualitative methods. A complete list of TIGs can be found on the AEA website. In addition to TIG membership, members can also provide input by volunteering in a variety of working groups that coordinate various aspects of the association’s activities. AEA also has numerous affiliated local and professional associations recognized as having similar missions. AEA’s bylaws outline the organization’s legal obligations as a nonprofit entity. The most recent edition of the bylaws as of January 2016 was the one that took effect in January 2011.

Professional Guidelines for Evaluators

The Evaluation Research Society had adopted a set of standards for program evaluation in 1982, but no standards or guidelines were officially adopted by AEA when it was formed. In 1992, a task force was created to draft a set of guiding principles for evaluators. This task force consisted of William Shadish, Dianna Newman, Mary Ann Scheirer, and Christopher Wye. An initial draft was sent to all AEA members in 1993, and a final draft was approved in 1994, resulting in the *Guiding Principles for Evaluators*, a general set of principles to inform evaluators’ practice in the field. The principles were reviewed and revised in a process throughout 2002 and 2003, and revisions were accepted by AEA membership in 2004.

As of January 2016, the 2004 version of the principles was the most recent. The AEA endorses five guiding principles: systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare. These principles are written as broad guidelines intended to guide evaluators' professional practice and apply to all types of evaluation, and they are not intended to serve as professional standards. Each of the overarching principles listed here has three to seven subprinciples. A complete listing of the principles can be found on the AEA website.

Activities

Since its inception in 1986, AEA has sponsored an annual conference called evaluation, which features presentations within various topical strands. It also offers a 3-day summer evaluation institute that provides professional development and training sessions. AEA publishes two journals: the *American Journal of Evaluation* and *New Directions for Evaluation*. Membership in AEA also includes access to *Evaluation Review* and *Evaluation and the Health Professions*.

AEA also maintains a blog with daily evaluation tips, a series of webinars that provide information on evaluation tools, virtual professional development courses, member discussion groups, and a newsletter. AEA contributes a representative to the Joint Committee on Standards for Educational Evaluation, which issues the *Educational Evaluation Standards*, the *Program Evaluation Standards*, and the *Personnel Evaluation Standards*. As of January 2016, AEA offers eight awards to recognize individuals in the categories of promising new evaluators, service, evaluation advocacy and use, evaluation practice, evaluation theory, outstanding evaluation, enhancing the public good, and research on evaluation. AEA also offers several fellowships for graduate students and faculty members and has a program in which it collaborates with evaluation organizations in other countries.

Jennifer A. Brussow

See also Evaluation Capacity Building; Evaluation; Formative Evaluation; Personnel Evaluation; Program Evaluation; Summative Evaluation

Further Readings

American Evaluation Association. (n.d.). About AEA. Retrieved from http://www.eval.org/p/cm/ld/fid=4

Kingsbury, N. (1986). Coming together: Evaluation network and evaluation research society share common business agendas at Evaluation'85 leading to the American Evaluation Association. American Journal of Evaluation, 7(1), 107–110. doi:10.1177/109821408600700118

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.). Thousand Oaks, CA: Sage.

Websites

American Evaluation Association: www.eval.org

Jonathan A. Plucker Jonathan A. Plucker Plucker, Jonathan A.

Lorraine Blatt Lorraine Blatt Blatt, Lorraine

American Psychological Association American psychological association

73

75

American Psychological Association

The American Psychological Association (APA) is an organization dedicated to advancing the field of psychology and using psychology to contribute to a wide range of issues facing society. APA was founded in 1892 when psychology was just developing as a discipline, primarily as an outgrowth of philosophy. This entry discusses the activities and structure of APA and how it has contributed to the field of education, in particular, through its education directorate and collaborations with other organizations.

APA has grown from 31 members at its founding to over 117,500 members in 2016. APA members include psychology researchers, educators, clinicians, consultants, and students who span 54 different divisions of psychology. APA is divided into four directorates that each focus on a topic critical to APA’s strategic plan: practice, public interest, science, and education.

The directorates engage in specific efforts toward research, advocacy, policy, and outreach that meet the various goals of each directorate. The practice directorate aims to increase awareness of and access to mental and behavioral health services in addition to developing and maintaining guidelines for practitioners and recipients of psychological services. The public interest directorate is focused on combating inequality and promoting social justice and human welfare. The science directorate seeks to support the discipline of psychology in a variety of ways, including providing training and funding for those studying or working in the field of psychology. Finally, the education directorate works both to improve psychology education and to apply valuable psychological research findings to educational practices.

Education at APA

The mission of the education directorate is to advance “the science and practice of psychology for the benefit of the public through educational institutions, programs, and initiatives” (APA, n.d.). It seeks to achieve this goal by supporting both education within psychology and the application of psychology to education. Regarding education within psychology, the directorate serves as the national accreditation organization for training in psychology through the APA Commission on Accreditation and promotes and monitors continuing education for psychologists. In addition, the directorate supports the teaching of psychology in high school through the Teachers of Psychology in Secondary Schools. Furthermore, the Center for Workforce Studies produces reports on the status of the psychology profession, including reports that provide data on the presence of psychologists in higher education.

Regarding the application of psychology to education, the directorate sponsors and conducts a range of activities, including the creation of resources and modules for teachers on topics such as student learning and diversity, the nature and enhancement of creativity, and student behavior and classroom management. One such resource is the Top 20 principles from psychology for PreK–12 teaching and learning, a report that describes 20 principles about teaching and learning drawn from the psychological research literature. The principles are organized into five categories including student thinking and learning; motivation; the relationship of social context, interpersonal relationships, and emotional well-being to learning; classroom management; and assessment. The Top 20 principles report has been translated into several languages and is used in both K–12 schools as a professional development resource and within college courses as a reading on advances in educational psychology.

Many of the education directorate’s application-focused activities are products from the Center for Psychology in Schools and Education and special working groups such as the Coalition for Psychology in Schools and Education and the Coalition for High Performance. A mix of APA and external funding supports these groups.

Cutting across both areas of effort—psychology education and the application of psychology to education—the education directorate advocates for policy and funding for psychological science and education and maintains a robust outreach effort, including a strong social media presence.

However, not all education-related activities at APA occur within the education directorate, as other directorates and several APA divisions focus on education as well. In particular, Division 15 (Educational Psychology) has a strong K–12 focus, and to a lesser extent, Division 10 (Society for the Psychology of Aesthetics, Creativity, and the Arts) is involved in K–12 education. These and other divisions publish journals featuring education research, hold mini-conferences for researchers and practitioners, and provide resources for educators. Other organization-wide activities also have an impact on education, such as the development of an ethics program and corresponding educational programs and resources to help practitioners understand the ethical code of conduct.

Additionally, APA collaborates with other education-focused organizations, with a good example being the Standards for Educational and Psychological Testing, a project of the American Educational Research Association, APA, and National Council on Measurement in Education. The standards, which have been published jointly by these professional organizations since 1966, have become the professional standards for educational assessment and are used in several countries.

Jonathan A. Plucker and Lorraine Blatt

See also American Educational Research Association; *Standards for Educational and Psychological Testing*

Further Readings

American Educational Research Association, American Psychological Association, … National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Psychological Association. (n.d.). Education Directorate. Retrieved from http://www.apa.org/ed/

Lucariello, J., Graham, S., Nastasi, B., Dwyer, C., Skiba, R., Plucker, J. A., &

Lee, G. M. (2015). Top 20 principles from psychology for preK–12 teaching and learning. Washington, DC: American Psychological Association, Center for Psychology in Schools and Education.

Joseph Madaus Joseph Madaus Madaus, Joseph

Walter Keenan Walter Keenan Keenan, Walter

Salome Heyward Salome Heyward Heyward, Salome

Americans with Disabilities Act Americans with disabilities act

75

77

Americans with Disabilities Act

The Americans with Disabilities Act of 1990 (ADA) is a civil rights law that prohibits discrimination against individuals with disabilities, including students in K–12 and higher education. The law is an extension of Section 504 of the Rehabilitation Act of 1973 (Section 504), which prohibits discrimination against those with disabilities in programs receiving federal funding. This entry first explains the ADA, then describes its impact on K–12 education, institutions of higher education, and standardized examination and high-stakes testing agencies.

To be eligible for protection under the ADA, an individual must have a physical or mental impairment that substantially limits a major life activity. In the years following the law's passage, several federal courts, including the U.S. Supreme Court, narrowly construed this definition of disability, resulting in limited coverage to individuals with disabilities. Congress responded by passing the Americans with Disabilities Act Amendments Act of 2008 (ADAAA), which expanded the definition of disability to ensure broad coverage.

The ADAAA made clear that to be considered a substantial limitation, an impairment need not prevent or significantly restrict the ability to perform a major life activity. It also expanded the definition of major life activities, prohibited the consideration of ameliorative effects of mitigating measures when determining disability status (except for ordinary eyeglasses and contact lenses), and expanded the definition of auxiliary aids and services necessary to assist individuals with disabilities. The ADAAA states that individuals must be provided with reasonable accommodations or modifications to ensure participation in programs. However, an accommodation that results in a fundamental alteration of the program or undue burden is not considered reasonable.

Impact on K–12 Education

The ADA mandates protection against discrimination for students with disabilities in Grades K–12. While many K–12 students with disabilities are eligible for special education services under the Individuals with Disabilities Education Act (IDEA), the educational progress of other students with disabilities might not be impacted to a level that services are required under IDEA. However, these students are still eligible for protection from discrimination under the ADA and may be eligible for individually appropriate accommodations and auxiliary aids. Additionally, ADA regulations may provide services beyond what is required under IDEA. For example, the ADA standard concerning communication may require auxiliary aids for a deaf student beyond what is required under IDEA. It should be noted that a student who is eligible for services under the IDEA is also eligible for protection under the ADA and Section 504; however, as noted, not all students covered under the ADA and Section 504 are eligible for services under the IDEA.

The ADAAA legislation made clear that determination of coverage does not demand extensive analysis. A school should first evaluate whether the student requires special education services and then determine whether the student is entitled to reasonable modifications of policies, practices, or procedures even if special education services are not necessary. For example, a student with attention-deficit/hyperactivity disorder may require modification to length of homework assignments.

Students with disabilities may not be unnecessarily segregated from other students. Programs or services that segregate students must provide opportunities for integration with students without disabilities to the maximum extent appropriate. The ADA is not limited to educational activities but rather applies to all services, programs, and activities provided by the school district. Additionally, the ADA applies to private as well as public schools.

Institutions of Higher Education

Institutions of higher education are obligated to provide reasonable accommodations and auxiliary aids to ensure equal access to postsecondary education programs for students with disabilities. Unlike at the K–12 level, postsecondary students must self-report disability, provide documentation of disability, and request accommodations through the appropriate campus disability contact person. Accommodations are then determined by assessing the impact of the documented disability on the ability to participate in the educational program.

The ADAAA stipulates that reading, writing, thinking, speaking, concentrating, and communicating are major life activities and that previous academic achievement does not necessarily mean that a student does not experience substantial limitation of a major life activity. Instead, the condition, manner, or duration it takes an individual to perform an activity as compared to most people in the general population should be considered. For example, a person with a learning disability will often be substantially limited in learning, reading, and thinking as compared to most people.

Examples of reasonable accommodations include extra time on exams, screen readers, note takers, audio lecture recordings, and reduced course load. Because equal access to information is necessary to participate in postsecondary education, textbooks, readings, and website information must be available in formats that are compatible with common adaptive technology as higher education incorporates more online instruction and provision of course information. Accommodations that fundamentally alter essential academic requirements, impose an undue burden, or impose a direct threat to the health or safety of the student or others are not reasonable.

Standardized Examination and High-Stakes Testing Agencies

The ADA applies to agencies that provide standardized exams and high-stakes tests for applications, licensing, certification, or credentialing for secondary, postsecondary, professional, or trade purposes. These entities must provide accommodations and auxiliary aids to individuals with disabilities to best ensure results accurately reflect aptitude or achievement levels rather than an individual’s impairment. Documentation required to obtain testing accommodations must be reasonable and limited to the need for requested accommodations

accommodations.

Proof of past testing accommodations in similar test settings is generally sufficient to support current accommodations. A candidate should generally receive the same testing accommodations previously received under the IDEA or Section 504 in Grades K–12 or received in postsecondary education without requiring further documentation. An absence of prior formal testing accommodations does not preclude a candidate from receiving accommodations. Submission of documentation from qualified professionals based on evaluation and careful consideration should be sufficient documentation. Agencies must respond in a timely manner to requests for accommodations, and applicants should have a reasonable opportunity to provide additional documentation when needed. Annotating or “flagging” accommodated test scores is prohibited.

Joseph Madaus, Walter Keenan, and Salome Heyward

See also Ability–Achievement Discrepancy; Accessibility of Assessment; Accommodations; Intellectual Disability and Postsecondary Education; Learning Disabilities; Special Education Identification; Special Education Law

Further Readings

Heyward, S. (2011). Legal challenges and opportunities. New Directions for Higher Education, 154, 55–64.

U.S. Department of Education—Office of Civil Rights. (2012). Questions and answers on the ADA amendments act of 2008 for students with disabilities attending public elementary and secondary schools. Retrieved January 10, 2016, from http://www2.ed.gov/about/offices/list/ocr/docs/dcl-504faq-201109.html

U.S. Department of Justice—Civil Rights Division—Disability Rights Section. (2009). A guide to disability rights laws. Retrieved January 10, 2016, from http://www.ada.gov/cguide.htm

U.S. Department of Justice—Civil Rights Division—Disability Rights Section. (2014). ADA requirements—Testing accommodations. Retrieved January 10, 2016, from http://www.ada.gov.regs2014/testing_accommodations.html

Legal Citations

Americans with Disabilities Act, 42 U.S.C. §§ 1210 1 *et seq*.

Individuals with Disabilities Education Act, 20 U.S.C. §§1400 *et seq*.

Daniel Tan-lei Daniel Tan-lei Shek Shek, Daniel Tan-lei

Li Lin Li Lin Lin, Li

Amos

Amos

77

81

Amos

Amos is a computer program for performing structural equation modeling (SEM) and mean and covariance structure analysis. Its full name is IBM SPSS Amos, with Amos standing for “analysis of moment structures.” It was developed by the Amos Development Corporation, which is now owned by the IBM SPSS Corporation. Because of its easy-to-use functions, it has become a popular SEM program. Many educational researchers use it to validate measures and test hypotheses. This entry describes the basic features and functions of Amos and illustrates its application in education research.

Features and Functions of Amos

Amos includes a graphical interface (Amos Graphics) and a nongraphical programmatic interface (Program Editor). While one can work directly on a path diagram in Amos Graphics, one can work directly on equation statements using syntax in Program Editor. Amos Graphics offers users a palette of tools and drop-down menus for analysis, while Program Editor provides a platform for analysis using VB.NET or C# scripts. The Amos package also includes a file manager, a seed manager for recording seed values in simulations of random sampling (e.g., bootstrapping), a data file viewer, and a text output viewer.

Amos is capable of performing confirmatory factor analysis (CFA), path analysis, multigroup analysis, multitrait–multimethod model, and multilevel analysis (e.g., latent growth curve model). In education research, Amos has commonly been used to (a) validate the factorial structure of an educational assessment instrument (single-group CFA), (b) test the measurement equivalence of a scale across different groups (multigroup CFA), (c) test a theoretical model (path analysis), and (d) examine the developmental trajectory of learning attributes (latent growth curve model).

Amos provides estimation with full information maximum likelihood to handle missing data, which is common in education research. Rather than imputing missing values, full information maximum likelihood estimates a likelihood function for each individual case based on the information from all the observed proportion of data. Full information maximum likelihood is theoretically robust and outperforms ad hoc methods such as listwise deletion, pairwise deletion, or mean imputation for addressing incomplete data. Besides conducting SEM with data that are measured on a continuous scale with multivariate normal distribution, Amos also provides Bayesian estimation to fit for ordered categorical data and allows users to conduct bootstrapping to tackle nonnormality.

With its graphic interface, Amos allows users who have little statistical knowledge of SEM to perform analysis efficiently. However, as researchers should take responsibility to conduct appropriate data analyses, it is highly recommended that they acquire an understanding of the corresponding concepts and practices of SEM. Amos can be purchased from the IBM SPSS website with its user’s guide free to download.

Illustration of Amos Applications

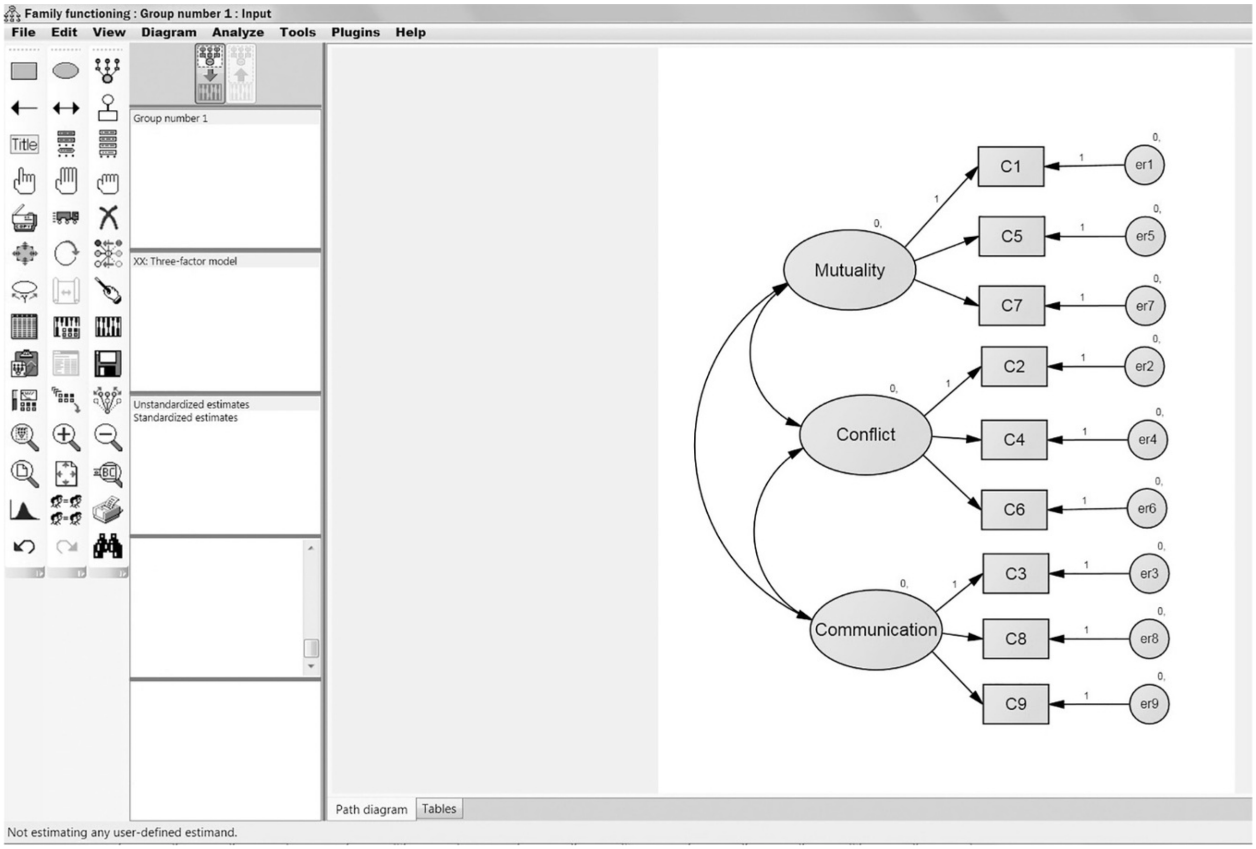
To demonstrate the application of Amos, this section provides an example of the use of Amos Graphics to perform CFA. The goal of this analysis is to validate the factorial validity of a scale measuring family functioning, which is often linked to student well-being. In this scale, it is hypothesized that three components of family functioning—family mutuality, family conflict, and family communication—are assessed by 3 items, respectively. These three components are theoretically correlated with each other. A sample of 1,000 seventh graders was used. Usually, the procedure of SEM analysis includes five steps: (1) model specification, (2) data specification, (3) calculation of estimates, (4) model evaluation, and (5) model modification (if necessary).

Model Specification and Data Specification

Model Specification and Data Specification

The first step of CFA was to construct a hypothesized model and then read a data set. A CFA model was drawn by using Indicator icon to create three latent factors (indicators) with three observed variables each and Covariate icon to create three covariance paths between the latent factors. To meet the demand of identification of model that each indicator must have a scale, Amos creates an indicator model with one factor loading automatically set to be 1. Next, a data file was imported by clicking on Data icon. Amos reads data in several database formats, including Microsoft Excel spreadsheets and SPSS databases, and text. Finally, the observed variables, factors, and measurement errors were labeled (Figure 1).

Figure 1 An illustration of the confirmatory factor analysis (CFA) path diagram

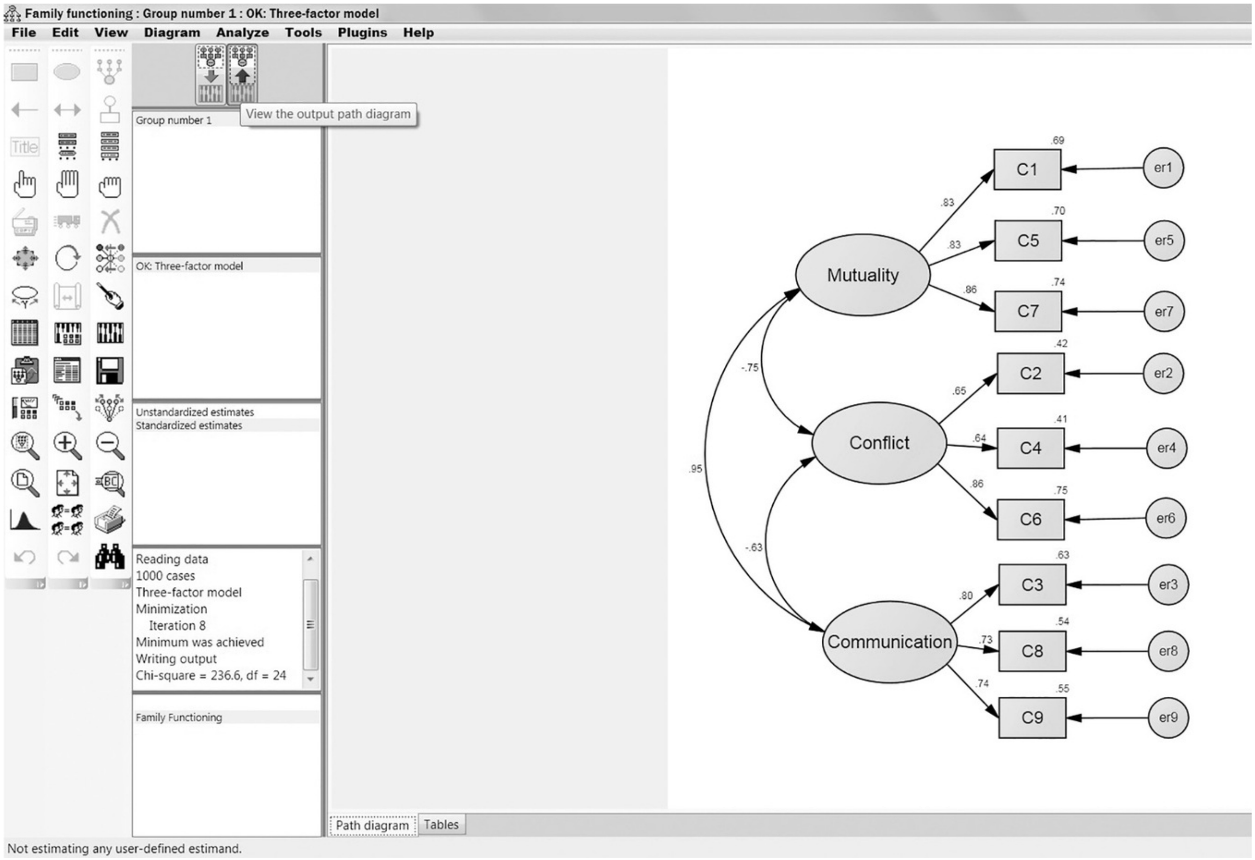


Not estimating any user-defined estimand.

Calculation of Estimates

The maximum likelihood estimation approach was chosen in this case. Amos also offers users other approaches including unweighted least squares, generalized least squares, Browne’s asymptotically distribution-free criterion, scale-free least squares, and Bayesian estimation. By clicking on the *Calculate Estimates* icon, the results were generated. The estimation results can be viewed in Amos Output while the estimates of parameters can be viewed on the screen by clicking on the *View Output Path Diagram* icon (Figure 2). The Amos Output shows three sets of information: model summary, model variables and parameters, and model evaluation.

Figure 2 Standardized parameter estimates of the confirmatory factor analysis (CFA) model



```
Reading data
1000 cases
Three-factor model
Minimization
   Iteration 8
Minimum was achieved
Writing output
Chi-square = 236.6, df = 24
```

Model Evaluation

Usually, a model will be evaluated based on the adequacy of the parameter estimates and the model as a whole. First, parameter estimates were evaluated based on three criteria:

1. Feasibility of the parameter estimates: any incorrect sign or value?
2. Appropriateness of the standard errors: any error that is very small (i.e., close to zero) or very large?
3. Statistical significance of the parameter estimates: any statistically insignificant parameter estimate in regression weights, intercepts, covariances, and variances which may be regarded as unimportant?

There was no specific problem with the parameters of this example.

Next, goodness of fit between the hypothesized model and the sample data was evaluated by referring to the model fit indices. In Amos, model fit is reported for the hypothesized model, as well as two additional models: saturated model and independence model. The saturated model is the most general model without any independence model. The saturated model is the most general model without any constraints placed on the population moments. In contrast, the independent model is the most restricted model with all the observed variables assumed to be uncorrelated with each other. An ordinary hypothesized model should lie in between these two models with better model fit. Amos provides eight groups of fit measures, as follows.

1. Minimum discrepancy between hypothesized model and sample data: for example, CMIN/DF (chi-square (*df*)), *p* (*p* value for chi-square test)
2. Measures based on the population discrepancy: for example, root mean square of error of approximation (RMSEA).
3. Incremental indices—comparative indices with comparison to a baseline model: for example, normed fit index, Tucker–Lewis index (TLI; i.e., nonnormed fit index), and comparative fit index (CFI).
4. Measures of parsimony—evaluating the simplicity of the model: for example, parsimony ratio.
5. Parsimony-adjusted measures: for example, parsimony-adjusted normed fit index
6. Information theoretic measures used for model comparison: for example, Akaike information criterion and Bayesian information criterion.
7. Goodness-of-fit index and related measures: for example, goodness-of-fit index and adjusted goodness-of-fit index.
8. Miscellaneous measures: for example, root mean square residual.

Conventionally reported fit measures include CMIN/DF ($\chi^2(df)$), CFI, TLI, and RMSEA. According to the rule of thumb, CMIN/DF > 2.00 represents a poor fit. For CFI and TLI, values > .90 represent an acceptable fit and > .95 a good fit. For RMSEA, values < .08 represent an acceptable fit and < .05 a good fit. Nonetheless, due to high sensitivity to large sample size, CMIN/DF is often used in model comparison rather than single-model evaluation. For the current example, results of the different indices are as follows: $\chi^2(24) = 9.860$ (not good), CFI = .957 (good), TLI = .919 (acceptable), and RMSEA = .094 (not good). Further action can be taken to improve the model fit.

Model Modification

Amos provides modification indices (MIs) to detect model misspecification. MI represents the expected decrease in overall CMIN (χ^2) value if one certain parameter is to be freely estimated in a subsequent run. Users can decide whether they want to modify the model when an MI value is large. However, any modification should be theoretically justifiable. MI cannot be computed with missing data, which is an obvious limitation of Amos. Rerunning the program using a sample with expectation–maximization imputation for the missing data revealed that the largest MI value rested in the correlation between errors of Items 8 and 9, MI = 118.190. This result suggested that Item 8 and Item 9 measured an additional construct that was not represented by the latent factor—family communication. When the two errors were allowed to be correlated, the model fit increased: $\chi^2(23) = 4.123$ (not good), CFI = .986 (good), TLI = .978 (good), and RMSEA = .056 (acceptable). Although correlating errors usually leads to improved fit, it is possibly at the cost of biased estimates of model.

Other Applications

In addition to CFA, Amos is capable of testing more complex models. When the covariances among latent factors are hypothesized to be explained by another factor, one can test a higher order CFA model. Daniel Tan-lei Shek and Lu Yu demonstrated how to test a second-order CFA model by assuming three latent factors (i.e., perception of program, perception of implementers, and perceived effectiveness of program) to be explained by a single higher order factor of subjective evaluation of program. They added a latent factor by using *Oval* icon, linked the first-order factor and second-order factor by adding a single-headed arrow from second-order factor to first-order factor (using *Path* icon), and finally, added a factor disturbance to each first-order factor by using *Error* icon, as the first-order factors become endogenous factors while the second-order factor becomes an exogenous factor.

In addition, Amos is often used to test factorial invariance of a measure. As illustrated by Yu and Shek, in 2014, full invariance includes configural invariance (i.e., invariance of factorial structure), structure invariance (i.e., invariance of factor covariance and factor variance), and measurement invariance (i.e., invariance of factor loadings, intercepts, and errors). To test the factorial invariance, users can use an important function of Amos—comparing nested models (i.e., a pair of models in which one can be obtained by constraining the parameters of the other). Users first need to create nested models using the function of “manage models,” post constraints on the parameters in the constrained model via simple syntax, and finally compare the constrained model with the unconstrained model.

With the two models being estimated simultaneously, Amos provides comparison of model fit in terms of CMIN (χ^2), normed fit index, IFI, RFI, TLI, and χ^2 difference test, in which researchers often reply on χ^2 difference test for judgment. In their case, when the factor loadings of the subscale of perceived effectiveness of program were constrained to be equal across program implementers of three grades of secondary school, $\Delta\chi^2$ increased relative to that of the unconstrained model, but the increase is not statistically significant. This finding suggested that the subscale was metric invariant across the samples. If the constraints lead to a significant increase in χ^2, it would suggest that the constraints make the model fit worse, and thus, the invariance cannot be established.

Amos can also perform path analysis with or without latent factors. For example, in 2014, in Yu and Shek's illustration of path analysis, they tested a mediation model (i.e., family functioning predicts Internet addiction directly and via the effect of positive youth development) and compared it with several alternative models. For instance, in one alternative model, family functioning has an indirect effect on Internet addiction via positive youth development without a direct effect. Amos provides information on direct effects, indirect effects, and total effects in the output. The general procedure of path analysis is similar to that of CFA mentioned earlier, yet there are several issues worth noting. First, it is convenient to test alternative models with different relationship among variables in Amos, whereas the alternative models should be conceptually meaningful. Second, if latent factors are used, the measurement model of the latent factor should be tested (i.e., CFA) before performing path analysis. Third, Amos provides bootstrapping to confirm the mediation effect in the path analysis.

Daniel Tan-lei Shek and Li Lin

See also Bayesian Statistics; Bootstrapping; Confirmatory Factor Analysis; Measurement Invariance; Path Analysis; SPSS; Structural Equation Modeling

Further Readings

Arbuckle, J. (2013). Amos 22.0 user's guide. Chicago, IL: SPSS.

Blunch, N. J. (2008). Introduction to structural equation modeling: Using SPSS and Amos. Thousand Oaks, CA: Sage.

Byrne, B. M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. International Journal of Testing, 1(1), 55–86.

Byrne, B. M. (2013). Structural equation modeling with AMOS: Basic concepts, applications, and programing. New York, NY: Routledge.

Kline, R. B. (2010). Principles and practice of structural equation modeling (3rd ed.). New York, NY: Guilford Press.

Shek, D. T. L., & Yu, L. (2014). Confirmatory factor analysis using AMOS: A demonstration. International Journal on Disability and Human Development, 13(2), 191–204.

Yu, L., & Shek, D. T. L. (2014a). Family functioning, positive youth development, and internet addiction in junior secondary school students: Structural equation modeling using AMOS. International Journal on Disability and Human Development, 13(2), 227–238.

Yu, L., & Shek, D. T. L. (2014b). Testing factorial invariance across groups: An illustration using AMOS. International Journal on Disability and Human Development, 13(2), 205–216.

Julius Sim Julius Sim Sim, Julius

Analysis of Covariance Analysis of covariance

81

85

Analysis of Covariance

Analysis of covariance (ANCOVA) is a statistical procedure that forms part of the general linear model. Indeed, it can be thought of as a combination of two other methods within this family of statistical models: analysis of variance (ANOVA) and linear regression. It represents the inclusion of a continuous predictor variable (covariate) within a standard ANOVA model, such that values on the outcome variable within the model are adjusted for values on the covariate. There are two main objectives of ANCOVA. First, it can be used in experimental designs to remove the effect of one or more confounding variables. Second, it serves to increase the sensitivity of a statistical test of the experimental factor in the statistical model. This entry discusses the form of the ANCOVA model, the functions of ANCOVA, assumptions of the analysis, using ANCOVA outside experimental contexts, and other considerations in the use of ANCOVA and alternative measures.

Form of the ANCOVA Model

The model for ANCOVA, in the case where there is just a single covariate, is:

$$y_{ij} = \mu + \tau_j + \beta z_{ij} + \varepsilon_{ij},$$

where y_{ij} is the outcome score for participant i in group j; μ is the overall mean score on the outcome variable in the study; τ_j is the effect of the experimental factor in group j; z_{ij} is the covariate score for participant i in group j; β is the regression coefficient for z (estimated from the sample data); and ε_{ij} is the residual for participant i in group j. Note that removing βz_{ij} would leave the basic ANOVA model. Normally, the term , where is the overall mean covariate score within the study, is used rather than z_{ij} so that the constant term in the model is set at the overall mean for the outcome variable, giving this model:

$$y_{ij} = \mu + \tau_j + \beta\left(z_{ij} - \bar{z}\right) + \varepsilon_{ij}.$$

It follows that when ANCOVA is performed, each participant's score is adjusted in relation to the covariate—it is the "hypothetical" score the individual would have if all participants had the mean value of the covariate, . Accordingly, the mean for each group in the experiment is also an adjusted mean. The adjusted mean for group *j* is:

$$\bar{y}_j' = \bar{y}_j - \beta\left(z_j - \bar{z}\right),$$

where and are the adjusted and unadjusted means, respectively, for group *j*; is the mean of the covariate for group *j*; and is the overall covariate mean.

Functions of ANCOVA

Statistical Control

As noted earlier, ANCOVA can be used to remove the confounding effect of an extraneous variable in an experimental study. For example, a study might be set up to examine the effect of two methods of learning on test performance, in which students are randomized to the two methods of learning; the first method is mainly student-centered learning (SCL), whereas the second consists more of teacher-directed learning (TDL). If, however, age is also associated with test performance, and if there is additionally an imbalance in age across the two randomized groups, age is a potential confounding variable. Age would thereby provide an alternative explanation of any between-group difference in test performance that is observed, so that this difference could not be confidently attributed to the different methods of learning. One cannot be sure that the groups would still have differed in terms of test performance if they had not also differed in terms of age. If, however, age is included as a covariate in the statistical model, the students' test scores would be adjusted for age, removing the confounding effect of this variable.

The precise effect of this adjustment will depend upon the magnitude and direction of the correlation between age and test scores and of the imbalance in age across the experimental groups. Suppose first that age is positively correlated with test score—*older* students tend to have *higher* scores, and also that the students in the SCL group are on average older than those in the TDL group. At the end of the study, if we took no account of age, the mean difference in test scores might favor the SCL group (i.e., this group had higher scores on average than the TDL group). However, because the students in this group were older than those in the TDL group, the apparent superior performance of the SCL group could be attributable in part to their age. The effect of the learning method has been confounded by the students' ages—their test scores have been biased upward. If we introduce age as a covariate, the mean age of the two experimental groups is statistically equalized. As the effect of age has been removed, the resulting mean difference will be adjusted downward, and this smaller difference will have a larger associated *p* value (and if the unadjusted mean difference had previously been statistically significant, it might no longer be so after this adjustment). Another way of looking at this is to think of students in the SCL group as having an unfair advantage because of their higher average age and thereby obtaining inflated test scores by comparison with the TDL group. The use of the covariate removes this bias.

Conversely, if the students in the SCL group were *younger* than those in the TDL group, the mean between-group difference in test scores would again be confounded if no account were taken of age. In this instance, however, the superiority of the SCL approach would be biased downward if not adjusted for age. The positive effect of the SCL method on test performance would be counteracted by a negative effect of the students' younger mean age in this group—their test scores would have been biased downward. In other words, instead of starting with an unfair advantage, they would start with an unfair disadvantage. Introducing age as a covariate would remove this bias. The mean between-group difference in test scores would be larger following adjustment and would have a smaller *p* value (and in the process may become statistically significant where it had not been prior to introduction of the covariate).

The process whereby covariates are selected for the purpose of statistical control should be specified in advance of the analysis. Otherwise, the analyst might be tempted to use as covariates, by trial and error, those variables that give the conclusion that the analyst wants. In addition, it should be remembered that steps can be taken at the design stage (e.g., matching or stratified randomization) to control for known potential confounding variables.

Statistical Power

The other primary objective of ANCOVA is to increase the precision of between-group estimates, thereby producing narrower confidence intervals around these estimates and increasing the sensitivity, or power, of a statistical test on the estimates. The *p* value in an ANCOVA model is derived from the *F* ratio and its associated degrees of freedom. The *F* ratio has, as its numerator, the variance in the outcome that is explained by the factor of interest in the experiment (in the current example, the methods of learning). The denominator for the *F* ratio is the unexplained variance in the outcome variable; this is the variance that is not attributable to the factor—the experimental error variance. The larger the ratio of explained to unexplained variance, the higher the value of *F*, and the lower the associated *p* value.

By introducing a covariate that is correlated with the outcome variable, some of the previously unexplained variance is now explained by the covariate and is thereby removed from the experimental error variance (the greater the correlation of the covariate with the outcome variable, the more variance is explained). As a result, the denominator of the *F* ratio is now smaller, the *F* ratio increases, and a smaller *p* value is produced. This will be achieved even if the groups do not differ on the covariate. So, in this example, even if students in the SCL groups had precisely the same mean age as those in the TDL group, providing age is correlated with test performance, the sensitivity of the statistical test on the two methods of learning would increase following the introduction of age as a covariate.

A situation in which it is particularly helpful to use ANCOVA in this way is where there are pretest scores on the outcome variable. Pretest scores tend to have a fairly high correlation with posttest scores, and the proportion of variance in the outcome variable that is explained by the covariate is correspondingly large. For example, a correlation of 0.5 or greater between pretest and posttest scores is quite common, and in such a situation, owing to the increased power derived from using the pretest scores as a covariate, the required sample size can be expected to be about 25% lower than that required for an unadjusted analysis on the posttest scores.

The cost of adjusting for a covariate is a loss of one degree of freedom, but except in very small studies, and unless the covariate accounts for negligible variance, this is amply recompensed by the increased sensitivity of the statistical test.

test.

Assumptions of the Analysis

The basic assumptions of ANCOVA are a combination of those for ANOVA and those for linear regression:

1. The level of measurement of both the outcome variable and the covariate is interval or ratio.
2. The predictive relationship between the covariate and the outcome variable is linear.
3. The covariate is a fixed variable and measured without error. Covariates are rarely fixed, but a covariate that is a random variable can normally be used provided that Assumption 7 in this list is satisfied.
4. The residuals are independent (i.e., the value of one residual does not influence, and is not influenced by, the value of any other residual).
5. The residuals have homogeneity of variance (homoscedasticity).
6. The residuals are (approximately) normally distributed—this assumption, which only applies to the residuals, not to the covariate, is required for hypothesis tests and confidence intervals. With larger sample sizes, this assumption becomes less stringent.
7. The residuals are uncorrelated with the covariate.

There are two additional assumptions that are specific to ANCOVA. The first is that the regression slope, β, should be equal in the two groups (the homogeneity of regression slopes, or parallelism, assumption). In the present example, if the relationship between age and test performance (β) differs between the study groups, the degree of adjustment of test scores should also differ between the groups, requiring a differing value of β for each group. However, the adjustment carried out within the ANCOVA model is in terms of the overall regression slope for the whole sample, which would clearly be inappropriate if the group-specific slopes are different.

The assumption of homogeneous regression slopes can be tested by constructing a scatterplot of the covariate and the outcome variable and fitting separate regression lines for the groups; the extent to which these lines are parallel can be judged visually. In addition, the assumption can be tested statistically. This involves a test of the interaction between the covariate and the grouping variable, which will tell us whether the relationship between the covariate and the outcome variable differs significantly across groups. Accordingly, a statistical model is constructed with the outcome variable, the grouping variable, the covariate, and a term representing the interaction between the grouping variable and the covariate.

A nonsignificant interaction supports the assumption of parallel regression slopes—although it should be remembered that all statistical tests of model assumptions are sensitive to sample size, the results of the test should be interpreted alongside visual assessment of the scatterplot. If the assumption of parallel regression slopes is considered to be untenable, one possibility is to categorize the covariate and include it as a set of dummy variables. In the process, some information in the covariate will be lost, and it will perform less effectively as a control variable or as a means of increasing statistical power, but the requirement for homogeneous regression slopes will have been avoided.

The second assumption—which is more a design assumption than a strict statistical assumption—is that the covariate should not be affected by the experimental factor. This normally has to do with the time at which a modifiable covariate is measured. Let us adapt the previous example, such that the covariate concerned is not age but anxiety. If we were to measure the students' anxiety after introducing the two methods of learning, it could be that these methods might differentially affect the students' anxiety (one method of learning might create greater self-confidence in a test situation and thus lower levels of anxiety in the group concerned). The implication of this is that the anxiety scores will contain within them part of the effect of the different methods of learning, so that when we adjust for anxiety we will at the same time adjust for the intervention effect. It is therefore important that any modifiable covariates are measured before the introduction of the experimental factor.

It should also be noted that issues of collinearity may arise if there are a number of covariates. With regard to its effect on *p* values, however, collinearity is often a less serious problem than in multiple linear regression, as the statistical significance of the covariates is not normally of interest when their function is simply that of statistical control.

Using ANCOVA Outside Experimental Contexts

The examples given earlier were from a randomized experiment. In a quasi-experimental design, the intervention groups are not formed by randomization but are preexisting, and often naturally occurring, groups. For example, a study might be based on testing two methods of instruction on male versus female students or on psychology versus sociology students. In this situation, the assumptions of ANCOVA need especially close attention (e.g., the homogeneity of regression slopes assumption may be more readily violated, and measurement error in the covariate may have more serious implications, than in a randomized experiment). Additionally, the results of the analysis should be interpreted carefully. For example, it might be argued that by adjusting the study groups in relation to a covariate, the “statistical” groups that are compared in the hypothesis test differ inappropriately from the “natural” groups on which the study was intended to be based. Additionally, the groups might be adjusted to an overall mean covariate score that would be unrepresentative of the individual groups if their mean scores on the covariate are at a considerable distance from the overall mean. It is helpful in such situations to present the results of an unadjusted analysis alongside those from the ANCOVA, as a sensitivity analysis.

Another issue that arises when employing ANCOVA with predetermined groups is closely linked to the earlier point about adjusting for an intervention effect in an experimental design. In an experiment, differences in covariate values between groups occur by chance in the process of randomization. In other instances, however, a chosen covariate may be intrinsically related to the factor that defines the groups to be compared. For example, a researcher intending to compare the academic performance of two group of students from different years—third grade and fourth grade, for example—might wish to adjust the comparison for age and sex. The adjustment for sex is probably appropriate, ensuring that differing proportions of boys and girls in the two groups do not confound the comparison of performance scores. However, age is likely to be correlated with many of the characteristics that distinguish third-grade students from fourth-grade students (e.g., age-related changes in problem solving, verbal reasoning, or abstract thinking), and an adjustment for age is likely to remove much of the difference between the two groups that the researcher wishes to test.

Other Considerations

ANCOVA employs a numerical (continuous or interval level) covariate. It is important to remember, however, that much the same effect, in terms of adjustment and/or statistical power, can be accomplished by introducing nominal or ordinal variables into the statistical model in a similar way. These might be thought of as also being covariates, although the term *ANCOVA* would only be used to describe the model when at least one such variable is numerical.

Another method that is sometimes used to control for pretest differences in an experimental study is the use of change (or gain) scores—that is, each participant's pretest score is subtracted from the participant's posttest score. However, this method is generally considered to be inferior to ANCOVA as a means of controlling for between-group differences in pretest scores because, unlike ANCOVA, it does not take into account the phenomenon of regression to the mean and may lead to biased estimates. Furthermore, the use of change scores will often lead to a less powerful statistical test than if ANCOVA were used.

ANCOVA provides an effective means of statistical adjustment for potential confounding factors—but it can only do so in respect of confounders that have been identified and measured by the investigator. Accordingly, ANCOVA is an adjunct to, not a substitute for, design features that control for confounding, in particular, randomization. Random allocation serves to balance all potential confounders across experimental groups, irrespective of whether they have been identified as confounders, irrespective of whether they have been measured, and irrespective of whether they are even measurable.

Julius Sim

See also Analysis of Variance; Gain Scores, Analysis of; Multicollinearity; Multiple Linear Regression; Simple Linear Regression

Further Readings

Egbewale, B. E., Lewis, M., Sim, J. (2014). Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: A simulation study. BMC Medical Research Methodology 14, 49.

Huitema, B. E. (2011). The analysis of covariance and alternatives (2nd ed.). Hoboken, NJ: Wiley.

Keppel, G., & Wickens, T. D. (2004). Design and analysis: A researcher's handbook (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

Miller, G. A., & Chapman J. P. (2001). Misunderstanding analysis of covariance. Journal of Abnormal Psychology, 110, 40–48.

Milliken, G. A. (2002). Analysis of messy data: Vol 3. Analysis of covariance. Boca Raton, FL: Chapman … Hall.

Rutherford, A. (2011). ANOVA and ANCOVA: A GLM approach (2nd ed.). Hoboken, NJ: Wiley.

Edward L. Boone Edward L. Boone Boone, Edward L.

Analysis of Variance

Analysis of variance

86

89

Analysis of Variance

Often researchers are confronted with determining whether the means of two or more groups differ. The analysis of variance (ANOVA) technique is a parametric hypothesis test to answer this question. ANOVA seeks to partition the overall data into components that correspond to variance explained by the groupings and variance that is unexplained by the groupings. Often the groups are defined by which treatment has been given to each of the experimental units in the group. In cases where the experimental units are randomly assigned to the treatment groups, ANOVA can be used to show causation. This entry discusses the basic principles of ANOVA and its organization, extensions, its use in contrasts and post hoc tests, and its assumptions.

Basic Principles

The simplest case of ANOVA is the one-way ANOVA where the groups are varied across only one factor and each group has the same sample size. Suppose there are *k* groups and within each group there are *n* samples taken from each group for a total sample size of *nk*. For notation, let y_{ij} be the measurement of outcome of interest from the *j*th sample in the *i*th group. We let μ_i denote the population mean of group *i*. In this notation, the ANOVA null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = L = \mu_k .$$

This hypothesis corresponds to the state where all of the means μ_i are equal to each other and hence do not differ. The alternative hypothesis in this case is:

$$\mathrm{H}_A : \text{at least two } \mu_i \text{ differ.}$$

If the ANOVA test rejects H_0 in favor of $H_{A,}$ this means there is enough evidence to conclude that the group means are truly different.

To accomplish this, ANOVA partitions the overall variance. The overall variance is simply the sample standard deviation squared of all of the data regardless of treatment group. In our notation, we would have:

$$S^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2}{nk-1},$$

where is the overall mean. In our notation, is the sample mean for the *i*th group. Here, the denominator is not useful in partitioning the groups and will be discarded to create the sum of squares total, denoted by SS_{TO} and is given by:

$$SS_{TO} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$$

By simply adding and subtracting the group sample mean in the SS_{TO} and doing some algebra (some algebra details have been omitted), one can obtain:

$$SS_{TO} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet} + \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}[(y_{ij} - \bar{y}_{i\bullet})^2 + 2(y_{ij} - \bar{y}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$$

$$+ (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2\,]$$

$$= \ldots$$

$$= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}_{\text{Error}} + \underbrace{\sum_{i=1}^{k} n(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}_{\text{Treatment}}$$

$$= SS_E + SS_T\,.$$

Notice by doing this, the SS_{TO} can be expressed as the sum of a term associated with error and a term associated with the treatment group. This is the essence of ANOVA, partitioning the SS_{TO} into meaningful components. Furthermore, each of the components is itself a sum of items that are squared; hence, the names sum of squares error, SS_E, and sum of squares treatment, SS_T are often given to the components. Note that in this one-way ANOVA scenario, SS_T is often called the sum of squares between and the SS_E is often called the sum of squares within and are denoted by SS_B and SS_W, respectively.

Similarly, for the one-way ANOVA case with equal sample sizes, the total degrees of freedom, $df_{TO} = nk - 1$, associated with the SS_{TO} can also be partitioned into degrees of freedom error, $df_E = n(k – 1)$, and degrees of freedom for treatment, $df_T = k – 1$. As with the sum of squares, the degrees of freedom also add together nicely $df_{TO} = df_E + df_T$.

Although all of this algebra may seem not to address the original answer, using the components above a signal-to-noise ratio can be created by:

$$F^* = \frac{SS_T / df_T}{SS_E / df_E}.$$

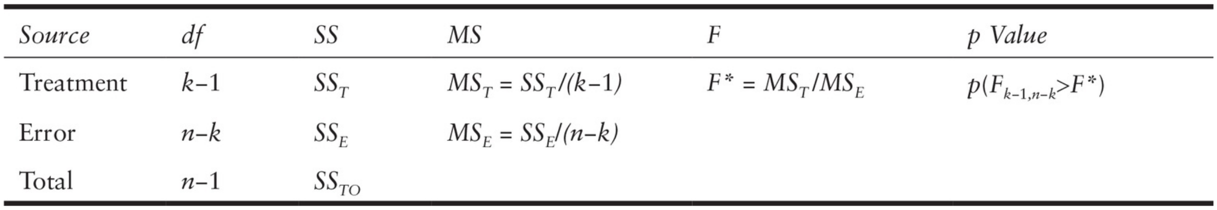
Notice that F^* is a fraction (ratio) with the observed “variance” associated with the treatment in the numerator and “variance” associated with the error in the denominator. Here, the * in the superscript is to denote that this value is an observed value that is calculated from the data. As with all signal-to-noise ratios, if $F^*<1$, then there is more noise than signal, and hence, there isn’t much evidence for H_A. If $F^*\approx 1$, then there is about the same amount of signal as noise and again not much evidence for H_A. However, if F^* is much greater than 1, then there is a lot of signal and little noise giving evidence toward H_A. Often F^* is called the *F*-statistic to differentiate it from the *F* distribution.

The question then becomes how big does F^* need to be for there to be enough evidence toward H_A that one could consider it statistically significant? Fortunately, F^* has a probability distribution associated with it, namely the *F* distribution. Recall, the *F* distribution is defined by both its numerator and denominator degrees of freedom, denoted by df_{num} and df_{den}, respectively. In this case, df_{num} is simply df_T and df_{den} is df_E. For a hypothesis test of H_0 versus H_A with a Type I error rate α if the calculated F^* statistic is greater than the $100\times(1-\alpha)$ quantile of the *F* distribution with the associated degrees of freedom, then the F^* is deemed to be “big enough” to be considered statistically significant. Hence, if one rejects H_0, then there are differences among the treatment groups. However, this test does give where the differences are, only if differences exist. To determine where the differences are a multiple comparison procedure would need to be performed after the ANOVA test.

Organization

Because there is a considerable amount of computation needed to calculate an

ANOVA test, the intermediary calculations are typically organized into what is called an ANOVA table. Table 1 shows the structure of the one-way ANOVA table.



Source	*df*	*SS*	*MS*	*F*	*p Value*
Treatment	$k-1$	SS_T	$MS_T = SS_T/(k-1)$	$F^* = MS_T/MS_E$	$p(F_{k-1,n-k}>F^*)$
Error	$n-k$	SS_E	$MS_E = SS_E/(n-k)$		
Total	$n-1$	SS_{TO}			

While in the one-way ANOVA case, the table seems simplistic and may not be clear why we would use this format; in the multiway ANOVA case, organization is paramount for both calculations and the ability to find the test one is looking for.

Extensions

One of the key advantages to the ANOVA approach is that it can be extended to more than a single treatment factor. The technique can be developed for two or more treatment factors where the individual treatments can be tested as well as the interactions between the treatments. For this work, only the two-way table will be presented with its corresponding formulae. Note that it is assumed that all treatments are considered fixed effects meaning that the levels of the treatments were not obtained at random but instead specified by the researcher before the experiment.

In the two-way ANOVA case, there are two treatments, Treatment A and Treatment B, where Treatment A has *a* treatment levels and Treatment B has *b* treatment levels. Here, each the treatment combinations are applied to experimental units. Furthermore, we will assume that each treatment combination is applied to the same number, *n*, of experimental units. This is a balanced case where the formulae are much easier to write. Let y_{ijk}, denote the value of the *k*th observation in the *i*th factor level of Treatment A and the *j*th factor level of Treatment B. In this notation, the following means will be needed to partition the variation. The overall mean: ; the mean of treatment combination consisting of the *i*th factor level of Treatment A and the *j*th factor level of Treatment B: . There will be $a \times b$ of these; the mean of the *i*th factor level of Treatment A: . There will be *a* of these; the mean of the *j*th factor level of Treatment B: . There will be *b* of these.

As in the one-way ANOVA setting, the total variation, SS_{TO}, can be partitioned into variation associated with the treatments Treatment A, Treatment B, the interaction between treatments, and error, namely SS_A, SS_B, SS_{AB}, and SS_E, respectively.

$$SS_{TO} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\cdots})^2$$

$$= nb\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 + na\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$$

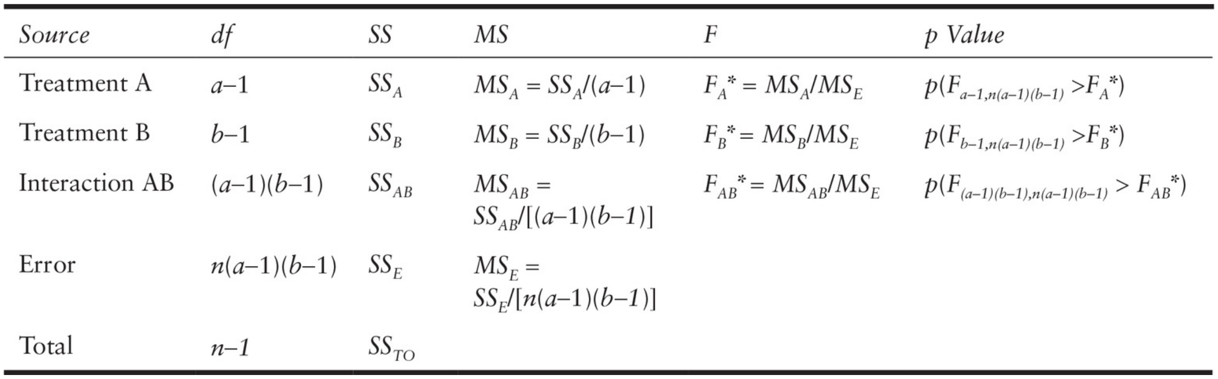
$$+ n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$$

$$+\sum_{i=1}^{a}\sum_{i=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\cdot})^2$$

$$= SS_A + SS_B + SS_{AB} + SS_E.$$

For simplicity, the algebraic steps have been omitted. From the equations just presented, one can see that the computations for the two-way ANOVA setting are considerably more tedious than the one-way ANOVA setting as many items need to be kept track of.

Table 2 gives the ANOVA table for the two-way ANOVA setting. Notice that three tests are included in the table as given by the three *p* values on the right side of the table. In the two-way setting, the interaction term is considered first as it gives an indication of whether the two treatments act in conjunction with each other. If the treatments do interact, then the main effects tests do not accurately isolate the effect of the treatments.



Source	*df*	*SS*	*MS*	*F*	*p Value*
Treatment A	$a–1$	SS_A	$MS_A = SS_A/(a–1)$	$F_A{}^* = MS_A/MS_E$	$p(F_{a-1,n(a-1)(b-1)} > F_A{}^*)$
Treatment B	$b–1$	SS_B	$MS_B = SS_B/(b–1)$	$F_B{}^* = MS_B/MS_E$	$p(F_{b-1,n(a-1)(b-1)} > F_B{}^*)$
Interaction AB	$(a–1)(b–1)$	SS_{AB}	$MS_{AB} =$ $SS_{AB}/[(a–1)(b–1)]$	$F_{AB}{}^* = MS_{AB}/MS_E$	$p(F_{(a-1)(b-1),n(a-1)(b-1)} > F_{AB}{}^*)$
Error	$n(a–1)(b–1)$	SS_E	$MS_E =$ $SS_E/[n(a–1)(b–1)]$		
Total	$n–1$	SS_{TO}			

Furthermore, ANOVA can be extended to any linear model setting, including linear regression, randomized complete block designs, fractional factorial designs, analysis of covariance, and repeated measures ANOVA. This flexible approach to partitioning the variance to determine which treatment factors may be significant is extremely useful and is readily available in statistical software packages such as SPSS, SAS, STATA, R, Minitab, and JMP. Most statistical software packages will provide the appropriate ANOVA table upon request.

Contrasts and Post Hoc Tests

Although ANOVA is extremely powerful for testing whether differences exist among the means, it does not identify where the differences in the group means are to be found. To determine which group means are different, one could employ either a contrast test or one of many post hoc tests. Contrast tests are specified before any experimentation begins and are used to test specific combinations of group means. Although contrast tests are far more powerful than post hoc tests, many researchers find them difficult to correctly specify the desired contrast test.

Post hoc tests ubiquitously used in research, despite the lack of power compared to contrast tests. There are many post hoc tests on the group means such as Fisher’s least significant difference, Bonferroni correction, Tukey’s honestly significant difference, Dunnett’s test, and many others. These post hoc tests attempt to test a large number of differences between group means and are often called multiple comparison procedures. The fact that they are conducting multiple tests while attempting to control the Type I error rate is where these multiple tests while attempting to control the Type I error rate is where these procedures lose statistical power compared to a predefined contrast test.

Assumptions

As with all statistical analyses, some assumptions are necessary and ANOVA is no different. Because ANOVA is a linear model, it has the same assumptions as regression analysis: normality, independence, and constant variance of residuals. Normality can easily be assessed using Q-Q plots and tested via tests such as Kolmogorov–Smirnov, Shapiro–Wilks, and Anderson–Darling. The normality assumption can be relaxed when the study is an experiment where random assignment to treatment group has been utilized. In this case, randomization theory can be used and normality is no longer a needed assumption.

Constant variance can be evaluated using side-by-side box plots of the residuals where each box plot corresponds to the residuals for a particular treatment combination and is tested via Levene's test, Bartlett's test, or Hartley's test. Independence is more difficult to assess and test as one would need to know the structure of the dependence such as temporal dependence or spatial dependence. Typically, the researcher should design the experiment in such a way that independence would be guaranteed by the experimental procedure versus testing for independence during the analysis. In cases where the assumptions are severely violated, the nonparametric alternative Kruskal–Wallis test may be employed which also results in a loss of statistical power.

Edward L. Boone

See also Analysis of Covariance; Bonferroni Procedure; Levene's Homogeneity of Variance Test; Multiple Linear Regression; Simple Linear Regression

Further Readings

Montgomery, D. (2012). Design and analysis of experiments (7th ed.). New York, NY: Wiley.

Ott, R. L., & Longnecker, M. (2015). An introduction to statistical methods and data analysis (7th ed.). Pacific Grove, CA: Brooks Cole.

Turner, J. R., & Thayer, J. (2001) Introduction to analysis of variance. Thousand Oaks, CA: Sage.

Claudia A. Gentile Claudia A. Gentile Gentile, Claudia A.

Analytic Scoring Analytic scoring

89

91

Analytic Scoring

Analytic scoring is a method of evaluating student work that requires assigning a separate score for each dimension of a task. Often used with performance assessment tasks, analytic scoring rubrics specify the key dimensions of a task and define student performance relative to a set of criteria across performance levels for each dimension. For example, analytic rubrics used to evaluate student essay writing often include the following dimensions: development of ideas, organization, language use, vocabulary, grammar, spelling, and mechanics.

Analytic rubrics used to evaluate students’ social studies reports might include the same dimensions but also dimensions specific to social studies: use of original source material, accuracy of information, quality of source material, and correct citations. The remainder of this entry describes the uses of analytic scoring in education and then looks at the benefits and challenges of this method.

Analytic scoring is used widely in education to evaluate students’ performance in various subject areas (such as reading, writing, speaking, mathematics, the sciences, social studies, world languages, physical education, industrial technology, and the arts). It is also used to evaluate students studying for professional careers in various fields (such as engineering, nursing, business, and teaching). Analytic rubrics have been developed and used at all grade levels, including early childhood, elementary, secondary, undergraduate, graduate, and postgraduate.

Analytic scoring is most often used when there is a need to assess how well students perform on individual dimensions of whole product or performance. Teachers, students, and/or evaluators use analytic rubrics to review the product or performance and assign ratings for each dimension, resulting in a set of subscores that can be combined to generate an overall score. Each dimension can subscores that can be combined to generate an overall score. Each dimension can be weighted equally or the weights on dimensions can vary, depending on the importance of each dimension to the successful accomplishment of the task. It is important to note that the relative importance of each dimension and the definition of successful performance on each dimension may vary with the specific topic or task. Thus, analytic scoring rubrics need to be customized for each performance task.

In recent years, analytic scoring has also been used to develop automated essay scoring systems. Researchers use human analytic ratings on student writing to develop automated models of the key dimensions of writing and then to test the validity of the automated models.

Benefits

Because analytic scoring identifies the key dimensions of a performance task and defines performance along a developmental continuum for each dimension, it is an approach to evaluating student work that provides an effective mechanism for identifying students’ strengths and weaknesses, more so than do alternate methods of scoring, such as holistic scoring. Holistic scoring involves examining multiple dimensions of students’ performance and then assigning a single overall score to capture the level of that performance.

Holistic scoring is a very efficient way of identifying students at the upper end of the scale (who excel on most or all dimensions) and those at the lower end of the scale (who struggle on most or all dimensions). However, for the majority of students who perform variably across dimensions, a single score is not very informative. Instead, the multiple dimension scores on analytic scoring rubrics provide students and teachers with specific information about students’ performance that can be used to individualize students’ learning plans and to monitor students’ progress across time.

Research on the use of analytic scoring rubrics has found that they provide valid judgments of complex competencies and that the analytic domains capture meaningful variation in student performance. In addition, with proper rubric construction and training, experts can be trained to reach a high level of interrater agreement when using analytic scoring rubrics, and the multiple scores that result from using an analytic rubric positively contribute toward test reliability (more so than does a single holistic rating). There is some evidence that rating on multiple traits increases task generalizability, so that fewer that rating on multiple traits increases task generalizability, so that fewer performance tasks are needed on an assessment.

Moreover, the use of analytic scoring rubrics has been found to promote learning by (a) making expectations and criteria clear, (b) providing a common language for teachers and students to discuss the subject, (c) facilitating teacher feedback to students, and (d) supporting students’ self-assessment. Students report that the feedbacks from analytic scoring rubrics are helpful.

Challenges

Some of the measurement challenges that surround the use of analytic scoring include concerns about whether unique information is provided by analytic scores. The dimensions defined by the rubric are often highly correlated and, thus, do not represent independent information about students’ knowledge and/or skills. For example, in the evaluation of students’ writing, the overall length of an essay correlates highly with many of the key dimensions of writing (such as the development of ideas, which requires a certain length of writing, and organization of ideas, which cannot be fully employed unless the essay has at least three paragraphs).

A related concern is the halo effect—when raters assign multiple analytic scores to a student performance or product, do they allow the rating of one dimension to influence their rating of the other dimensions? There is evidence to suggest that the analytic rating of student writing may be prone to the halo effect, so that the number of actual independent dimensions may be fewer than the number of dimensions on the rubric. To explore how well analytic scoring rubrics measure competencies across a number of dimensions, researchers recommend the use of factor analysis, which can provide rubric developers with valuable information so that each dimension on the rubric corresponds to one unique competency, enhancing the efficiency and effectiveness of the rubric.

Claudia A. Gentile

See also Holistic Scoring; InterRater Reliability; Performance-Based Assessment; Reliability; Rubrics

Further Readings

Bennett, R. E. (2015, March). The changing nature of educational assessment.

Review of Research in Education, 39(1), 370–407.

Hammond, L. D., & Adamson, F. (2014). Beyond the bubble test: How performance assessments support 21st century learning. San Francisco, CA: Jossey-Bass.

Lai, E. R., Wolfe, E. W., & Vickers, D. (2015, February). Differentiation of illusory and true halo in writing scores. Educational and Psychological Measurement, 75, 102–125.

McMillan, J. H. (2012). SAGE handbook of research on classroom assessment. Thousand Oaks, CA: Sage.

Eric T. Beeson Eric T. Beeson Beeson, Eric T.

Andragogy

Andragogy

91

92

Andragogy

The term *andragogy* refers to a set of principles and assumptions about adult learners, the learning environment, and the learning process. Educational research, measurement, and evaluation require a firm understanding of the underlying instructional theories guiding best practices. This entry provides an overview of andragogy including its core assumptions of the learner and learning environment, key outcomes and criticism, and methods of assessment.

Originating from the Greek root *andra* (meaning adult) and *agogus* (meaning to lead), the concept of andragogy can be traced back to Alexander Kapp, a German educator, in the early 1800s; however, it was not until the late 1960s that andragogy was popularized by the work of American educator Malcolm Knowles. Although typically associated with adult learning, andragogy describes *adulthood* as a psychological, rather than a chronological, milestone in which the learner develops a self-concept striving toward independence.

Knowles defined andragogy as the “art and science of helping adults learn” (1980, p. 43). Central to andragogy are six core assumptions that adult learners:

1. need to know the why, what, and how of the educational experience;
2. strive toward a self-concept of independence, autonomy, and self-actualization;
3. have invaluable resources from their previous experiences that can enrich their current educational endeavors;
4. develop readiness to learn based upon the relevance of the current scenario to their current developmental tasks;

5. have an orientation to learning that is grounded in real-world scenarios of personal importance; and
6. are internally motivated by goal attainment and problem resolution.

Furthermore, andragogy outlines four assumptions about the learning environment. These assumptions are as follows:

1. The teacher is a facilitator of a coconstructed experience of learning focused on self-directedness, autonomy, and self-actualization;
2. Instructional methods such as experiential exercises, problem-and case-based learning, role-playing, simulations, Socratic questioning, and field experiences help students identify gaps between what they know, what they don't know, and strategies for how to fill in these gaps;
3. Real-world scenarios are the organizing structure for the learning process; and
4. Scenarios should be scaffolded according to the desired learning outcomes and current developmental level. Andragogical methods focus on the development of cognitive complexity and self-directed learning skills rather than the simple remembering of facts.

The evaluation of student learning outcomes when using instructional methods grounded in andragogy can be evaluated objectively through methods such as multiple choice exams but are best assessed using multiple strategies including portfolios, case presentations, role-playing, and clinical scenario exams. Andragogical methods are especially effective in increasing learners' situational interest, cognitive complexity, clinical reasoning, lifelong learning skills, satisfaction, long-term retention, performance on free-recall tasks, performance on short answer and essay tests, ratings by supervisors on clinical observations, and performance on clinical or case-based portions of exams. Andragogical methods may be less effective when the short-term recognition of facts and concepts is needed for multiple-choice and true–false portions of exams.

Eric T. Beeson

See also Instructional Theory; Learning Theories; Long-Term Memory; Portfolio Assessment; Self-Directed Learning; Short-Term Memory

Further Readings

Bolton, F. C. (2006). Rubrics and adult learners: Andragogy and assessment.

Assessment Update, 18(3), 5–6.

Harden, R. M., & Davis, M. H. (1998). The continuum of problem-based learning. Medical Teacher, 20(4), 317–322.

Knowles, M. S. (1980). The modern practice of adult education: From pedagogy to andragogy (revised and updated). Englewood Cliffs, NJ: Cambridge Adult Education.

Knowles, M. S., Holton, E. F.III, & Swanson, R. A. (2005). Adult learner: The definitive classic in adult education and human resources development (6th ed.). Burlington, MA: Elsevier.

St. Clair, R. (2002). Andragogy revisited: Theory for the 21st century? Retrieved from ERIC database. (ED468612) Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. Interdisciplinary Journal of Problem-Based Learning, 3(1). Retrieved from http://dx.doi.org/10.7771/1541–5015.1046

Taylor, B., & Kroth, M. (2009). Andragogy's transition into the future: Meta-analysis of andragogy and its search for a measurable instrument. Journal of Adult Education, 38(1), 1–11.

Stephen G. Sireci Stephen G. Sireci Sireci, Stephen G.

Alejandra Garcia Alejandra Garcia Garcia, Alejandra

Angoff Method

Angoff method

92

95

Angoff Method

This entry describes the Angoff method for setting standards on educational tests and how it can be used to set valid standards on educational tests. *Standard setting* refers to the process used to establish cut scores on educational tests that are used to classify test takers into categories such as “pass,” “fail,” “proficient,” “advanced,” and other categories generally referred to as achievement levels. Many educational tests, such as licensure tests professionals are required to pass to become licensed and high school graduation tests that students must pass to receive a high school diploma, require these standards.

Most people in modern society have taken tests based on which the standards are set. However, it is not widely known as to *how* those standards were set. The most popular method is the Angoff method and its variations.

In 1971, William Angoff wrote a seminal chapter called “Scales, Norms, and Equivalent Scores” in a book on educational measurement. In the chapter, he described how test developers transform students’ responses to test items into standardized scores and how they maintain equivalence of these score scales over time.

In describing how to incorporate meaning into the score scale by setting “pass/fail” standards on the scale, Angoff described a method suggested by his colleague Ledyard Tucker. This process involved having subject matter experts (SMEs) think about the “minimally competent” test taker; that is, the test taker who “just barely” has the sufficient knowledge and skills required to pass the exam (sometimes referred to as the “borderline” candidate). The task for the SMEs was to review each test item and judge whether the minimally competent test taker would answer the item correctly. The passing score suggested by each SME is calculated by simply summing the number of items the SME predicted would be correctly answered by the minimally competent candidate and then averaging that score across the SMEs.

Angoff added a footnote to his description of Tucker’s “yes/no” method and suggested instead of judging whether the minimally competent test taker would or would not correctly answer the item, the SMEs could estimate the *probability* the minimally competent test taker would correctly answer the item. Those probability ratings could then be summed, and averaged over SMEs, to derive the passing standard. The process he suggested in that footnote became known as the Angoff method and quickly became the most popular method for setting standards on educational tests.

Like all test-centered standard-setting methods (i.e., methods where SMEs review and rate test items), the Angoff method involves several steps. These steps include (a) discussing the knowledge and skills of the minimally competent test taker, (b) reviewing the test items, (c) providing a probability rating for each item, (d) discussing all or a subset of those ratings, and (e) revising the original ratings as the SMEs regard necessary. The final cut score is based on the revised ratings in Step (e).

As a simple illustration of the Angoff method, imagine a test with 100 items. If an SME reviewed each item and estimated the minimally competent test taker would have a 0.50 probability of answering each item correctly, the SME-suggested passing score would be 50 (i.e., 0.50 × 100 items). Of course, no SME would assign the same probability value to all items because items vary in their difficulty. Thus, our example is oversimplified to illustrate how the cut score is calculated for a single SME. The final cut score would be averaged over all SMEs.

The process described thus far refers to items that are scored dichotomously, which means 1 point for a correct answer and 0 for an incorrect answer. However, many educational tests use items that are scored on longer scales (e.g., an essay worth 6 points). Also, many educational tests, such as the National Assessment of Educational Progress in the United States, have more than two pass/fail standards. For these and other reasons, there have been many “modifications” of the Angoff method “modifications” of the Angoff method.

Modifications of the Angoff Method

Modified versions of the Angoff method have been introduced and widely used to (a) set standards on tests with polytomously scored items, (b) set standards on tests involving more than two standards (e.g., classifying students into categories such as “basic, “proficient,” and “advanced”), and (c) increase the agreement among SMEs. The term *modified* indicates the original method has been altered for a specific application. In an extensive chapter on standard setting, Ronald Hambleton and Mary Pitoniak hypothesized there may be more than 100 variations of the Angoff method. The “traditional” Angoff method comprises five steps:

1. train the SMEs on the process,
2. facilitate a discussion of the minimally qualified (borderline) test taker,
3. collect the first round of ratings of SMEs,
4. SMEs discuss first round ratings, and
5. SMEs revise their ratings based on the discussions.

In the case of polytomously scored items, rather than making a probability rating for each item, SMEs estimate the mean score they expect the borderline test taker to achieve on each item. For tests that involve more than two standards, modifications include having the SMEs make separate judgments for each item for each standard. To increase agreement among the SMEs, additional rounds of discussion are used.

A related modification is to give the SMEs statistics describing the difficulty levels of the items after they make their initial ratings. These statistics are thought to provide a “reality check” for the SMEs. For example, if an SME thought borderline candidates had a high probability of answering an item, but the item statistics suggested very few examinees answered the item correctly, the SME may take a deeper look at the item to understand why and possibly revise the rating.

Providing item statistics to SMEs is somewhat controversial. The Angoff method is intended to produce a content-based (criterion-referenced) standard, which is why SMEs who are familiar with the content being tested and the students being assessed are selected as judges. By relying on empirical information from test takers, the standard may become norm referenced. Studies have found that SMEs have a difficult time setting appropriate standards without empirical data, but critics counter item statistics can overly influence SMEs, leading to standards that are driven by item difficulty, rather than by SME judgment regarding what constitutes appropriate achievement.

Research has shown SMEs tend to make more inaccurate predictions of borderline test takers' performance on relatively easy or difficult items, when they are not provided empirical data in the form of item statistics. Some researchers point to this as an inherent problem in the Angoff method that reduces its utility as a standard-setting procedure.

Another modification of the Angoff method, and one that is not contentious, is facilitating several rounds of discussion among the SMEs. Such discussion allows SMEs to consider different viewpoints about what makes an item difficult and how the knowledge and skills of borderline test takers are exhibited in performance on an exam. Many studies have shown these discussions reduce variance (increase consensus) among SMEs. In fact, agreement among the SMEs is one of the criteria based on which standard-setting studies are evaluated. However, the type of feedback provided, and when it is provided, can impact how SMEs modify their ratings across rounds.

Evaluating and Validating Angoff Standard-Setting Studies

The validity of a standard-setting study is typically evaluated using procedural, internal, and external validity evidence. Procedural evidence refers to the quality of the standard-setting study, starting with recruiting qualified panelists and training them well, and proper execution of the study. Internal evidence focuses on the consistency of results, ideally estimating the variability in the standards set, if the study were replicated.

External evidence refers to the degree to which the classifications of examinees are consistent with other performance data. Examples of external validity evidence include classification consistency across different standard-setting methods, tests of mean differences across examinees classified in different achievement levels on other construct-relevant variables, and the degree to which external ratings of test takers' performance are consistent with their test-based achievement-level classifications based achievement-level classifications.

There are several actions standard setters can do to build validity into a standard-setting study, as opposed to evaluating the validity of the standard after it is set. One important consideration is the number and quality of the SMEs. Research suggests at least 10 SMEs be used but more is better to reduce the standard error of the (mean) cut score. Equally as important, if not more important, than the number of SMEs are their qualifications and representativeness. SMEs should be fully proficient in the knowledge and skill areas measured on the test, and they should represent the relevant population (e.g., students, teachers, licensed professionals) with respect to subdiscipline areas of expertise and demographic factors.

Another important consideration is the quality of the training. SMEs should be required to take at least some test items, without the answer keys, to get an appreciation of the difficulty of the exam. They should also practice rating items and discuss the items to make sure they are on task and they understand the notion of the minimally competent test taker.

After gathering initial Angoff ratings, validity can be built into the process by having SMEs discuss and review their ratings. These discussions often illuminate item features SMEs may have missed when initially rating the items and will correct any coding errors or other errors they may have made. Finally, surveys of SMEs during and at the conclusion of the study can help evaluate how well they understood their tasks and the factors they used in making their judgments.

The Influence of the Angoff Method

The Angoff method is not the only method for setting standards on educational tests, but it is often the method to which others are compared. The legacy of the Angoff method is that it illustrated how standards can be set on educational tests by aligning the standard to experts’ judgments of what is considered to be “above standard” performance. Rather than awarding passing scores and other achievement levels based on how well test takers perform relative to one another, the Angoff method sets an “absolute” standard that all test takers can achieve.

By successfully implementing the Angoff method, standard setters can have confidence the standards they set will be valid. However, successful implementation of the method requires competent and representative SMEs who are carefully trained, who understand their tasks and the “minimally competent” test takers, and who can provide reliable and valid ratings. Surveys of SMEs, and comparing test takers’ achievement-level classifications to their performance on other measures of their knowledge and skills, can help evaluate the quality of the results from an Angoff standard-setting study.

Stephen G. Sireci and Alejandra Garcia

See also Body of Work Method; Bookmark Method; Cut Scores; Ebel Method; Modified Angoff Method; National Assessment of Educational Progress; Standard Setting

Further Readings

American Educational Research Association, American Psychological Association, … National Council on Measurement in Education. (2014). Standards for educational and psychological tests. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 508–600). Washington, DC: American Council on Education.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. International Journal of Testing, 13, 65–85.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational measurement (4th ed.). Westport, CT: American Council on Education/Praeger.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425–461.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. C. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 119–157). Mahwah, NJ: Erlbaum.

Sireci, S. G., Hauger, J. B, Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 grade 12 National Assessment of Educational Progress mathematics test. Applied Measurement in Education, 22, 339–358.

Diana Joyce-Beaulieu Diana Joyce-Beaulieu Joyce-Beaulieu, Diana

Anxiety

Anxiety

95

97

Anxiety

The two distinguishing hallmarks of anxiety disorders are an emotional state of fear and worry that result in diminished functioning. Fearfulness may be based on an actual experience or fostered by cognitive distortions that result in the misperception of a threat. Consequently, a physiological response occurs that ranges from disconcerting to debilitating and may result in aggressive or avoidant behaviors to escape the distress. The worry associated with anxiety creates a persistent state of angst or apprehension that is sustained over time. Without treatment, anxiety can negatively impact personal well-being, academic achievement, employment, and lifelong accomplishments. This entry discusses types of anxiety disorders as well as measurement options.

Anxiety Diagnosis

The *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* delineates 11 disorders in the anxiety domain. All of the disorders share common features of fear and worry; however, they differ by the circumstance that triggers the anxiety, the type of response when anxious, and the thought distortions that maintain the anxiety.

Anxiety Onset in Early Childhood

Although there is variability in age of onset, two of these disorders are most likely to first occur during early childhood: separation anxiety disorder and selective mutism. The distinguishing feature for separation anxiety disorder is anxiety related to separation from individuals one has formed a deep emotional bond with (e.g., a parent). Symptoms may include fear about losing the primary caregiver or fear that grievous harm will occur to the caregiver. The resulting behaviors may include refusing to leave a caregiver, extreme distress when leaving home, and psychosomatic complaints. These behaviors can result in school absenteeism as well as a high incidence of school nurse visits for perceived physical complaints, thus missed instruction time and lower achievement.

Selective mutism has a pattern of only speaking in select social circumstances that are familiar and comfortable (e.g., at home) and only with certain individuals (e.g., parents, siblings). Children with selective mutism often avoid speaking at school, to teachers, and even to classmates. This behavior can result in limiting social skills development and also prohibit accurate classroom monitoring and assessment of skills.

Anxiety Onset in School-Age Children

Specific phobia and social phobia are anxiety disorders that most often occur among school-age children (i.e., those aged 7–15). Specific phobia involves a manifest significant fear of particular objects or circumstances, such as dogs, spiders, frogs, or high places (e.g., balconies). The level of fear accompanying exposure to the specific phobia may result in avoidant behaviors or hinder the individual from participating in activities. For example, a child afraid of heights may refuse to access a stairwell or elevator, thus mobility is compromised. For an adolescent with a fear of animals’ participation in required science labs may be problematic.

A social phobia is characterized by fearfulness of social settings wherein individuals may be observed by others and there is a perception that they will be negatively evaluated by others. This apprehension results in significant discomfort and sometimes avoidance of the social interaction. Individuals with social phobia may avoid meeting new people or even quit talking in groups, which can narrow their social networks. The effects of social phobia can be particularly devastating when purposeful evaluation is expected (e.g., a student class presentation assignment).

Anxiety Onset in Early Adulthood

Panic disorder, agoraphobia, and generalized anxiety disorder diagnoses are most likely to be made during early adulthood (i.e., aged 20–35). Panic disorder involves unexplained and very sudden (i.e., within minutes), overwhelming fear arousal. A number of possible intense physiological reactions are present (e.g., racing heart rate, profuse sweating) and may give the individual an unwarranted sense of high alert or impending doom. Repeated panic episodes can result in individuals significantly restricting their own social opportunities and educational or career aspirations out of fear of inducing a panic attack.

Agoraphobia is characterized by significant fear of being trapped or unable to escape specific situations (e.g., confined space, bus). Individuals with agoraphobia may avoid public places and transportation, resulting in self-seclusion. Generalized anxiety disorder is characterized by broader multiple fears sometimes including aspects of daily life (e.g., work). Generalized anxiety disorder can result in disturbed sleep, tense muscles, and irritability and may lower overall quality of life.

Other Anxiety Disorders

The last four diagnoses do not have a dominant age of onset. Substance/medication-induced anxiety disorder occurs as a result of substance use or withdrawal. The particular anxiety symptoms and intensity will vary based on the type of substance (e.g., alcohol, cocaine) that has induced the effects. Anxiety disorder due to another medical condition is diagnosed when anxiety symptoms are present; however, they are better understood as a result of a medical condition (e.g., seizure disorder). Other specified anxiety disorder and unspecified anxiety disorder diagnoses are warranted when anxiety symptoms are present but not pervasive enough to meet the full criteria of another anxiety disorder. Additionally, several other mental health disorders outside of the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* anxiety disorders domain (e.g., obsessive-compulsive disorders) have anxiety-related symptomology.

Measurement for Anxiety

Measurement of anxiety symptoms is often accomplished through norm-referenced anxiety rating scales completed by teachers, parents, or self-report. These measures offer an objective comparison of frequency and intensity of specific symptoms. Interview methods also are helpful in identifying temporal sequence of symptom onset, specific triggers for anxiety, and thought patterns that may perpetuate worry. Observations afford the opportunity to understand anxiousness within a context and measure demonstrated behaviors.

Clinicians also may measure small changes in anxious feelings during counseling sessions through a self-reported Subjective Units of Distress Scale, a number scale created in collaboration with the patient that denotes level of stress. Additionally, the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* offers online symptom measures through the publisher’s website. Together, these measures can inform treatment options and progression of symptoms.

Diana Joyce-Beaulieu

See also Asperger’s Syndrome; Autism Spectrum Disorder; *Diagnostic and Statistical Manual of Mental Disorders*

Further Readings

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

Mash, E. J., & Barkley, R. A. (2014). Child Psychopathology (3rd ed.). New York, NY: Guilford Press.

Sattler, J. M. (2014). Foundations of behavioral, social, and clinical assessment of children (6th ed.). La Mesa, CA: Author.

APA

See American Psychological Association

Timothy Franz Timothy Franz Franz, Timothy

APA Format

APA format

97

100

APA Format

The publication manual of the American Psychological Association (APA) sets the standard for writing in psychology and has also become the standard for writing in many other behavioral and social sciences disciplines such as education, nursing, and business. The sixth edition of the manual was published in 2009.

The guidelines set forth in the APA manual address how to communicate complex scientific writing, including writing style and the mechanics for formatting a paper. Writing in APA format involves two different tasks. The first of these is style, or the quality of the prose, and the second is the mechanics that includes requirements such as margin size, section headings, and how to give proper credit to others. This entry discusses APA guidelines for style and mechanics and lists other types of resources that can be used to understand APA format and other aspects of academic writing.

APA Writing Style: Writing Well

This entry gives only brief information about writing style, as there are many resources about how to improve writing style. Further, Chapters 3 and 4 of the APA manual detail how to write clearly (e.g., avoiding bias in language and writing in the active voice) as well as the expectations of grammar in scientific writing (e.g., using punctuation in text). Every manuscript is, of course, unique. However, in academic writing, certain conventions and elements are generally expected; these include that the manuscript has a thesis, that it be unified and coherent, and that it follow rules of grammar.

Thesis

A thesis, which is sometimes called a research question (though it is not written as a question but rather a statement) in an empirical paper, is the driving force behind any paper. A thesis, in general, is a clear statement of the paper's purpose. The thesis may be written in a way where it is implicit or explicit, but making a thesis explicit can help the author in writing the manuscript by creating a guide to what should be included.

Unity and Coherence

Once a paper has a clear thesis, the next step in writing is to make sure that the paper hangs together. Like any other paper, a manuscript written in APA format must have unity and coherence, which allows readers to remain focused on the relevant topic. Unity is when a paper logically flows from topic to topic throughout the manuscript, and each paragraph only contains a core idea. Specifically, the paper is united across all topics and each is linked to the thesis.

Coherence is when the ideas within a paragraph are presented in a rational order, and each is necessary in supporting the single idea presented in that paragraph. Thus, a paragraph should start with a topic sentence (that links to the thesis), include evidence supporting the topic, and then end with a concluding sentence that leads into the next paragraph. Papers include unity and coherence flow from the thesis to the conclusion, and this improves writing style and readability.

Grammar

Chapters 3 and 4 of the APA manual, as well as many other writing guides such as William Strunk Jr. and E. B. White's *The Elements of Style*, detail the requirements and expectations of proper grammar and writing with the style expected for scientific papers. Some of the concepts in Chapter 3 include organization, clarity, removing bias from language, and smoothness of expression. Chapter 3 ends with grammar and usage recommendations for scientific writing, such as avoiding the passive voice.

Chapter 4 of the APA manual is titled "The Mechanics of Style" and refers to "the rules or guidelines a publisher observes to ensure clear, consistent presentation in scholarly articles" (p. 87). This chapter describes many of the basic tools for writing, including punctuation marks and how to present statistics.

APA Format Mechanics

The mechanics of APA format are designed to improve the format, flow, and readability of review and empirical papers and to put those ideas into a format that a publisher may use for subsequent publication as necessary. These mechanics are somewhat different for empirical papers (those reporting findings of original research) and nonempirical papers. According to the APA manual, nonempirical work includes papers such as literature reviews, theoretical papers, and case studies.

There is little guidance in the APA manual regarding the overall flow of nonempirical papers. This is not surprising because the variety of topics for nonempirical paper can vary considerably. On the other hand, a considerable proportion of the APA manual is devoted to writing empirical papers, where there are more consistent expectations for the order of the material presented (see Chapter 2 of the APA manual).

Empirical manuscripts, regardless of the content and audience, are for the most part designed to answer, in order, the following five questions:

1. What did this project do? (Introduction)
2. Why did this project do this? (Introduction)
3. How did this project do this?
4. What did this project find?
5. What does it all mean?

Introduction

The first part of the introduction of an empirical paper typically answers the question *what did this project do?* The second, and often much more lengthy, part of the introduction section answers the question *why did this project do this?* To accomplish this, an introduction typically includes the following information: A brief summary that frames the project and explains why it is important, the thesis, a review of the past literature that has examined this issue, a logical explanation that explains the specific goals of the current project (usually also embedded in the relevant literature). Finally, a good place to end an introduction is with a brief paragraph that provides an overview of the method used to conduct the empirical study.

Method

Although there is some variability about the subsections that comprise a method section, the main point of the section is to explain the research process in enough detail so that another researcher may replicate the project. Thus, this section answers the question *how did this project do this*? The first part of this section should describe the design. This gives the readers a framework for understanding the remainder of the method section. After the design statement, the method section usually describes the sample for the project. The third part of the method section typically describes the materials (usually referring to things such as paper-and-pencil or online questionnaires) and/or apparatus (usually referring to physical materials, such as the type of computer and monitor). Finally, many studies end the method section with a description of the procedure, which is a step-by-step description of how the study occurred.

Results

The results section answers the question *what did this project find?* The purpose of this section is to explain just the facts with little to no interpretation. The goal for the results section is to explain the findings, including the descriptive information (e.g., themes if qualitative, or means if quantitative) as well as any relevant inferential statistics. Chapter 5 of the APA manual provides considerable detail about how to report numbers and statistics, including examples of tables and figures.

Discussion

The discussion section of a manuscript is where a researcher moves from the specific work in the study to some ideas to go beyond that study and answers the question *what does it all mean?* As in the introduction, the content in the discussion section varies from manuscript to manuscript. A discussion section can begin with a brief summary of what happened in the project (basically in one to three paragraphs restating what was stated in the results section). In short papers, this may seem redundant. Many papers, however, have multiple hypotheses and a short summary can help a reader better understand the disparate findings.

Next, the author should link the findings to past research. If the results, for the most part, support the hypotheses, this part of a manuscript may be a brief reiteration of the introduction. Most research, however, has at least some findings that are unexpected. In these cases, writing the discussion is often a more difficult process because it needs to explain why the findings occurred and, as in the introduction, ground that information in the past literature. This time, however, the explanations need to be based on new logic and literature. In this case, there are usually two major categories of explanations. The first category is methodological: That something about the method turned out to be a poor test of the theories and ideas (e.g., the sample was inappropriate or too small). The second category is theoretical: The ideas captured in the introduction were not properly derived. A well-written discussion where some of the predictions are not supported should include information that covers both categories.

After explaining the results and linking them to past research, a discussion section should move beyond the findings. This can occur by recommending future research (to respond to methodological limitations and/or extend theory) as well as the real-life implications of the findings. An empirical manuscript then ends with a conclusion that ties the findings back to the initial thesis and, as in the beginning of the introduction, take the reader back to the overall importance of the findings.

Giving Credit

According to the APA manual (2009), “scientific knowledge represents the accomplishments of many researchers over time. A critical part of the writing process is helping readers place your contribution in context by citing the researchers who influenced you” (p. 169). There are two parts to giving credit: in-text citations and references. The in-text citations indicate where in a manuscript an author has discussed past work. A reference section appears toward the end of an manuscript written in APA style and lists the papers cited in the text; only the papers that are cited in the text should appear in the references (hence, the reference section is not a bibliography or listing of all resources). Chapter 6 of the APA manual discusses how to format in-text citations. In addition, the APA created an online source called APA Style that can help authors with formatting citations and references from Internet sources. Chapter 7 of the APA manual discusses how to format the reference section.

APA Format and Writing Resources

This entry provides a very brief explanation of APA format. There are many other resources to help authors, including, of course, the APA manual itself. Many templates, or sample papers, can be found online. These provide a detailed visual guide that an author can use in formatting a manuscript. Some of the most popular templates and checklists are listed in the Further Readings section of this entry, along with summaries of the APA manual and other guides to academic writing.

Timothy Franz

See also Abstracts; American Psychological Association; Journal Articles; Literature Review; Methods Section; Results Section

Further Readings

American Psychological Association. (2016). Sample one-experiment paper. Retrieved June 26, 2016, from http://www.apastyle.org/manual/related/sample-experiment-paper-1.pdf

American Psychological Association. (2009). Concise rules of APA style. Washington, DC: Author.

American Psychological Association. (2009). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

Ashford University. (2013). APA essay checklist for students. Retrieved June 26, 2016, from https://awc.ashford.edu/cd-apa-checklist.html

Darley, J. M., Zanna, M. P., & Roediger, H. L.III (2004). The compleat academic: A career guide (2nd ed.). Washington, DC: American Psychological Association.

Hairston, M., Ruszkiewicz, J., & Friend, C. (2002). The Scott, Foresman handbook for writers (6th ed., pp. 791, 795, 801). New York, NY: Longman. Retrieved June 26, 2016, from https://www.slu.edu/Documents/student_development/student_success_center/

Houghton, P. M., & Houghton, T. J. (2009). APA the easy way: A quick and simplified guide to the APA writing style (2nd ed.). Ann Arbor, MI: XanEdu Publishing.

Off Campus Library Services, Indiana Wesleyan University. (2013). APA style checklist. Retrieved June 26, 2016, from http://www2.indwes.edu/apa/apastylechecklist.pdf

Office of Research and Public Service, The University of Tennessee, Knoxville. (n.d.). APA 6.0 templates for Microsoft Word. Retrieved July 1, 2016, from https://www.sworps.tennessee.edu/training/APA_6_0/resources/apa_doc_temp

Purdue University Online Writing Lab. (n.d.). Microsoft Word: Sample APA document. Retrieved June 26, 2016, from https://owl.english.purdue.edu/media/pdf/20090212013008_560.pdf

Rossiter, J. (2010). The APA pocket handbook: Rules for format … documentation. Port St. Lucie, FL: DW Publishing.

Schwartz, B. M., Landrum, R. E., & Gurung, R. A. R. (2016). An easy guide to APA style (3rd ed.). Thousand Oaks, CA: Sage.

Silva, P. (2007). How to write a lot: A practical guide to productive academic writing. Washington, DC: American Psychological Association.

Strunk, W., & White, E. B. (1999). The elements of style (4th ed.). Boston, MA: Allyn … Bacon.

Sword, H. (2012). Stylish academic writing. Cambridge, MA: Harvard University Press.

Websites

APA Style: www.apastyle.org

David Morgan David Morgan Morgan, David

Applied Behavior Analysis Applied behavior analysis

100

104

Applied Behavior Analysis

Applied behavior analysis is a growing profession devoted to the application of basic learning principles to socially significant behavior occurring in many natural environments, including the home, school, workplace, and other public venues. Founded on well-documented learning processes, such as respondent and operant learning, applied behavior analysis involves direct observation and recording of relevant target behaviors, systematic and continuous data collection, and implementation of interventions designed to address behavioral deficiencies or excesses.

Behaviors targeted for intervention with applied behavior analysis are of practical, not theoretical, concern, and are usually identified by pertinent stakeholders, their teachers, parents, siblings, peers, coworkers, or the clients themselves. Applied behavior analysis has a significant track record of evidence-based treatments, especially in the domain of developmental disabilities. This entry discusses the history and development of applied behavior analysis, its major features, and areas where it is increasingly being used.

History and Development of Applied Behavior Analysis

Applied behavior analysis emerged in the 1960s as an extension of the laboratory science of behavior founded by B. F. Skinner in the late 1930s. Skinner's research on operant conditioning, which he termed the experimental analysis of behavior, identified foundational principles of learning, including reinforcement, extinction, punishment, and stimulus control (generalization and discrimination). Although primarily responsible for the development of the basic science, Skinner himself saw clear implications of operant principles for behavior in the real-world settings and, in the late 1950s, embarked on a program of research aimed at identifying more effective instructional tactics for professional educators. Being an amateur engineer, Skinner fashioned early teaching machines capable of systematically programming instructional contingencies to enhance student mastery of academic concepts. Skinner’s laboratory analysis of behavior had revealed that any behavior could be conceptualized within the context of a three-term contingency, consisting of antecedent environmental events, the behavior of interest, and consequences that follow behavior.

In instructional design, academic materials, such as short written text or questions, served as antecedents, an active and objective response from the student served as behavior, and feedback regarding the accuracy of the student’s response served as consequential stimulation. Using standardized programs, Skinner was able to show that students were able to efficiently master a number of academic skills, including math and science, rapidly and fluently as a result of the frequent active responding and immediate feedback characterizing such programmed instruction.

By the 1970s, considerable research had been conducted on programmed instruction and other behaviorally based instructional methods, including Fred Keller’s personalized system of instruction. Meta-analyses of these research programs showed the instructional methods to be far more effective than traditional instructional methods, especially those dominated by instructor lectures. In fact, Project Follow Through, the largest educational experiment ever conducted, begun in 1967 as a part of President Johnson’s War on Poverty, amassed strong evidence of the effectiveness of behaviorally oriented instruction. When pitted against nearly a dozen alternative educational tactics, methods of behavioral instruction developed at the University of Oregon and the University of Kansas produced substantially larger student gains in both basic academic skill development and affective measures, such as self-concept.

By the late 1960s, applications of basic learning principles to various target behaviors and settings had grown sufficiently to justify a specialized journal, and in 1968, the inaugural volume of *Journal of Applied Behavior Analysis* was published. To this day, *Journal of Applied Behavior Analysis* remains the preeminent outlet for research in the field. By the 1970s, applications of behavioral principles had become common, especially in the areas of developmental disabilities and autism.

Autism, characterized by severe deficiencies in language and social behavior and excessive stereotypic behavior and self-injury, had historically proven unresponsive to efforts at traditional therapy, and many individuals with this diagnosis lived most of their lives in institutions. During the late 1960s and early 1970s, however, behavioral psychologist O. Ivar Lovaas developed systematic programs for addressing both the behavioral deficits and excesses of children with autism, providing the first evidence that behavioral interventions could effectively enhance the independence and autonomy of such clients. In the ensuing decades, a significant database emerged replicating Lovaas's work and establishing applied behavior analysis as the only evidence-based treatment for autism, as acknowledged by both the surgeon general of the United States and the health and education departments in states including California, Maine, and New York.

Training in applied behavior analysis now occurs at hundreds of colleges and universities worldwide, and practicing behavior analysts must hold either a bachelor's or master's degree, have taken significant coursework, have practical experience at the undergraduate and/or graduate level, and possess certification from the Behavior Analyst Certification Board, founded in 1998. Demand for trained behavior analysts increased in response to an increased prevalence of autism diagnoses during the first decade of the 21st century.

Major Features of Applied Behavior Analysis

Regardless of the specific client, setting, or behavior being addressed, behavior analysts conceptualize and implement assessment and intervention protocols consistent with certain basic principles. Donald M. Baer, Montrose M. Wolf, and Todd R. Risley articulated the principal dimensions of applied behavior analysis in 1968, and they remain central attributes of the profession today: *applied, behavioral, analytic, technological, conceptual, effective,* and *generalizable*.

Applied

Although informed by the empirical data generated by a basic science of behavior, applied behavior analysts develop and deliver interventions for socially significant behavior occurring in natural settings. The range of behaviors to which basic principles have been applied, and the circumstances under which interventions have been delivered, is truly remarkable, running the gamut of human behavior. A very brief list of such applications would include the following:

Teaching academic skills to both normally developing and developmentally delayed students.

Reducing stereotypic and self-injurious behavior in individuals on the autism spectrum.

Teaching both verbal and nonverbal communication skills to noncommunicative clients.

Enhancing physical exercise or medical compliance in medical patients.

Teaching basic self-care skills (e.g., dressing, cooking, and cleaning) to developmentally disabled clients.

Teaching fire and gun safety to children.

Teaching children to respond effectively and assertively to potential abductors.

Teaching basic job skills, including interviewing, eye contact, and conversational skills.

Improving peer, sibling, and/or coworker interaction skills.

Teaching effective use of contemporary technology, such as tablets, computers, household appliances, and entertainment technology.

Behavioral

Applied behavior analysis primarily targets behavior that can be readily observed and measured, as opposed to such private activity as thinking, imagining, perceiving, and so on. Although these “private events” can be conceptualized as behavior, actions that directly operate on the environment are both easier to observe systematically and more likely to have real-world, pragmatic value for the client. It is possible to devise observational and measurement tactics, for instance, for all of the behaviors that are part of the applications listed earlier, all of which can take on considerable social significance for the client. Direct measurement of behavior contributes importantly to the scientific status of applied behavior analysis and makes drawing inferences about intervention effectiveness more tenable.

Analytic

Behavior analysts deliver clearly articulated interventions to alter client behavior while simultaneously measuring target behavior in a continuous manner. By collecting data in real time, the behavior analyst is capable of identifying changes in behavior that are functionally related to the intervention. In order to establish that behavior change occurred in response to the treatment, and not some other variable, behavior analysts build multiple replications into treatment protocols, collecting relevant data systematically under all conditions. Replications can be carried out with the same client, sometimes across different settings or behaviors, and can also be carried out across multiple clients in order to establish the reliability of the functional relationship between interventions and behavior change.

Technological

In addition to collecting data systematically during behavioral interventions, behavior analysts describe their intervention tactics in clear and concise language and in a manner that could be readily carried out by others if necessary. Although many behavioral interventions are not carried out as part of a formal research process, behavior analysts do collect data throughout clinical protocols because making decisions about changes in treatment or about treatment effectiveness in the absence of supporting data is considered unethical. Because the actual behavior plans being implemented are described in significant detail, they can be replicated readily by other clinicians or researchers. This practice is characteristic of many mature sciences, especially those that have spawned applied technologies.

Conceptual

The interventions implemented by applied behavior analysts are not designed idiosyncratically by the individual clinician nor are they reflective of a commitment to an eclectic or generic behavioral science perspective. Instead, they are informed by a consistent dependence on the conceptual moorings of the basic science, the experimental analysis of behavior. This science, begun in the late 1930s, produced a cumulative database attesting to the role played by fundamental principles of behavior in natural settings. Such processes as reinforcement, punishment, extinction, generalization, and discrimination are known to underlie almost all adaptive behavior, both human and nonhuman. These foundational concepts subsequently define the parameters of the treatment plans developed and implemented by practicing behavior analysts.

Effective

As described in previous sections, ongoing data collection and analysis assist professional behavior analysts in determining whether interventions have produced effective outcomes for clients. In addition, behavior analysts purposefully seek out the opinions of important stakeholders, such as parents, siblings, peers, or coworkers, in evaluating client progress. The process of asking those who know the client best to offer feedback regarding the success of the behavior program is called social validation. This practice is important because these significant others will likely be substantial sources of reinforcement for the client, and reciprocal interactions between them and the client will have repercussions for the client’s long-term behavior change.

Generalizable

It has long been known that the effects of therapeutic interventions delivered in highly specialized environments, for instance schools or hospitals, often fail to transfer outside the treatment environment. Behavior analysts are especially adept at ensuring that such failures are minimized, as their professional skills include a working knowledge of tactics for enhancing stimulus control, including client generalization of learned skills across varying settings. Indeed, a behavioral intervention is not considered successful unless changes in client behavior have been formally assessed in a multitude of environments in which the behavior is likely to be important. Behavior analysts usually build into an intervention-specific tactics for ensuring generalization of client functioning in such environments.

Contemporary and Future Directions

Applied behavior analysis, initially incubated within the parent discipline of psychology, eventually emerged as a distinct profession. As is true of any relatively new profession, applied behavior analysis suffered early growing pains. National certification for professionals did not emerge until 2000, and although the Behavior Analyst Certification Board has now certified thousands of practitioners, the professional title is not yet familiar to others, including related professionals (e.g., psychologists, social workers, occupational and related professionals (e.g., psychologists, social workers, occupational and physical therapists) or the insurance companies responsible for remunerating most health-care professionals. This situation is changing and will no doubt continue to do so with the increasing number of certificants, national and international training programs, and states that have set up formal licensing boards to oversee the profession.

In addition to the basic growth of the profession, there has been a corresponding broadening of the kinds of behaviors, clients, and settings to which behavior analytic principles have been applied. The behavior principles targeted by applied behavior analysts are pervasive and influence nearly every action we take from the mundane to the profound. Consequently, applied behavior analysts have spread their professional wings in demonstrating the applicability of their principles to a variety of real-world circumstances, from weight loss to pet training.

In the area of education, contemporary behavior analysts can take advantage of the capabilities brought by the microcomputer and the Internet to use powerful sounds and graphics to reinforce academic responding. Computerized versions of Keller’s personalized system of instruction proved both easily adapted and successful in teaching college concepts and skills, including psychological principles and computer programming. In addition, preschoolers utilize behaviorally based computer reading programs. Effective academic contingencies, characterized by frequent active responding by the student, immediate and powerful feedback in real time, and nearly continuous assessment, are more realizable than ever before.

The use of basic behavior principles to encourage patients to adhere to medical regimens or exercise programs has long been a mainstay of applied behavior analysis. As in other behavioral domains, advances in technology have altered the landscape. In exergaming, for example, electronic games are designed to require high levels of physical exertion to make contact with the powerful reinforcers of the game environment. In studies of school-aged children, exergaming stations produced increased enthusiasm and higher levels of activity than stations employing more conventional physical education activities.

In addition, behavior analysts are leveraging the Internet, social media in particular, to create powerful support groups and social reinforcement for those facing a range of personal challenges, including medical treatments, substance abuse, weight loss, and gambling. In Europe, positive results were seen from a major effort at increasing schoolchildren’s consumption of high-quality foods, utilizing personalized token systems. Technologies exist today that allow any individual to monitor a large array of health indicators, such as steps taken, calories consumed, and heart rate, and contingency management programs developed by behavior analysts have helped to put teeth into individual resolutions and self-improvement programs.

David Morgan

See also ABA Designs; Behaviorism; Experimental Designs; Generalizability; Learning Theories; Reinforcement; Response Rate; School-Wide Positive Behavioral Support; Single-Case Research; Time Series Analysis

Further Readings

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1(1), 91–97.

Chance, P. (2006). First course in applied behavior analysis. Long Grove, IL: Waveland.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Applied behavior analysis (2nd ed.). Upper Saddle River, NJ: Pearson.

Fisher, W. W., Piazza, C. C., & Roane, H. S. (2011). Handbook of applied behavior analysis. New York, NY: Guilford.

Vargas, J. S. (2009). Behavior analysis for effective teaching. New York, NY: Routledge.

Elizabeth H. McEneaney Elizabeth H. McEneaney McEneaney, Elizabeth H.

Applied Research

Applied research

104

107

Applied Research

Applied research is an umbrella term that includes various kinds of systematic, empirical research that aims to solve particular problems. In the broad field of education, these problems are those that would arise not only in all aspects of teaching and learning across the lifespan but also from organizational and policy dimensions related to efforts to educate. Applied research seeks to ameliorate problems in education through collecting and analyzing data that directly inform organizational and institutional decision making.

As both public entities and private foundations devote increasing resources to educational programs, the evaluative role for applied research has grown, consistent with and at the same time as calls for accountability of schools and specific educational programs have increased. This entry presents research traditions that typically fall under the category of applied research, the relationship between applied and basic research, the methods and approaches used in applied research, the dissemination of applied research, and examples and criticisms of applied research.

Types of Applied Research

Many forms of research in education are within the purview of applied research. Evaluation studies are almost always applied research with the problem being to what extent a new or existing policy or program achieves intended goals (outcomes) and why (process). Researchers in educational measurement are frequently involved in applied research to develop measures of key constructs deemed important to the teaching and learning process while also investigating their validity and reliability, such as in the development of standardized testing instruments to measure learning in particular areas.

Paulo Freire popularized various forms of action research in education, including participatory action research. Community-based participatory research often are applied in their orientation, as well as design-based research (an early example is seen in education in Ann Brown’s “design experiments”) and the teacher (or educational practitioner) research movement highlighted by Joe Kincheloe, Marilyn Cochran-Smith, and Susan Lytle. An influential movement, recently spearheaded by Anthony Bryk and others, highlights the need for more “improvement science” that uses rapid, iterative cycles of Plan-Study-Do-Act to produce knowledge in naturalistic settings to enhance school efficacy in ways that can be used widely, that is, brought to scale. Knowledge produced through applied research in education may therefore be used to help develop new or revised products, strategies, procedures, or technologies.

Relationship Between Applied and Basic Research

The problem to be investigated in an applied research project could be selected by the researchers themselves or by nonresearcher stakeholders or clients. This can be compared to so-called pure or basic research in which the research problem is typically one that arises from a gap or contradiction identified in the existing research literature. Basic research always seeks to add to the theoretical research base, but applied research may not have that aim as a primary purpose.

Although the distinction between applied and basic research is useful in many ways, it is sometimes drawn too simplistically. In the process of solving problems and guiding decision making, applied research may also generate theoretical knowledge that is characteristic of basic research. It is also vital to understand the ways that these two forms of research have a reciprocal and iterative relationship over time within any particular field of education research. The design and development of an applied research project usually is supported at least in part by prior basic research results specifying key concepts and elaborating relationships between them, as well as pointing the way to appropriate research methods and designs.

Basic research, particularly in the field of education, often concludes with statements about the implications for practice of the research results, which often generates topics for applied research. In fact, many major funders of basic educational research, both government and private foundations alike, seek to improve “knowledge transfer” by requiring funded studies to consider implications for practice of the research, thus narrowing the distance from more theoretical, basic research to its application. In addition, there has been a reduction in the separation between the worlds of research and practice, such that practitioner-led applied research has become increasingly influential. Notions of reflective practice in education have come to include components of self-evaluation, such as a teacher systematically collecting and analyzing data on the effectiveness of a new way to teach a topic, that are in fact often small-scale examples of applied research. In short, just as the boundary between basic and applied research has blurred, understandings of who has the appropriate authority and expertise to conduct research, particularly applied, context-based research, has broadened.

Approaches and Research Methods

Applied research in education uses a range of approaches and research methods, employing various means to convey results to appropriate audiences. In solving a given problem, applied research can be exploratory, or it can aim to explain patterns (often with the aim of generating accurate predictions), or it can confirm an expected or intended result or prior finding. With the aim of finding a solution to a particular problem, applied research tends not to adhere to strong paradigmatic assumptions such as those outlined by Yvonna Lincoln and Egon Guba. Instead, the pragmatic imperative to solve the problem and guide decision making frames the work, leading some to call applied research “postparadigmatic.”

Although often drawing on quantitative research methods, applied research can be purely qualitative or invoke mixed method approaches. The primary goal of applied research is not necessarily to generate generalizable results, but rather for evidence-based findings to be immediately applicable to a particular context. This is especially true when the problem has been selected by a group of stakeholders or clients based on their particular needs and circumstances.

Dissemination

To produce what some might call “actionable knowledge” applied in particular educational contexts, applied researchers need to pay special attention to the audience for their work and to the manner in which findings are conveyed. This may to some extent explain the preponderance of quantitative methods in action research, and it also suggests that much applied research is disseminated not through academic journals but through practitioner-oriented publications and magazines, think tank white papers, position papers by state/provincial departments of education, blogs, and the like.

Forums such as professional conferences and journals allow teachers, other educators, and administrators to disseminate results of practitioner-led applied research projects in order to share knowledge about how to solve common problems with their relevant professional communities. These venues may or may not conduct blinded peer-review of the research. In fact, to the extent that applied research is truly intended to solve a particular problem in a particular context, there may be no need to disseminate results at all beyond that setting. For example, a school district conducting applied research to solve a problem of low parental involvement may share results with district personnel only. Nevertheless, some peer-reviewed journals in education are especially known for publishing applied research, including the *Journal of Education for Students Placed at Risk* and the *Journal of Applied Research in Higher Education*.

Examples and Quality of Applied Research

In the U.S. context, some notable examples of applied research have been a series of reports on the impact of the early childhood education program Head Start, the Coleman report on the effects of racially segregated education, and studies of the effects of school voucher programs on outcomes for urban youth. Each of these represents an effort to delve into a politically sensitive and contentious issue. In judging the quality of applied research projects in these and other areas, Alis Oancea and John Furlong have argued for a multidimensional framework, including not only quality in an epistemic sense common to all research, such as the use of ethical and robust research design and methods to produce trustworthy results, but also considering issues of value and capacity building. Quality applied research from this perspective is of value to users when it is responsive to their needs and contexts and is presented in an accessible manner.

To enable impact, applied researchers who are not themselves practitioners are wise to establish links with communities of practice early in the research process. Such a strategy is likely to enhance quality by improving the capacity building effect of the applied research, improving the receptiveness of researchers to take the complexity of practice into account, and allowing practitioners to incorporate lessons gleaned from applied research. Finally, the quality of applied research may also be judged by its value for money, that is, cost-effectiveness and transparent, rigorous accounting.

Concerns About Applied Research

Although applied research is widely embraced as a means for enhancing decision making, solving problems, and ultimately improving programs in education, some concerns do exist. Although the typical lack of paradigmatic grounding is not usually seen as an issue, the tendency for applied research to be atheoretical has been criticized in part because the lack of theory may produce a fragmented body of results. In quantitative applied research, the lack of theory to justify particular hypotheses can also lead to a higher likelihood of false-positive findings of statistical significance, for instance.

In an early statement of skepticism about applied research in education, David Cohen and Michael Garet challenged the tenet that applied social research, including applied research in education, provides authoritative knowledge on the costs and consequences of programs, noting that it, like basic research, does not typically reduce intellectual conflict about the value of one approach over another. Similarly, calls for applied research can sometimes be used to justify delays in program implementation. Finally, applied research sometimes functions to lend legitimacy to decisions that have already been made, following the decision making, rather than driving it.

Elizabeth H. McEneaney

See also Action Research; Data-Driven Decision Making; Design-Based Research; Improvement Science Research; Mixed Methods Research; Paradigm Shift

Further Readings

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. The Journal of the Learning Sciences, 2(2), 141–178.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). Learning to improve: How America's schools can get better at getting better. Harvard Education Press.

Cochran-Smith, M., & Lytle, S. L. (Eds.). (1993). Inside/outside: Teacher research and knowledge. Teachers College Press.

Cohen, D., & Garet, M. (1975). Reforming educational policy with applied social research. Harvard Educational Review, 45(1), 17–43.

Furlong, J., & Oancea, A. (Eds.). (2013). Assessing quality in applied and practice-based research in education: Continuing the debate. New York, NY: Routledge.

Greenwood, D. J., & Levin, M. (2006). Introduction to action research: Social research for social change. Sage.

Howell, W. G., & Peterson, P. E. (2006). The education gap: Vouchers and urban schools. Brookings Institution Press.

Kincheloe, J. L. (2012). Teachers as researchers (classic ed.): Qualitative inquiry as a path to empowerment. New York, NY: Routledge.

Schweinhart, L. J. (1993). Significant benefits: The high/scope Perry Preschool Study through age 27. Monographs of the High/Scope Educational Research Foundation No. 10. Ypsilianti, MI: High/Scope Educational Research Foundation.

U.S. Department of Health and Human Services, Administration for Children and Families (January 2010). Head Start Impact Study. Final Report. Washington, DC.

Gervase R. Bushe Gervase R. Bushe Bushe, Gervase R.

Appreciative Inquiry

Appreciative inquiry

107

110

Appreciative Inquiry

Appreciative inquiry (AI) is an organizational development method grounded in social constructionist theory that engages stakeholders in an inquiry into their collective strengths, assets, and what is working as a precursor to identifying what they want more of and how to achieve that. It has proven to be a popular and successful transformational change approach. Some researchers advocate its use as a participatory evaluation method under certain conditions, particularly when there is a desire to improve a program or process. This entry describes first the theory and practice of AI and then its use as a method of evaluation.

AI Theory

AI was originally developed in the mid-1980s by David Cooperrider, Frank Barrett, Ron Fry, and Suresh Srivastva of the Organizational Behavior Department at Case Western Reserve University, in response to the dominant use of problem-solving in *action research*, a method of improving social systems by involving system members in self-study. They noted the lack of new theory generated by action research and argued that it engendered an unhelpful bias toward seeing organizations as problems to be solved. They argued that using positivistic, scientific assumptions and methods to improve groups and organizations made the mistake of treating people like simple stimulus–response mechanisms, ignoring how so much of collective life is based on sensemaking, narratives, and beliefs about the future.

Cooperrider, Barrett, Fry, and Srivastva described how assessing groups and organizations against predetermined models of health or dysfunction tended to create the very issues they were supposed to uncover and argued, instead, that there were no inherently correct ways to organize; our methods of organizing are limited only by human imagination and our collective agreements. A method of study that was interested in improving organizations, they argued, would have to lead stakeholders to produce new ideas grounded in their collective hopes and desires and that would most likely emerge if they first inquired, appreciatively, into what gave life and vitality to their organization.

After about 15 years of experimentation and study, a set of five principles of AI were developed that are now widely accepted:

The constructionist principle

What we believe determines what we do, and thought and action emerge out of relationships. People coconstruct the organizations they inhabit through conversations and day-to-day interactions. The purpose of inquiry is to stimulate new thoughts, stories, and beliefs that create new choices for decisions and actions.

The principle of simultaneity

As we inquire into human systems, we change them. The seeds of change, what is discovered and learned, are implicit in the very first questions asked. Questions are never neutral, and organizations move in the direction of the questions most persistently and passionately discussed.

The poetic principle

Organizational life is expressed through language and narratives, the story lines people use to make sense of what is taking place. Words are not passive transmitters of meaning. The words and topics chosen for inquiry have an impact far beyond just the words themselves; they evoke feelings, understandings, and worlds of meaning. Always use words that point to, enliven, and inspire the best in people.

The anticipatory principle

Choices made today are guided by beliefs about the future. The creation of positive imagery of a desirable future on a collective basis and the design of actions to take toward that future, refashions anticipatory reality.

The positive principle

Momentum and sustainable change require positive emotions and social bonding. Hope, excitement, inspiration, camaraderie, and joy increase openness to new ideas and different people, creativity, and cognitive flexibility. They also promote trust and good relationships between people, particularly between groups in conflict, required for collective inquiry and participatory change.

AI Method

During the 1990s, the creators of AI resisted developing formulas for how to do AI, instead of encouraging adoption of the principles and experimenting with ways to implement them in practice. As a result, AI is practiced in numerous ways. However, by 2000, Cooperrider and Diana Whitney developed the 4-D model, a set of four phases (discovery, dream, design, and destiny/deployment) that is now widely utilized, while Jane Magruder Watkins and Bernard Mohr popularized a similar 4-I model that has been embraced by many using AI for evaluation. The model is as follows:

> *Initiate*: Decide how and when to introduce AI. Determine the overall focus of the inquiry. Decide on the appropriate structure and leadership for the inquiry.
> *Inquire*: Develop an interview guide and engage as many stakeholders as possible in a search for what is known about the program, process, group, or organization at its best.
> *Imagine*: Work with the information collected during the interviews to catalyze conversations about collective aspirations for the program, process, group, or organization’s future.
> *Innovate*: Engage stakeholders in proposing activities and projects to move in the direction of those aspirations. Develop and implement processes to encourage taking action and embedding successful innovations.

Various architectures for engagement have been used, but most studies of transformational change report using some variation of the so-called AI Summit, in which a large group of stakeholders go through the 4-Ds (or inquire, imagine, and innovate) over 2–4 days. In an ideal AI process, all the stakeholders would gather to inquire into the best of what they know about the focal topic, understand and express their collective aspirations for the focal topic, foster the emergence of small groups of motivated people with common ideas and interests to develop proposals or plans for actions, and leave the summit with clear ideas about what they will do next. AI Summits with thousands of attendees have been successfully hosted.

Appreciative Interviews

A key innovation of AI is to gather stories about stakeholders’ peak experiences at the beginning of interviews during the discovery or inquire phases. The generic AI interview asks, “Thinking about your history in this organization, please tell me the story about the time when you felt most alive, most involved, or most excited to be a part of this organization.” The story is probed to understand what brings life and vitality to the organization. The rest of the interview guide asks what they most value about the organization, what their dreams or wishes for the future of the organization are, and what they think needs to happen for the organization to move in that direction.

AI interviews often focus on something specific the organization wants to improve, such as customer service, sustainability, or product innovation. The interview is then constructed similarly to a generic AI interview but refocused accordingly. In general, opening the interview by asking for a personally meaningful, “best of” story about the focus of the inquiry is considered essential for an AI.

Interviews can be done by an individual or small team, but studies have found that getting stakeholders to interview each other increases the amount of change produced by helping to build relationships and catalyze changes in conversations and narratives that occur after the interviews. Many research projects and evaluation studies, however, use just an individual or small group to do the interviews and don’t actually do the other AI phases. Instead, they use the information collected as a qualitative data set that is analyzed using standard qualitative research methods. Most AI advocates believe these should not be called appreciative inquiries, but instead call them appreciative interviews.

Using AI for Evaluation

AI for evaluation has been used in a variety of situations, including to evaluate the effectiveness of foreign aid programs, social service program delivery, the effectiveness of training programs in corporations, audits for compliance with quality management system standards, and other organizational processes. There are many descriptions of the use of AI in educational settings, both from an evaluation perspective and from an organizational development perspective.

Proponents of the inclusion of AI as an evaluation method list it as a learning and development type of evaluation or within the category of participatory evaluation. At least five overall benefits have been advanced for using an appreciative approach to evaluation.

1. It generates information that has the maximum potential of being used. The inclusion of many stakeholders in conversations about what matters to them makes it more likely they will embrace the results of the evaluation and do something with it. The process itself can be motivating and energizing.
2. Better information can be gathered more quickly. Large group formats allow for the generation of large amounts of data in short time periods. Personal stories and scenarios provide very rich data for analysis. The collection of real stories from real users emphasizes what the users want as opposed to what designers believe they need.
3. It makes it more likely that groups in conflict or who do not trust each other, or do not trust the evaluators, will engage in the evaluation. By having people focus on what they like and want more of, people who might otherwise feel anxious or cautious about truthfully discussing problems and deficiencies are more likely to engage honestly.
4. For similar reasons to Number 3, it can be useful when there is a need to generate support for the evaluation, perhaps because of past evaluation failures, or fear or skepticism toward the evaluation. In situations where the group being evaluated has a history of oppression, asking for their stories of things at their best, and assessment of what works, is often experienced as being treated respectfully.
5. For any system that has democratic, pluralistic, and/or empowerment agendas, it is a more congruent evaluation approach. It can increase evaluation capacity of stakeholders and the system. AI’s use of storytelling makes it particularly effective in cultures with oral history traditions.

Being a sociorationalist, postmodern approach that challenges the validity of scientific assumptions for studying people, AI does not fit well with traditional criteria for assessing evaluations such as independence, neutrality, and minimal bias. AI advocates argue that it is impossible for anyone to enter a social system bias. AI advocates argue that it is impossible for anyone to enter a social system from a neutral stance and that there is no such thing as independence; by their very presence, evaluators are influencing the social systems in which they enter. While concerns have been expressed that a focus on the positive may undermine the appearance of neutrality and, therefore, the collection of valid information, experience in the field suggests just the opposite—that people are willing to be more honest when asked about their opinions of what works than when asked about problems and failures.

Another concern is that with its focus on the positive, AI will miss seeing and reporting important problems. Researchers report, however, that asking questions such as “what is your wish for this project?” or “what is your dream for this organization?” or “how could we improve this process?” elicit all the same issues that asking directly about problems would surface but without feelings of rancor or recrimination.

All advocates emphasize the use of AI in specific circumstances and not in others. There is widespread agreement that AI is worth considering when it matches the values and culture of those who will use the evaluation and when the purpose of the evaluation is to develop and improve whatever is being evaluated. Conversely, AI is not likely to be a useful method when the values and culture of the target group do not favor a participatory approach, or there is a desire for mainly quantitative data, and/or when one of the aims of the evaluation is to terminate a process or program.

It has also been noted that successful use of AI requires specialized knowledge and skill sets not normally associated with evaluation training. Training in AI as a change method is available, but training in use of AI as an evaluation method is rare.

Gervase R. Bushe

See also Action Research; Collaborative Evaluation; Constructivist Approach; Democratic Evaluation; Evaluation Capacity Building; Narrative Research; Postpositivism

Further Readings

Barrett, F. J., & Fry, R. E. (2005). Appreciative inquiry: A positive approach to building cooperative capacity. Chagrin Falls, OH: Taos Institute.

Bushe, G. R. (2012). Appreciative inquiry: Theory and critique. In D. Boje, B. Burnes, & J. Hassard (Eds.), The Routledge companion to organizational change (pp. 87–103). Oxford, UK: Routledge.

Coghlan, A. T., Preskill, H., & Catsambas, T. T. (Guest Eds.). (2003, Winter). New directions for evaluation, Vol. 100 (Special issue on appreciative inquiry in evaluation).

Cooperrider, D. L., Barrett, F., & Srivastva, S. (1995). Social construction and appreciative inquiry: A journey in organizational theory. In D. Hosking, P. Dachler, & K. Gergen (Eds.), Management and organization: Relational alternatives to individualism (pp. 157–200). Aldershot, UK: Avebury.

Cooperrider, D. L., & Srivastva, S. (1987). Appreciative inquiry in organizational life. In R. W. Woodman & W. A. Pasmore (Eds.), Research in organizational change and development, Vol. 1 (pp. 129–169). Stamford, CT: JAI Press.

Cooperrider, D. L., Whitney, D., & Stavros, J. M. (2008). Appreciative inquiry handbook (2nd ed.). Brunswick, OH: Crown Custom.

Dunlap, C. A. (2008). Effective evaluation through appreciative inquiry. Performance Improvement, 47(2), 23–29. doi:10.1002/pfi.181

Ludema, J. D., Whitney, D., Mohr, B. J., & Griffen, T. J. (2003). The appreciative inquiry summit. San Francisco, CA: Berrett-Koehler.

Preskill, H. S., & Catsambas, T. T. (2006). Reframing evaluation through appreciative inquiry. Thousand Oaks, CA: Sage.

Watkins, J. M., Mohr, B. J., & Kelly, R. (2011). Appreciative inquiry: Change at the speed of imagination (2nd ed.). San Francisco, CA: Pfeiffer-Wiley.

Phillip L. Ackerman Phillip L. Ackerman Ackerman, Phillip L.

Aptitude Tests

Aptitude tests

110

114

Aptitude Tests

The term *aptitude,* according to most dictionaries, is derived from the Latin term *aptitudo,* meaning fitness. The psychological use of the term is similar in that it has traditionally referred to a potential for acquiring knowledge or skill. Traditionally, aptitudes are described as sets of characteristics that relate to an individual’s ability to acquire knowledge or skills in the context of some training or educational program. There are two important aspects of aptitude to keep in mind. First, aptitudes are present conditions (i.e., existing at the time they are measured). Second, there is nothing inherent in the concept of aptitudes that says whether they are inherited or acquired or represent some combination of heredity and environmental influences. Also, aptitude tests do not directly assess an individual’s future success; they are meant to assess aspects of the individual that are indicators of future success. That is, these measures are used to provide a probability estimate of an individual’s success in a particular training or educational program. Although the meaning of *aptitude* is well delineated, there is much controversy over how to distinguish aptitude tests from other kinds of psychometric measures, specifically intelligence and achievement tests, partly because the major salient difference between intelligence, aptitude, and achievement tests has to do with the purpose of testing rather than with the content of the tests. What makes an assessment instrument an aptitude test rather than an intelligence or achievement test is mainly the future orientation of the predictions to be made from the test scores.

Historians generally date the movement of modern psychological testing from the 1905 work by Alfred Binet and Théodore Simon in developing a set of measures to assess intelligence. The Binet-Simon measures, and especially the English translation and refinement made by Lewis Terman in 1916, called the

English translation and refinement made by Lewis Terman in 1916, called the Stanford-Binet, are in widespread use even today. Few adults living in industrialized countries today have avoided taking at least one test of intelligence during their school years. Intelligence tests were designed with the goal of predicting school success. Thus, in terms of the definition of aptitude provided above, when the purpose of an intelligence test is prediction, then the intelligence test is essentially an aptitude test—although an aptitude test of general academic content (e.g., memory, reasoning, math, and verbal domains). Aptitude tests, however, sample a wider array of talents than those included in most general intelligence measures, especially in the occupational domain. By the late 1910s and early 1920s, dozens of different aptitude tests had been created for prediction of success in a variety of different occupations (e.g., auto mechanic, retail salesmen, waitress, telegrapher, clerk, Hollerith operator, musician, registered nurse).

It is important to distinguish between so-called trade tests and aptitude tests. The distinction rests more on the characteristics of the examinee population than on the content of the tests. That is, when all the examinees can be expected to have similar prior exposure to the knowledge and skills needed to perform well on the test, the test is essentially one of ability or aptitude. But when prior knowledge and skills have an important impact on the examinees’ success on the test, it is essentially an achievement test, or a measure of learned knowledge or skills, rather than an assessment of potential for acquiring such knowledge or skills. For psychologists who design aptitude tests, this is a critical concern. For example, the psychologist must be able to determine whether reading skills are an important determinant of test performance in order to present the test material in a paper-and-pencil format. Intelligence test developers assumed that individual differences in reading skills in young children were possible confounding influences, and so the developers created intelligence tests that did not require a child to know how to read or write. For assessing the aptitude of adults for an office clerk job, however, being able to read would be a prerequisite skill, so a paper-and-pencil aptitude test would certainly be appropriate.

Utility of Aptitude Tests

Aptitude tests are useful for the purpose of aiding educational or occupational selection when there are marked individual differences in the likelihood of success that are, in turn, determined by cognitive, perceptual, or physical abilities. The degree of utility of an aptitude test is determined by three main abilities. The degree of utility of an aptitude test is determined by three major factors: (1) the cost of training or education, (2) the correlation between the aptitude test scores and success on the educational or occupational criterion, and (3) the ratio of the number of applicants to the number of places to be filled. When training is expensive, the cost to the organization of having trainees fail can be an important factor in adopting an aptitude testing program for screening applicants. When training is brief or inexpensive, such as for retail sales or other entry-level positions, the value of aptitude testing is diminished because the cost of accepting applicants who fail is not as burdensome for the organization. The correlation between aptitude test scores and success measures will determine how accurate the prediction of success or failure is. The larger the correlation, the more accurate the prediction. Finally, when there are many more applicants than spaces to be filled, the aptitude test will be more effective in maximizing the overall success rate. In contrast, when there are few applicants for each position, and thus nearly all applicants are accepted, the ranking of applicants by aptitude becomes largely irrelevant.

Two Types of Aptitude Tests

The aptitude tests developed over the past century have generally bifurcated into two different types: jobspecific tests and multiaptitude batteries. Similar to the early aptitude tests described above, jobspecific aptitude tests are typically designed to determine which candidates are best suited to particular occupations. In theory, there can be as many different occupational aptitude tests as there are differentiable occupations. In practice, however, there are common aptitudes underlying many occupations. For example, different kinds of mechanical jobs (e.g., auto mechanic, electronics service repair, assembly worker) may all involve aptitudes for dexterity, fine motor coordination, visual perception, and so on. An organization that wishes to select employees for a particular occupational placement might attempt to identify (through job analysis) what particular aptitudes are needed for successful job performance. The organization, in order to select the applicants who are most likely to succeed in a training program, can then create an aptitude measure that samples these specific aptitudes. Alternatively, among the dozens of commercially available tests, the organization may find an off-the-shelf aptitude measure that covers the most important aptitudes for training success for the particular job.

The other kind of aptitude measure is the multiaptitude battery. These tests are used frequently in educational contexts, and some are used in large-scale employment testing situations. In the educational context, multiaptitude tests may be very general, such as the SAT, which was created in 1926 for selecting high school students for college and university placement. Today, the SAT is one of the most widely used aptitude test batteries in the United States and is administered to more than 1 million students each year. The original SAT assessed only two broad academic aptitudes: verbal and math. The most recent modification of the SAT also includes a writing component. Multiaptitude test batteries can also be designed to provide assessments across several different aptitudes. The first large-scale multiaptitude batteries for use in educational contexts were developed by Louis Leon Thurstone and Thelma Thurstone in the early 1940s and became known as the primary mental abilities battery. Another battery, the differential aptitude tests (DAT), was introduced and is still in use today. The DAT provides scores on eight different aptitudes (verbal, numerical, abstract reasoning, clerical speed and accuracy, mechanical reasoning, spatial relations, spelling, and language use).

There are many more such multiaptitude batteries that are administered in schools throughout the United States each year. Many of these tests do not have the term *aptitude* in their titles, but they are similar in content coverage and in the general purposes of testing. Such educational aptitude batteries are primarily used for counseling purposes. That is, the underlying premise for the utility of these tests is that they allow a parent or counselor to identify an individual student’s aptitude strengths and weaknesses. Usually, the test information is presented as a profile, a set of bar graphs that show where the student stands in respect to some norming group on each of the different aptitudes. Counselors may use this information to help guide the student in a way that either builds on the student’s strengths or attempts to remediate the student’s weaknesses. In practice, however, many of the different aptitudes assessed with these measures are themselves substantially positively correlated because of shared variance with general intelligence. When that happens, it is more difficult to provide a reliable differentiation among the individual’s strengths and weaknesses. This is one of the most intractable problems associated with the counseling use of multiaptitude test batteries.

Multiaptitude batteries for occupational selection tend to be somewhat more useful for selection and classification purposes. (Classification is the process of assigning particular individuals to specific jobs by matching the individual’s profile of aptitude strengths and weaknesses to the job requirements.) The two largest occupational multiaptitude test batteries used in the United States are the

Armed Services Vocational Aptitude Battery and the General Aptitude Test Battery. The Armed Services Vocational Aptitude Battery is used by the U.S. armed forces, and until recently, the General Aptitude Test Battery was used by federal and state employment agencies. In contrast to the multiaptitude batteries described above for educational contexts, these two tests are explicitly linked to a wide variety of specific occupations. For example, when individuals complete the Armed Services Vocational Aptitude Battery, they are each provided with a set of scores that determines their suitability for all the different entry-level occupations within the military. With that information, they can be classified into the occupation in which they are most likely to succeed.

Concerns About Aptitude Tests

Although aptitude tests have been shown to be quite effective predictors of future academic and occupational performance, they have been somewhat controversial because of the meaning inherent in the assessment of potential and because of a wide variety of group differences in performance on standardized aptitude tests. Experience with the SAT, for example, has indicated marked mean score differences between male and female test takers; between Black, White, Hispanic, and Asian American test takers and between socioeconomic status groups. Because the SAT is not traditionally considered to be taken by a representative or random sample of 16-to 18-year- olds (because those students taking the test essentially are self-selected college-bound individuals), group differences on the SAT do not provide direct evidence for overall group differences in academic potential. However, the differences between group means are significant and sometimes substantial, which has led many commentators to question whether and how much the test is associated with prior educational background and other demographic variables. Much of the difficulty centers around the term *potential* associated with aptitude tests, in contrast with achievement measures. That is, if these different groups differ only in terms of academic achievement, there would be perhaps less controversy than there is if the groups are determined to differ in terms of academic potential. Many testing organizations have in fact revised the names of their aptitude tests to remove the term that is associated with potential (e.g., the Scholastic Aptitude Test became the Scholastic Assessment Test in the 1990s and later became known as simply the SAT). At one level, such a change may be cosmetic, but at another level, it does show that testing organizations have come to recognize that one does not need to imbue a test with the notion of potential in order to make predictions about future academic or occupational performance. That is, there is nothing inherently problematic in using an intelligence or achievement test for the same purpose as an aptitude test as long as it taps the same underlying knowledge and skills that are critical for performance on the predicted criterion measure. Given that intelligence, aptitude, and achievement tests assess only current performance, it is ultimately the prediction aspect of a test that makes it an aptitude test. Furthermore, it is fundamentally impossible to know what an individual’s actual potential is for academic or occupational knowledge or skills because it is not possible to know what the universe of instructional or training programs may be. Should methods of instruction or training be improved at some time in the future, even those individuals with relatively lower aptitudes may show marked increases in performance. In that sense, the operational conceptualization of aptitude has to be in terms of whatever instructional or training methods are actually in use at any one time.

Over-and Underachievement

One aspect of aptitude tests that has been very much misunderstood is the notion of over-and underachievement. Typically, the term *overachiever* is given to individuals who have relatively higher scores on achievement tests than they do on aptitude tests, and the term *underachiever* is given to individuals who have relatively lower scores on achievement tests than on aptitude tests. However, given that both aptitude and achievement tests often assess the same underlying knowledge and skills, the choice of labeling one test or another an aptitude or achievement test is generally arbitrary. That means that one could just as easily assert that individuals have higher or lower aptitude in association with their achievement test performance, which makes little conceptual sense but is entirely consistent with the underlying properties of the tests. In fact, given the nature of statistical regression-to-the-mean phenomena, which are associated with taking the difference between any two measures, it is common for individuals with low scores on one test (e.g., aptitude) to have relatively higher scores on the other test (e.g., achievement), and similarly, individuals with higher than average scores on one test will have somewhat lower scores on the other test. The attribution that low-aptitude individuals are often overachievers and high-aptitude individuals are often underachievers is most often an artifact of this regression-to-the-mean phenomenon and thus does not provide any useful diagnostic information. Only extremely large differences between such scores (i.e., differences that significantly exceed the difference attributable to regression-to-the-mean effects) can provide any potential diagnostic information.

Phillip L. Ackerman

Note: Adapted from Ackerman, P. L. (2007). Aptitude Tests. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 1, pp. 39–43). Thousand Oaks, CA: SAGE.

See also Ability Tests; Achievement Tests; SAT; Stanford–Binet Intelligence Scales

Further Readings

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). New York, NY: Prentice Hall.

Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). New York, NY: Harper … Row.

Thorndike, R. L. (1963). The concepts of over-and underachievement. New York, NY: Bureau of Publications, Teachers College, Columbia University.

Ming Fai Pang Ming Fai Pang Pang, Ming Fai

Aptitude-Treatment Interaction Aptitude-treatment interaction

114

119

Aptitude-Treatment Interaction

Aptitude-treatment interaction (ATI), also known as attribute-treatment interaction or trait-treatment interaction, refers to the tenet that treatments or interventions that are well matched to learners' specific aptitudes, attributes, or traits are more effective in helping them to appropriate the object of learning, that is, what the learners are expected to learn. This entry further describes ATI and research in this area. It then details the evolution of ATI research, describes ATI research designs, and discusses implications for educational practice. Finally, it looks at future directions for ATI.

Early descriptions of ATI can be found in both Eastern and Western literature, such as in ancient Chinese and Hebrew writings, early Greek and Roman teachings, and early European philosophies. The basic premise of ATI in educational research is that no single treatment, which refers to any manipulable situational variable such as an instructional approach or a teaching resource, is best for every learner because differences in learners' aptitudes interact with the treatment, which in turn affects the treatment outcomes. In the presence of such an interaction effect, the research question, asking which treatment or intervention is better or more effective, becomes somewhat imprecise or unsophisticated. Instead, researchers in this area focus on which treatment is better or more effective for which group of learners and under what conditions and the underlying reasons or explanation for the existence of such relationships.

Research on ATI can help to determine whether and which particular treatments can be chosen or adapted to best fit specific groups of learners. The ultimate goal of ATI research is to identify and develop treatments that best match the aptitudes of different groups of learners to maximize their learning effectiveness. For instance, learners with high spatial aptitude learn better through an instructional approach that uses visual elaboration than one that uses textual instructional approach that uses visual elaboration than one that uses textual materials only; conversely, learners with high verbal aptitude learn more effectively with an instructional approach that incorporates verbal elaboration than one that uses mainly visual aids. Alternatively, learners with high mathematical aptitude learn music composition more effectively when the instruction focuses on mathematical concepts (e.g., understand the rules for writing chords to create harmony), while learners with high kinesthetic aptitude profit more from actually singing the melody when they start to compose music.

An ATI effect is said to occur if the extent to which the outcomes for two or more treatments, or one treatment over two or more trials, shows a statistically significant difference for learners who differ on one or more of the aptitude variables under investigation. Therefore, adapting the instruction to correspond with the aptitude of learners is optimal in the sense that it can realize the learning potential of each group of learners and at the same time produce better learning outcomes for all of the learners involved.

Evolution of ATI Research

The development of ATI research can be traced back to the seminal work by Lee Cronbach and Richard Snow in which they investigated how differences in some individual aptitude variables might demonstrate strong interaction effects with particular interventions or treatments. Over the years, this research area has been in a state of flux, in which researchers have defined and characterized the notion of aptitude in different ways, devised different kinds of treatments, and made use of different methods to assess the interaction.

Aptitude

Early studies define aptitude as any personal characteristics that increase or decrease the probability of success of the learner who receives a particular treatment. Some further argue that those personal characteristics must be measurable and hypothesized as critical for eliciting a positive response to the treatment. Some researchers contend that these characteristics or aptitudes should not be confined to general intelligence or a fixed set of cognitive abilities but should include noncognitive aptitudes such as personality, motivation, learning attitude, learning style, belief, self-regulation, self-efficacy, and emotion. Some recent studies posit that aptitude can be more content-specific, characterized by the learner’s prior knowledge in relation to the object of learning. In other words, studies have shown that the effectiveness of treatments or interventions is influenced by how many and what critical aspects of the object of learning the learners have already discerned and focused upon.

The personal characteristics or aptitudes that may interact with the treatments are conceived to be many and varied, but not infinite. As learning is a highly complex phenomenon, there is a growing trend for researchers to conceptualize aptitude as a combination of multiple aptitude variables and to analyze the higher order interactions between them. The different combinations of aptitude variables, which are often called aptitude complexes, interact with the treatment and affect the treatment outcomes in different ways. For example, a 2003 study with college students and adults found that a combination of three aptitude variables—self-concept, interest, and motivation—was strongly correlated with domain knowledge and ability measures.

Aptitude is widely recognized as having multiple dimensions, including cognitive, conative, and affective domains. In the early stages, research in this area was more focused on cognitive aptitude, such as general intelligence and cognitive learning styles. Most of these studies aimed to examine whether and in what ways learning outcomes depend on the degree to which the treatments or instructional designs match the learner’s specific cognitive aptitude. The most common and strongest ATI that emerged in early studies was related to general intelligence. Students with high general intelligence and who were relaxed and independent were found to benefit more from less structured learning environments, such as those using a student-centered instructional approach, inductive teaching methods, small group discussion, and discovery learning activities. The reverse was true for students with lower general intelligence and those with anxiety or a high need to conform. These students performed better in highly structured learning environments such as those using a teacher-centered instructional approach and didactic and lecture-based teaching methods.

However, ATIs are highly complex and context bound and can change rapidly, which is why they are so difficult to identify. In fact, no specific ATI effects have been sufficiently established to form the basis for designing instructional practices in the classroom on a large scale, which might be because some significant interactions remain unidentified. Some critics view students’ learning outcomes as so dynamic that they cannot be based on cognitive aptitudes alone and argue that cognitive learning styles may vary within the individual when they attempt different tasks or encounter different situations.

Research has increasingly recognized the importance of elements of the conative domain (such as locus of control, self-regulation, and motivation) and affective domain (such as emotion, anxiety, and self-efficacy) for ATI and has explored the possible interactions of cognitive-conative-affective aptitude complexes with treatment effects. For instance, studies have attempted to investigate the power of emotions and intentions in guiding and managing cognitive processes and how they affect the treatment outcomes. Although the complex interplay between cognitive-conative-affective aptitudes is acknowledged by researchers, the growth and decline of cognitive abilities such as memory, attention, and so on, demonstrate an inverted U-shaped developmental trajectory across the life span in contrast to affect and conation. It is important to bear in mind that aptitudes should not be seen as fixed; they are dynamic in nature and represent the degree of readiness of the learner to learn and to perform well in a particular situation or in a fixed domain at a particular time and can be fostered and changed over time.

Treatment

In education, treatment always refers to the creation of the learning condition or environment in which the learners are situated. Researchers may introduce different instructional approaches or use different teaching resources to create different learning conditions for learners who have different aptitudes. They can then examine the interaction effect between the learning conditions and the aptitude of the learners in terms of the treatment outcomes, which are usually measured by pre-, post-and delayed posttests.

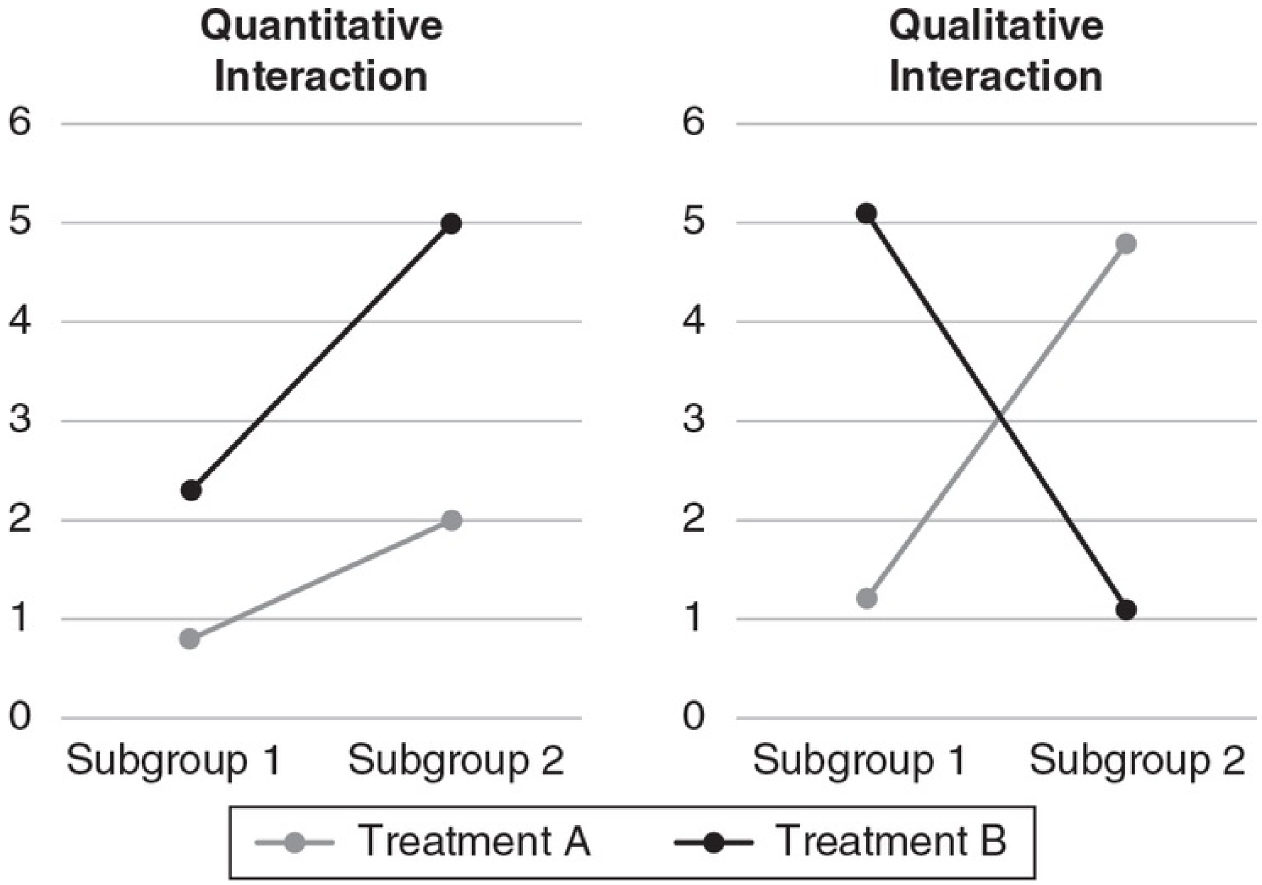
In terms of the types of treatments used, the most frequently manipulated treatment variables in earlier studies were structure and elaboration. The typical manipulation for structure involved providing one group with a more self-directed learning environment and another group with a more teacher-controlled, lecture-based learning environment. Elaboration was typically manipulated by providing one group with analogies or some other means of clarifying or elaborating the learning materials, while the other group typically received only the learning materials without any further elaboration. Other experimental manipulations included providing one group with deductive training and the other with inductive training or comparing a group of learners who engaged in learning primarily through a small group setting to another group who participated in a large group setting.

Recent ATI research has covered a broader range of treatments, including some innovative teaching strategies and curriculum designs. For instance, in one study, one group was provided with static pictures for recognizing rotated spatial structures, while the other group was offered animations. Another study examined the efficacy of a motivation-enhancing treatment, attributional retraining, to support students who were at risk because of a high failure-avoidance orientation (i.e., the tendency to maintain self-worth by avoiding failure).

Interaction

ATI occurs when the degree to which results for two or more treatments, or one treatment over two or more trials, shows a statistically significant difference for learners who also differ on one or more aptitude variables under investigation. Treatment-subgroup interactions may be assessed by either quantitative or qualitative methods. In earlier studies, most researchers made use of quantitative measures in which the direction of the difference between the treatment alternatives in terms of treatment effectiveness was the same for all subgroups, but they differed in the extent of the difference (see Figure 1); that is, the difference was only in magnitude.

Figure 1 An example of quantitative and qualitative treatment-subgroup interactions



Source: Doove, Van Deun, Dusseldorp, and Van Mechelen (2015, p. 3).

Qualitative measures have become increasingly popular in recent years. The difference in treatment effectiveness may be in different directions for different subgroups. This is a statistical difference, as illustrated in the regression slopes in Figure 1. However, the qualitative treatment-subgroup interactions have greater practical value in suggesting the most appropriate treatments for different subgroups of learners for optimal treatment outcomes.

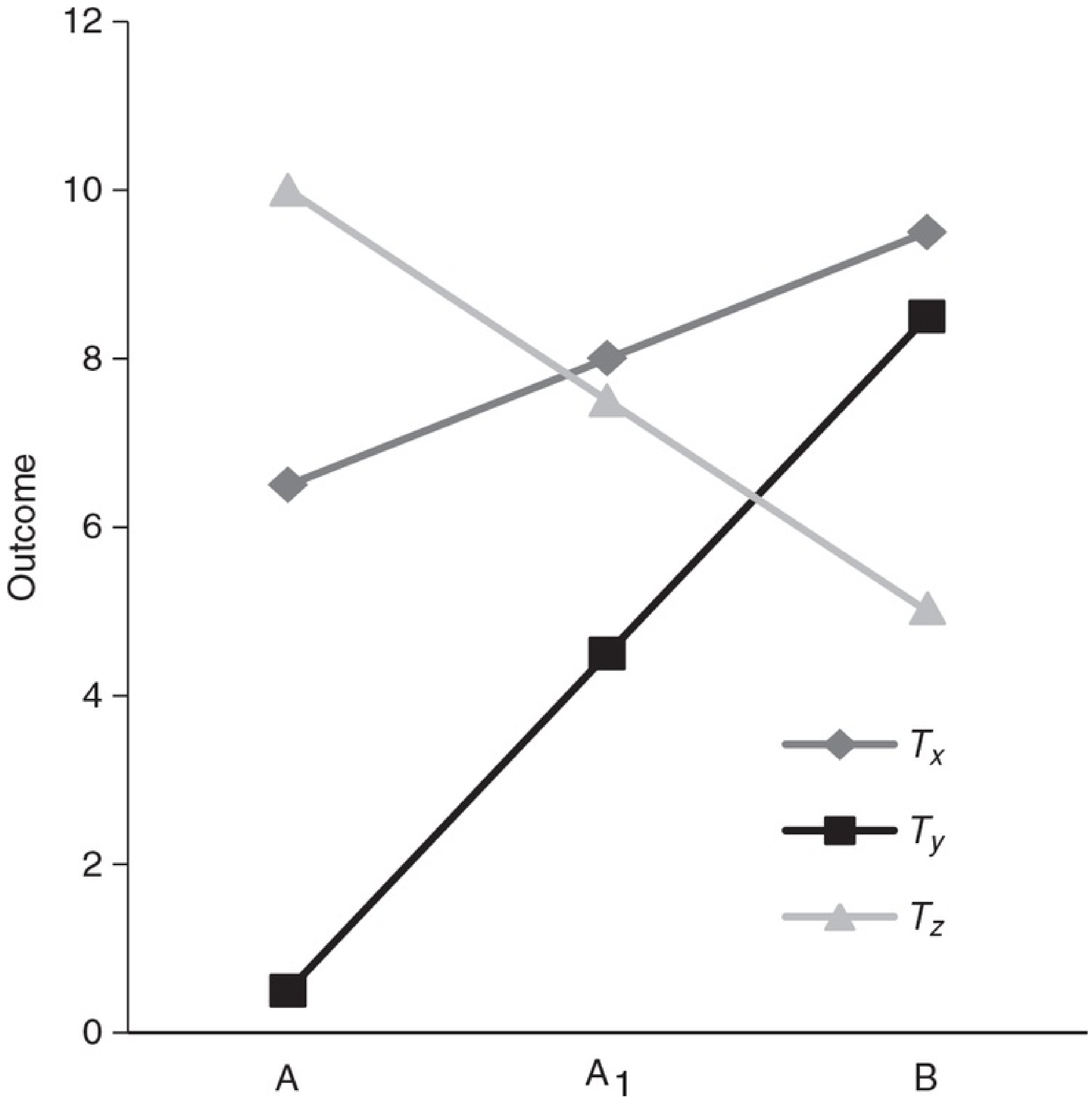
Nevertheless, these early studies found no consistent ATIs in education and training, perhaps due to the small sample sizes in many of the studies. Some studies had further problems due to the large number of variables they measured and analyzed in an attempt to identify every main effect and interaction possible. The diversity of the treatments used, coupled with the relative lack of data points to detect any patterns of ATIs for many of the treatments, made it difficult to draw conclusions about the potential interactions with personal characteristics. With the advent of new methods of assessing ATIs and more robust research designs, it is encouraging that recent studies have found more support for ATIs.

ATI Research Designs

ATI research aims to identify and develop treatments that work best for particular groups of learners with certain aptitudes to optimize the treatment outcomes. For ATI research to be meaningful, it should be driven by plausible hypotheses rather than simply treating the study as a hit-or-miss exploration of statistical associations. The most commonly used research designs for ATI are (a) the standard experimental design, (b) the regression discontinuity design, and (c) the change curve (or growth curve) design.

The *standard experimental design* is the most common and comprises a simple, randomized, controlled experiment, in which two or more groups receive the same treatment and their learning outcomes are assessed with respect to different levels of the aptitude(s) under investigation. Figure 2 illustrates the ATI effect, in which treatment T_z is shown to be more effective than treatment T_y for persons with Aptitude A, but there is no difference between the two treatments for persons with Aptitude B.

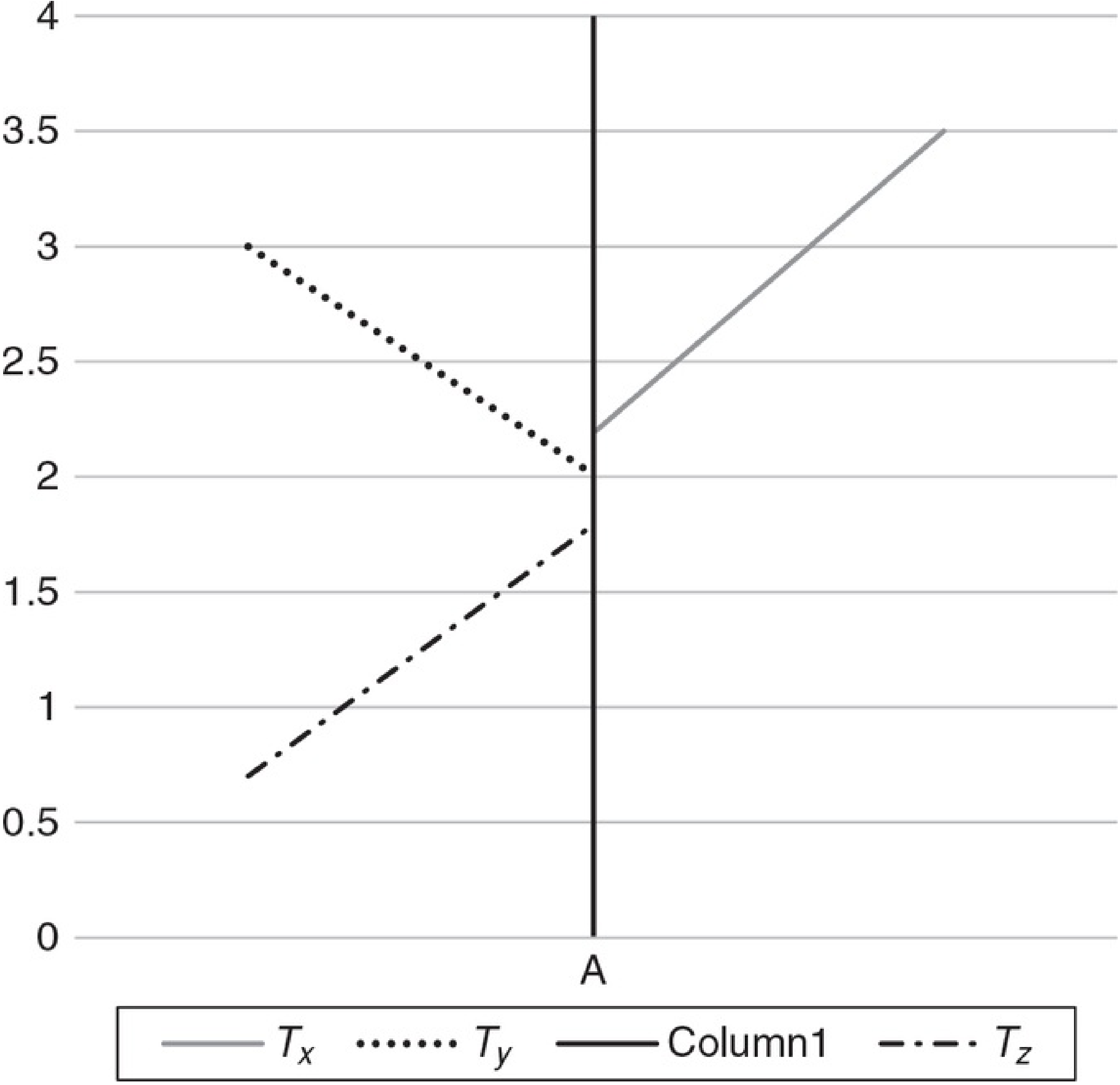
Figure 2 Standard ATI research design



Source: Adapted from Snow (1991, p. 208).

The *regression discontinuity design* can be used when randomization cannot be carried out. In this design, learners are assigned to different treatments on the basis of a cutoff score on the aptitude(s) to be investigated. A treatment effect is observed if there is a discontinuity in the outcome measurement, as illustrated in Figure 3, where treatment T_X is more effective than the other treatments for learners above aptitude level A.

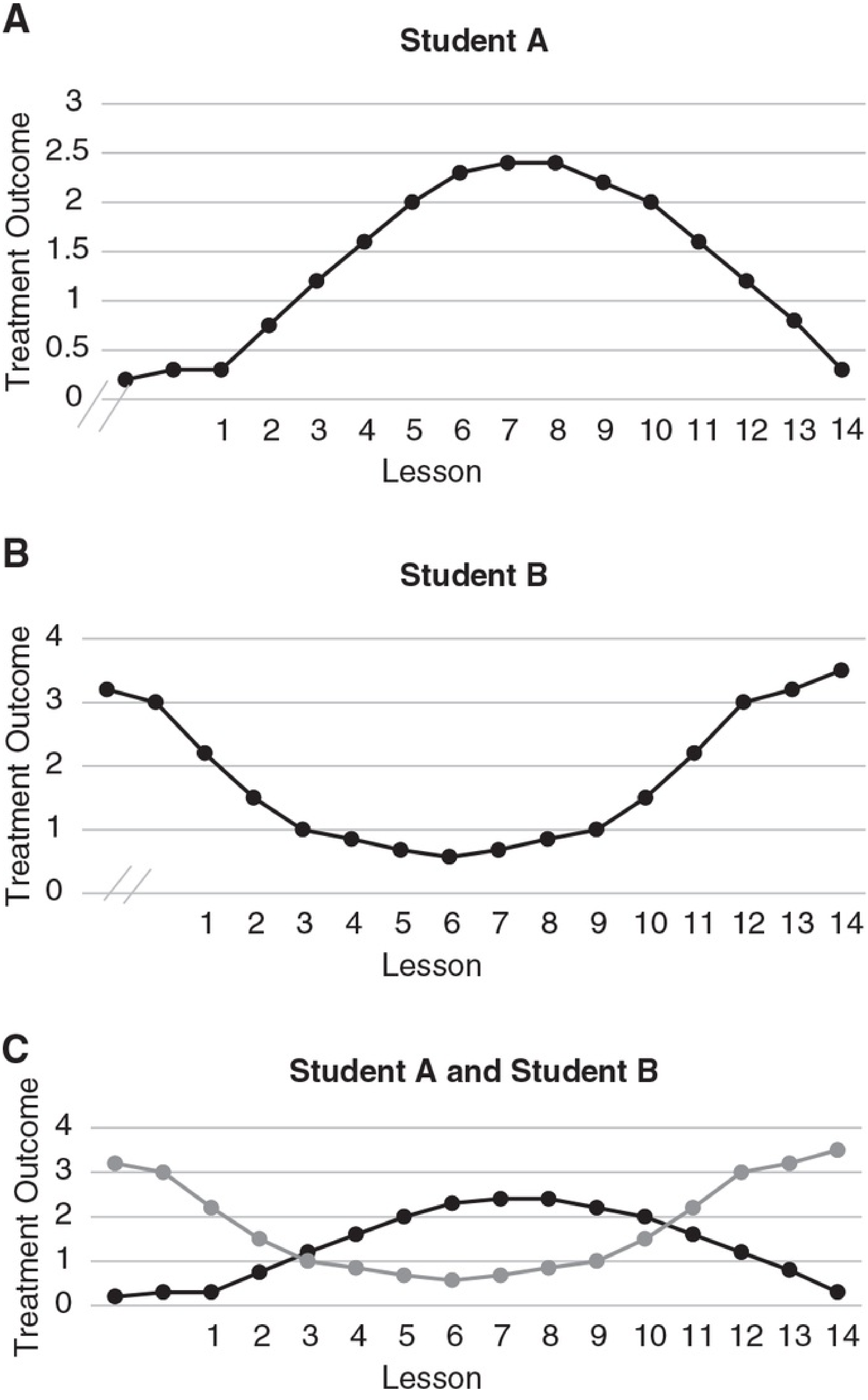
Figure 3 Regression discontinuity ATI research design



Source: Adapted from Snow (1991, p. 208).

In the *change (or growth) curve design*, the change in the outcome variable after learners receive the treatment is observed and analyzed over time, using a growth curve (see Figure 4). This design allows the data of individual learners to be shown and analyzed, and the design does not require a comparative trial control, which may not be feasible or ethical in some educational settings.

Figure 4 Change curve (or growth curve) ATI research design


A
Student A
Treatment Outcome
3
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Lesson
B
Student B
Treatment Outcome
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Lesson
C
Student A and Student B
Treatment Outcome
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Lesson

ATI designs require large sample sizes (i.e., at least 100 subjects per treatment) to ensure adequate statistical power to detect a moderately strong ATI with power of .90. In terms of treatment duration, longer interventions (i.e., at least 10 sessions or more) are recommended to obtain reliable results. Furthermore, to ensure adequate statistical power, it is advisable to have at least two different treatments, as interaction effects need to be shown to occur above and beyond the additive influence of the main effects. In terms of aptitude variables, the aptitude to be investigated should be strongly associated with the outcome for one intervention but not the others. It is important to note that studying one aptitude at a time while disregarding other aptitudes may result in an unwarranted oversimplification. However, studying too many aptitudes or treatment components simultaneously may result in research findings that are infeasible, if not impossible, to interpret.

Implications for Educational Practice

ATI research has motivated a professional movement to introduce and promote differentiated curricula and differentiated instruction to cater for individual differences and learners' diversity in schools. The rationale behind differentiated curricula and instruction is that learners with different attributes or aptitudes benefit from different curriculum contents and/or instructional approaches, which is grounded in the ATI tenet. Teachers are recommended to design, customize, and adapt their curricula and instruction to match the different aptitudes and learning needs of different groups of learners in schools or classes.

Unlike ATI researchers, teachers may not have the time and expertise to design and administer rigorous diagnostic tests to identify and analyze the aptitude(s) of their students. Instead, they tend to ascertain and assess learners' aptitudes and learning needs based on informal classroom observations and analysis of student work samples and test performances. Insights from ATI research findings offer useful inputs and guidance to the teaching profession when devising plans and actions to provide pedagogical support for differentiation to their students.

Furthermore, a number of recent educational practices, such as learning progressions, response to intervention, and data-driven instructional decision making, were influenced by ATI research findings. Learning progressions involve learners progressing along a pathway of increasing proficiency or competency at their own pace. To help them master the target level of proficiency, teachers need to adjust and align their instruction and pace in accordance with the evolving progress of the learners at different points along the pathway.

In response to intervention, early intervention with customized support is given to those learners who are at risk of failing. An initial performance/capacity assessment is conducted to identify students who may encounter learning challenges. These students are then given individualized pedagogical support, while the teacher closely monitors their progress. The tight coupling between the performance/capacity assessment and instruction forms a feedback loop, which follows the ATI principle of matching aptitude with appropriate instructional support in an optimal manner.

Data-driven instructional decision making is premised on the understanding that there is a genuine need to use various kinds of data to inform practice and continuously improve the quality of education. As with ATI research, the data can be from cognitive, conative, and affective domains. Teachers who obtain regular and frequent data related to the aptitudes of the learners are more likely to make well-informed pedagogical decisions and appropriately tailor their instruction to the aptitude profiles of the learners and thus realize the learning potential of every learner.

Future Directions

ATI theory has grown and evolved over the years. It remains the foundation for a broad range of research studies in learning and instruction and for a number of educational policies and practices around the world. The tenet of identifying the interaction between learners’ aptitudes and alternative treatments or instructional interventions, and subsequently creating the learning conditions that match the aptitudes of the learners to achieve optimal learning outcomes, is well supported by different stakeholders in the educational arena.

With recent developments in educational neuroscience research, it is anticipated that new light will be shed on the underlying brain functions of learners with different aptitudes when engaged in the same or different treatments or interventions. This would make significant contributions to the field of ATI, both theoretically and practically.

Ming Fai Pang

See also Interaction; Regression Discontinuity Analysis; Two-Way Analysis of Variance

Further Readings

Caspi, O., & Bell, I. R. (2004). One size does not fit all: Aptitude × Treatment Interaction (ATI) as a conceptual framework for complementary and alternative medicine outcome research. Part II—research designs and their applications. The Journal of Alternative and Complementary Medicine, 10(4), 698–705.

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods. New York, NY: Irvington.

Doove, L. L., Van Deun, K., Dusseldorp, E., & Van Mechelen, I. (2015). QUINT: A tool to detect qualitative treatment–subgroup interactions in randomized controlled trials. Psychotherapy Research, 1–11, 3.

Pang, M. F., & Marton, F. (2013). Interaction between the learners’ initial grasp of the object of learning and the learning resources afforded. Instructional Science, 41(6), 1065–1082.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. Journal of Consulting and Clinical Psychology, 59, 205–216.

April Galyardt April Galyardt Galyardt, April

Areas Under the Normal Curve

Areas under the normal curve

119

121

Areas Under the Normal Curve

The normal distribution, also called the Gaussian distribution, is a probability distribution that arises in many different natural processes. For example, the height of adult organisms in most species follows a normal distribution. The normal distribution is important in statistics because many estimators have sampling distributions that are asymptotically normal; this includes means, medians, proportions, and all maximum likelihood estimators. This means that the null distribution for many hypothesis tests is a normal distribution and that the *p* value for these tests is given by the area under a normal curve.

This entry first discusses the general definition of probability as area, providing both a formal calculus-based definition and a less formal intuitive explanation. Then it discusses the relationship between the area under the normal curve and hypothesis testing.

Probability as Area

The normal distribution is characterized by its unimodal, symmetric “bell shape” and is defined by the density function:

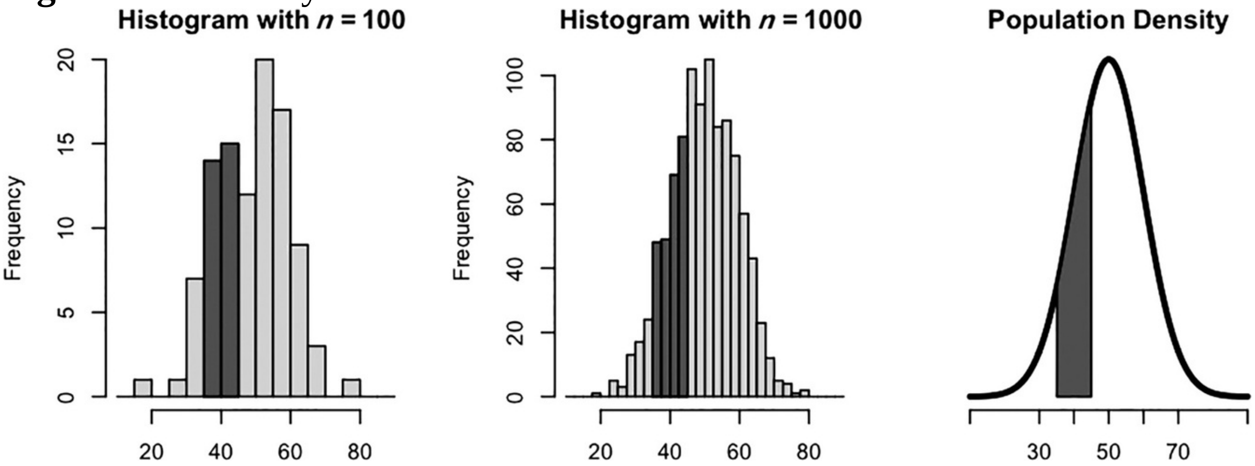
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$$

The probability that a particular event occurs is based on the area under this curve. Formally, the probability that a normally distributed random variable falls within a particular range is given by the integral of the density function over that range.

$$\mathrm{Prn}(a \le X \le b) = \mathrm{ab}1\sigma^2\pi\ e - (x-\mu)22\ \sigma^2 dx.$$

Intuition for this idea, which does not rely on calculus, can be developed through analogy to histograms: Consider an exam where the scores are normally distributed around 50 with a standard deviation of 10. Then if we administered the test to 100 people, we might see 29 scores between 35 and 45. This is shown as the shaded area in the left of Figure 1; the scores within this range represent 29% of the total area.

Figure 1 Probability and area



If we instead administered the test to 1,000 people, we might see 247 scores that fall between 35 and 45, and thus the shaded area in the center of Figure 1 is 24.7% of the total area in that histogram. If we could let the sample size go to infinity, and sample the entire population, then the proportion of scores within that range would be the area under the density curve, shown on the right of Figure 1, which is 24.2% of the total area under the curve.

Cumulative Distribution Functions

This definition of probability as the area under a curve holds in general. Other distributions (e.g., *F* distribution, chi-square distribution, and *t* distribution) have their own density functions, and the probability that an observation from one of these distributions would fall within a particular range corresponds to the area under the respective density curve. In general, the area under a density curve is expressed with a cumulative distribution function:

$$\mathrm{Fa} = \mathrm{Pr}\ (X \leq a) = -\infty \mathrm{af}\ (x)dx.$$

Calculating the Area Under a Normal Curve

The integral for the area under the normal curve has no closed form (meaning there is no simpler way to write the formula), and therefore calculating probabilities without an aid is potentially quite time-consuming. Statistical software packages (such as R, SPSS, and SAS) have built-in functions to calculate probabilities from common distributions, including the normal distribution. Before the use of such software became widespread, it was common to look up the probabilities in a table. In the 2010s, tables of normal probabilities remain common primarily for pedagogical purposes in introductory courses.

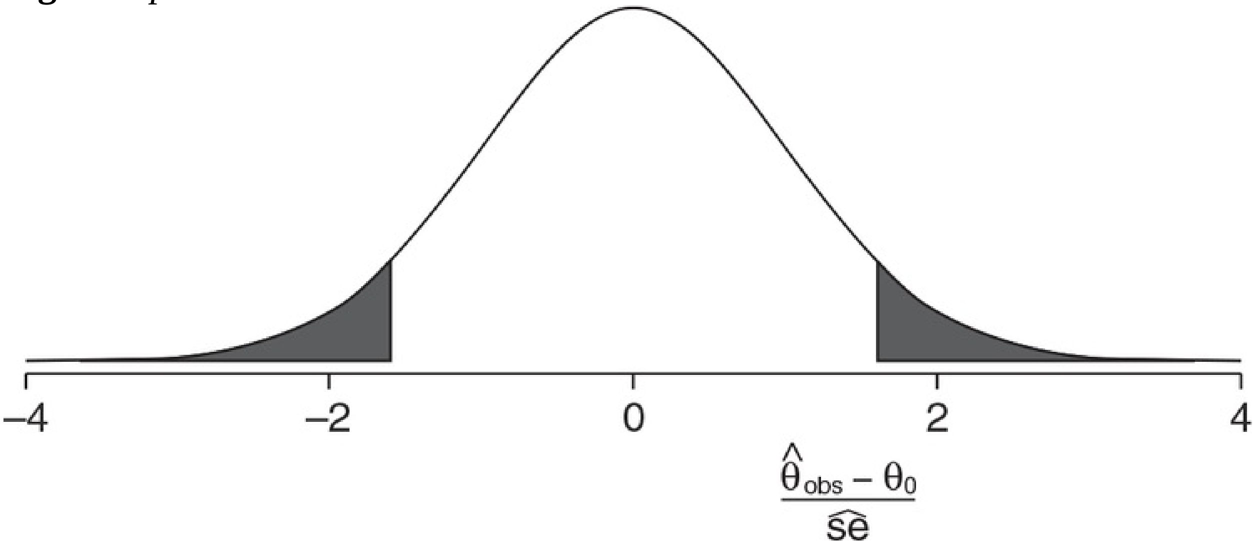
On a normal curve, about 34% of scores will fall between the mean and 1 standard deviation above the mean, with another 34% between the mean and 1 standard deviation below the mean. Between 1 standard deviation and 2 standard deviations on either side of the mean, there is room for about 14% of scores. This leaves roughly 2% of scores farther than 2 standard deviations below and above the mean. These are roughly rounded off estimates (it is important to remember) because the normal curve is infinite and never quite touches the *x*-axis. There are tiny, but greater-than-zero probabilities of scores occurring as one moves farther and farther from the middle.

Relationship to Hypothesis Testing and *p*-Value Calculation

Many estimators for statistical parameters have asymptotically normal sampling distributions. For example, the sample mean *x* is normally distributed around the population mean (μ); and in linear regression, the estimated regression coefficients (β) are normally distributed around the “true” regression coefficients (β). This property is the critical component for many hypothesis tests, since this means that the null distribution for these estimators will be a normal distribution.

It is possible to build a Wald test for any estimator that is asymptotically normal, that is, whenever θ is an estimator of θ and $\theta - \theta_{se} \rightarrow N(0,1)$. In general, a *p* value is the probability that the test statistic would be bigger than the observed value if the null hypothesis was true. In the Wald test for $\theta = \theta_0$, this is the probability $P\theta - \theta_{0se} > \theta_{\text{obs}} - \theta_{0se}$, where θ_{obs} is the observed value of the estimator, as shown in Figure 2.

Figure 2 *p* Values and area



Since the sampling distribution is normal, this probability is given by the area under the normal curve. Since all maximum likelihood estimators have normal sampling distributions, this Wald test (sometimes referred to as a *z* test) arises frequently.

April Galyardt

See also Central Limit Theorem; Distributions; Histograms; Hypothesis Testing; Maximum Likelihood Estimation; Normal Distribution; *p* Value; *t* Tests

Further Readings

Agresti, A., & Franklin, C. A. (2013). Statistics: The art and science of learning from data (3rd ed.). Upper Saddle River, NJ: Pearson.

Wasserman, L. (2004). All of statistics. New York, NY: Springer.

Xiaodi Zhou Xiaodi Zhou Zhou, Xiaodi

Danling Fu Danling Fu Fu, Danling

Asian Americans and Testing Asian americans and testing

121

124

Asian Americans and Testing

This entry examines the performance of Asian American students on tests and looks at the cultural traditions and other factors thought to be behind their generally strong academic performance. Asian Americans are Americans of Asian descent or the first-generation immigrants from Asia. Asian Americans, particularly children of those who immigrated to the United States after World War II, are perceived as high achievers in their scholarly work and are stereotyped as good test takers in U.S. schools.

Asians began coming to the United States in the early 19th century, but most have come since the passage of the Immigration Act of 1965, which eliminated the use of immigration quotas based on national origin. According to 2014 U.S. Census Bureau estimates, the total population of Asian Americans of 20 ethnic groups has reached 19 million, about 5.6% of the entire U.S. population. This percentage is up from less than 1% in 1965. The six largest groups of this population are Chinese (23%), Filipinos (20%), Indian (17%), Korean (10%), Vietnamese (10%), and Japanese (9%). Over 60% of Asian Americans are immigrants.

The most recent wave of Asian immigrants has been different from those who came in the 19th century escaping poverty to work as manual laborers on the transcontinental railroad. Many who have arrived since the 1980s, particularly those from China and India, have been immigrant investors or graduate students gaining entry to institutions for higher education. These Indians and Chinese newcomers are on average well-educated and prosperous, whereas other Asian immigrants, especially those coming as refugees from Southeast Asia with limited English proficiency, often struggle to adapt their lives in the United States.

Academic Performance

On the whole, Asian Americans have relatively high levels of academic performance and educational attainment. As of 2015, 54% of Asian Americans had at least a bachelor’s degree, compared to 33% of the general population of the United States. In addition, despite representing only 5.6% of the total population, Asian Americans receive 25–30% of the National Merit Scholarships and make up more than 30% of the math and physics Olympiad teams and Presidential scholars.

Relatively speaking, Asian Americans perform well on intelligence tests and standardized assessments. This population is seen as academic high achievers and industrious students who, on average, tend to outperform their White peers. Yet relatively high academic performance is by no means the norm for all Asian Americans. Many Asian immigrant students struggle in school and drop out at relatively high rate, particularly those from Southeast Asia.

Cultural Traditions and Social Context

Asian cultures tend to place high value on intellectual prowess as expressed on exams. Confucius, an ancient Chinese philosopher whose influence spread across Asia like that of Socrates and Plato in Europe, stressed that working with one’s intellect was preferable to working with one’s hands. The testing culture in China began during the Tang Dynasty (618–907), when imperial exams began to act as the gatekeeper to choose the best scholars for lucrative positions in the government and bureaucracy, the respected route for social upward mobility. This tradition of testing infiltrated other nearby societies, as the neighboring cultures of Korea, Japan, and Vietnam also began adopting such practices. In essence, ever since the first millennium, Asian cultures have viewed performance on tests as an important criterion of demonstrating competence.

There are three main schools of thought on the reasons for the superior academic performance of Asian American students. The first focuses on their families. There is evidence to suggest Asian families who immigrated to the United States tend to have attained higher levels of education, which may also lead to their marital stability and high incomes. These factors contribute to great family support for the academic achievement of Asian American youths. Considering children's accomplishment as parental success and family honor, Asian parents tend to put much of their resources and effort into their children's education and academic pursuits.

In addition, Asian students tend to have authoritarian parents who demand academic excellence and believe that great effort brings great success. These parents tend to push their children to work long hours, send them to tutoring programs and test–prep classes, establish community heritage schools for their children to study their home languages, and encourage their children to join chess clubs or participate in any activities that they consider would sharpen their children's minds and learning habits. With such familial support and motivation, Asian students come into school with a systemic foundation for scholastic achievement.

The second explanation of Asian academic performance involves intrinsic factors, such as motivation and self-control. Being from a culture that believes in sacrificing the present for the future and forgoing personal interest for family honors, Asian students are disciplined at a young age to fight against any distraction from their concentration on academic work. Studies have found that these internally oriented factors contribute to one's attentiveness to tasks, persistence to finish tasks, and patience for boring drills and memorization exercises. These abilities serve as major assets when taking exams, allowing great focus on the details of the questions asked and stamina throughout the duration of the exam. Such personality traits common among Asian American students contribute to strong performance on tests.

Asian Americans' academic effort may also be attributed to their immigrant status. There are several explanations for this "immigrant paradox." First, immigrants are self-selected for leaving their homeland for better opportunities and future success in their chosen host country. As newcomers devoid of much social or political capital, they may see education as the most efficient means for upward mobility in a foreign land. Even without the guarantees of benefits of societal support, educational attainment may be seen as a path toward brighter futures. When nothing is guaranteed, Asian Americans may see education as an unbiased arbiter rewarding diligence and effort with a higher standard of living.

After a century of marginalization, Asian Americans, often seen as "foreigners" or "outsiders," have come not to expect anything, but to work hard to reach their goals. As indicated earlier, their collective history in the country began on the lower rungs of society, as largely invisible to the rest of society. To offset these barriers, Asians see academics and traversing the gatekeeper of exams as the best route to societal recognition.

Chief among these expectations is postsecondary education, as expectations for college entrance are markedly higher among Asian students as compared to their White peers. Standardized tests are the means to attain that goal, and so, Asian American families may expend more effort and capital in test preparation than other racial and ethnic groups. Asian students, as a result of this cultural and familial push, tend to see tests as more indicative of their worth and essential to their future success than do their White peers. As a result, they may strive harder to excel on these exams and take them more seriously than others. Compounded with the rich testing tradition, these factors compel and propel Asian test takers to outperform their counterparts.

Test-Driven Tradition and Its Drawbacks

Each Asian American ethnic group brings its unique characteristics and traditions to American society. Yet, all these ethnic groups seem to prioritize education and intellectual superiority, which may be seen as an assured guarantee of success. The Asian societies that best conform to the stereotype of superior test performance are Singapore, Hong Kong, Korea, Japan, Taiwan, and China. Three cultures, Singapore, Korea, and China, are highlighted in this section.

First, Singapore’s entire educational system throughout the year is solely geared to the end of the year exams given throughout the country. Teachers mainly teach according to the textbooks, giving worksheets and drills to reinforce the material. Semantic memory of facts and specifics through sheer memorization and rehearsal are prioritized. As such, all learning leads to eventual demonstration of taught knowledge on exams. This acclimates students to perform well on standardized exams.

As another example, South Korea also has had a long history of standardized examinations to assess student achievement. In fact, the nation utilizes one, standardized national exam for consideration for university admittance. This test prioritizes rote memorization and information recall and does not effectively gauge examinees’ more fluid intelligences, such as creativity or analytical thinking. Also, because of these pressures from standardized achievement exams, most South Korean students (84%) attend extracurricular private educational academies to bolster their scores.

Finally, in China, standardized exams are also given much emphasis. Owing to its more direct Confucian roots and the perpetuation of well-established traditions of examination, Chinese students are also educated in a framework geared toward test performance. From elementary school onward, students are given unified competency exams to classify and rank students with classmates and compete with peers in the same province. There are entrance exams for middle and high school, and especially for college. College entrance exams are an entire family affair, as family members devote time and energy to provide the most conducive climate for achievement on these high-stakes exams. During the annual national college exam dates, some Chinese cities even limit traffic to ensure the proper conditions for students to take their exams.

Asian nations have developed as highly competitive test-driven societies. Responsible parents prepare their offspring to compete from the moment of their birth. As such, a large part of education in these Asian nations is on assisting in performance on these high-stakes tests. For example, Chinese parents are apt to spend exorbitant amounts of capital on test preparation services, which for the most part, teach tricks and guessing strategies to outsmart the test companies. In fact, test takers often have little to no idea what the questions are specifically asking but use test-taking and guessing strategies so adeptly as to approximate the correct answers. Test preparation and tutoring services have become one of the most lucrative businesses in China.

There are validity issues with standardized test performance after years of test-focused education. First, these tests, rather than measuring the intended indices, may partly gauge students' test-taking abilities. Strategies such as how to memorize, formulate responses, and guess right answers on tests can improve scores dramatically without marked difference in intellect. Often those teachers who know how to guess the test items correctly are considered the most able teachers, and the test–prep cram schools are most favored when they can provide students with the practice exams that closely approximate testing items.

A second issue confounding validity is that of cheating. Giving the high stakes of test scores, cheating can be common in many Asian countries. For example, there are cases where test takers will memorize the entire standardized exam and print it in review books to assist future test takers. Specifically, in South Korea print it in review books to assist future test takers. Specifically, in South Korea and China, cheating cases associated with standardized tests have caused cancellation and nullification of SAT scores by the Educational Testing Service or the College Board.

For many Asian American parents and students, the beliefs, expectations, and practices about academics common in their home cultures have carried over to their lives in the United States. Although high achievers who are seen as diligent and disciplined, Asian Americans are cast as great workers but face stereotypes that they lack the audacity, ingenuity, and other qualities of leaders. These stereotypes are believed to contribute to Asian Americans' underrepresentation in leadership positions in many areas of society. For instance, a 2015 report said that while Asian Americans make up 27% of the professional workforce in Silicon Valley technology companies, they comprise just 14% of executive positions.

Xiaodi Zhou and Danling Fu

See also Achievement Tests; African Americans and Testing; Cultural Competence; Culturally Responsive Evaluation; Gender and Testing; High-Stakes Tests; Intelligence Tests; Standardized Tests

Further Readings

Ghymn, E. M. (Ed.). (2000). Asian American studies: Identities, images, issues past and present. Peter Long International Academic Publishers.

Hsin, A., & Xie, Y. (2014). Explaining Asian Americans' academic advantage over whites. Proceedings of National Academy of Science, 111, 16–21.

Lee, J., & Zhou, M. (2015). Asian American achievement paradox. Russell Sage Foundation.

Ryan, C. L., & Bauman, K. (2016, March). Educational attainment in the United States: 2015. Current population reports. Washington, DC: U.S. Census Bureau.

Suárez-Orozco, C., Rhodes, J., & Millburn, M. (2009). Unraveling the immigrant paradox: Academic disengagement and engagement among recently arrived immigrant youth. Youth Society, 41, 151–185.

Zhang, Y. (2003). Immigrant generational differences in academic achievement and its growth: The case of Asian American high school students. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Tony Attwood Tony Attwood Attwood, Tony

Asperger’s Syndrome

Asperger’s syndrome

124

127

Asperger’s Syndrome

Asperger’s syndrome is a neurodevelopmental disorder named after the Viennese pediatrician Hans Asperger. This entry first discusses the initial conceptualization of the disorder and more recent changes in how it is conceptualized. It then details the characteristics of the Asperger’s syndrome, the tendency of children with the disorder to be teased and bullied, and the outcomes for those with the disorder as adults.

In the late 1930s, Hans Asperger, a Viennese pediatrician, noticed that some of the children referred to his clinic had a distinct and unusual profile of abilities. Despite having intellectual ability within the normal range, the children had a limited ability to have a reciprocal social interaction with peers and adults, difficulty reading body language, and conspicuous delays in social reasoning, as well as difficulties making and maintaining friendships. Other characteristics were an intense interest in a specific subject, difficulty coping with change, a tendency to impose routines and rituals, and extreme distress in response to specific sensory experiences. Asperger considered that the profile was an expression of autism, and we now conceptualize Asperger’s syndrome as an autism spectrum disorder (ASD).

The term *Asperger’s disorder* was first included in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM)*, published in 1994. There has subsequently been a great deal of research defining the characteristics and evaluating intervention strategies for children at school and in the home and for adults at college and in the workplace. In May 2013, the American Psychiatric Association published the fifth edition of the *DSM* (*DSM*-

5) and replaced the term *Asperger's disorder* with the diagnostic term *autism spectrum disorder level 1, without accompanying intellectual or language impairment*, a lengthier and more cumbersome term. The rationale is that ASD can be conceptualized as a dimensional rather than a categorical concept and that a single umbrella term of ASD, with specific information on the level of expression, is more accurate and consistent with the research literature and clinical experience. However, the term *Asperger's syndrome* is still used by clinicians, parents, teachers, therapists, and those with an ASD. The general public and media also continue to use the term. For simplicity and continuity, this entry uses the term *Asperger's syndrome*.

Key Characteristics

The *DSM-5* diagnostic criteria refer to persistent deficits in social communication and social interaction across multiple contexts, with deficits in social–emotional reciprocity, nonverbal communication, and the development, maintenance, and understanding of relationships. The underlying assumption is that someone who has Asperger's syndrome has difficulty "reading" social situations. The deficits in social–emotional reciprocity can be expressed by a tendency to be withdrawn, shy, and introspective in social situations, avoiding or minimizing participation or conversations; or, conversely, actively seeking social engagement and being conspicuously intrusive and intense, dominating the interaction and being unaware of social conventions such as acknowledging personal space. In each example, there is an imbalance in social reciprocity.

There is a third strategy for coping with difficulties with "reading" social situations and that is to avidly observe and intellectually analyze social behavior, subsequently achieving reciprocal social interaction by imitation and by using an observed and practiced social "script" based on intellectual analysis rather than intuition. This is a compensatory mechanism often used by girls who have Asperger's syndrome, who are thus able to express superficial social abilities. In addition, adults who have Asperger's syndrome can gradually learn to read social cues and conventions, such that any deficits in social–emotional reciprocity may not be conspicuous during a brief social interaction.

One of the characteristics of Asperger's syndrome is a difficulty reading someone's body language, facial expressions, gestures, and voice to indicate their thoughts and feelings and then incorporating that information in the conversation or interaction. We conceptualize this difficulty as an expression of impaired theory of mind.

Asperger’s syndrome is also associated with a signature language profile. This can include impaired pragmatic language abilities (i.e., the “art” of conversation) with a tendency to engage in monologues and a failure to follow conversational rules. There may also be literal interpretations, with a tendency for the person to become confused by idioms, figures of speech, and sarcasm. There may also be unusual prosody, for example, a child may consistently use an accent based on the voice of a television character or an adult may speak with an unusual tone, pitch, or rhythm. All these characteristics affect the reciprocity and quality of conversation.

Another diagnostic characteristic is restricted, repetitive patterns of behavior, interests, or activities. This can include insistence on sameness, inflexible routines, and the acquisition of information on a specific topic. Parents and teachers are often concerned that routines and rituals are imposed in daily life, with the child showing great agitation if prevented from completing a particular routine or ritual at home or in class. There is a determination to maintain consistency in daily events and high levels of anxiety if routines are changed.

The special interests, which can occur throughout both childhood and the adult years, often involve the acquisition of information and knowledge on a specific topic and are unusual in terms of intensity or focus. Each interest has a “use by date,” ranging from hours to decades, and research has indicated that the interest has many functions, such as being a “‘thought blocker” for anxiety, an energy restorative after the exhaustion of socializing, or an extremely enjoyable activity. Special interests can also create a sense of identity and achievement, as well as provide an opportunity for making like-minded friends who share the same interests. The sense of well-being associated with the interest can become almost addictive, such that the interest begins to dominate the person’s time at home; this may lead to genuine concern that it is preventing engagement in any other activities.

The *DSM-5* includes reference to sensory sensitivity as one of the hallmark characteristics of ASD. This has been a characteristic of Asperger’s syndrome that has been clearly and consistently described in autobiographies and recognized by parents and teachers. Sensory sensitivity can be a lifelong problem, with sensitivity to distinct sensory experiences that are not perceived as particularly aversive by peers. These can include specific sounds, especially “sharp” noises such as a dog barking or someone shouting; tactile sensitivity on a specific part of the body; and aversive reaction to specific aromas, light intensity, and other sensory experiences. In contrast, there can be a lack of sensitivity to some sensory experiences, such as pain and low or high temperatures. The child or adult can feel overwhelmed by the complex sensory experiences in particular situations, such as shopping malls, supermarkets, birthday parties, or the school playground. Sometimes, social withdrawal is not due simply to social confusion but to the need to avoid sensory experiences that are perceived as unbearably intense or overwhelming.

Additional Characteristics

Mood Disorders

While the child may have considerable intellectual ability and academic achievement, there is invariably confusion and immaturity with regard to understanding and expressing feelings and a vulnerability to developing signs of an anxiety disorder or depression. There may also be a need for guidance in the management and expression of anger and affection. The theoretical models of autism developed within cognitive psychology and research in neuropsychology and neuroimaging provide some explanation as to why children and adults with Asperger’s syndrome are prone to secondary mood disorders.

The term *alexithymia* is used to describe a characteristic associated with Asperger’s syndrome, namely an impaired ability to identify and describe feeling states. Children and adults with Asperger’s syndrome often have a limited vocabulary of words to describe feeling states, especially the subtler or complex emotions, and will need education in perceiving and expressing emotions. Over the last decade, there has been the development and evaluation of a new range of cognitive behavior therapy programs for parents and teachers to help those who have Asperger’s syndrome understand and express emotions at home and at school.

Cognitive Abilities

Children with Asperger’s syndrome often have an unusual profile of cognitive abilities. Some young children may start school with academic abilities above their grade level, such as advanced literacy and numeracy that may have been their grade level, such as advanced literacy and numeracy that may have been self-taught through watching educational television programs, using educational computer programs, or avidly looking at books and reading about a special interest. Some young children with Asperger’s syndrome appear to easily “crack the code” of reading, spelling, or numeracy; indeed, these subjects may become their special interest and a subsequent talent. In contrast, some children with Asperger’s syndrome have considerable delay in these academic skills, and an assessment of their cognitive abilities suggests specific learning disorders, especially dyslexia. There are more children with Asperger’s syndrome than one might expect at the extremes of cognitive and academic ability.

At school, teachers often recognize that the child has a distinctive learning style, often being talented in their understanding of the logical and physical world, as well as noticing details and remembering and arranging facts in a systematic fashion. However, the child can be easily distracted, especially by noises or social activity in the classroom, and when problem solving, appears to have a “one-track mind” and a fear of failure. As the child progresses through the school years, teachers may identify problems with organizational abilities, especially with regard to homework assignments and essays. They also note that the child appears to not follow advice, look to peers for guidance, or learn from mistakes. End-of-year school reports often describe a conspicuously uneven profile of academic achievement, with areas of excellence and areas that require remedial assistance.

The research on IQ profiles indicates that children with Asperger’s syndrome tend to have good factual and lexical knowledge. Their highest scores are often on the subtests that measure vocabulary, general knowledge, and verbal problem solving. In the visual reasoning subtests, children with Asperger’s syndrome can achieve relatively high scores on the Block Design and Matrices tests. However, the profile can also include slower processing speed and impaired auditory working memory.

Movement and Coordination

As much as children with Asperger’s syndrome have a different way of thinking, they can also have a different way of moving. When walking or running, the child’s coordination can be immature, sometimes with an idiosyncratic gait that lacks fluency and efficiency. On careful observation, there can be a lack of synchrony in the movement of the arms and legs, especially when the person is running. Parents often report that the young child needs considerable guidance in running. Parents often report that the young child needs considerable guidance in learning activities that require manual dexterity, such as tying shoe laces, dressing, and using eating utensils. The movement and coordination problems can be obvious to the physical education teacher and other children during PE classes and sports and in playground games that require ball skills. The child with Asperger's syndrome can be immature in the development of the ability to catch, throw, and kick a ball. One of the consequences of not being successful or popular at ball games is the exclusion of the child from some of the social games in the playground. Such children may choose to actively avoid these activities, knowing they are not as able as their peers.

Teachers and parents can become quite concerned about difficulties with handwriting. The individual letters can be poorly formed and the child may take too long to complete each letter, causing delay in completing written tasks. While other children in the class may have written several sentences, children with Asperger's syndrome are still deliberating over the first sentence, trying to write legibly, and becoming increasingly frustrated or embarrassed about their inability to write neatly and consistently.

Teasing and Bullying

Children who have Asperger's syndrome are frequently targets of teasing, bullying, rejection, and humiliation. This can have a devastating effect on self-esteem and is a major cause of depression in adolescence and a contributory factor for school refusal and school suspension for retaliation. Schools are becoming more aware of this problem and introducing programs to prevent teasing, bullying, and rejection specifically designed for children and adolescents who have Asperger's syndrome.

The Adult Years

The children who were diagnosed with Asperger's syndrome in the last 20 years are now becoming adults. Also, diagnosticians are increasingly receiving referrals for diagnostic assessments of mature adults who are the relatives of young children with Asperger's syndrome. Practitioners and researchers therefore are now exploring the challenges faced by adults in terms of tertiary education, employment, and relationships. There is a trend for the signs of Asperger's syndrome to become increasingly less conspicuous with maturity and support and the potential for the achievement of a successful career and a long-support and the potential for the achievement of a successful career and a long-term relationship.

Tony Attwood

See also Anxiety; Autism Spectrum Disorder; *Diagnostic and Statistical Manual of Mental Disorders*

Further Readings

Attwood, T. (2007). The complete guide to Asperger’s syndrome. London, UK: Jessica Kingsley Publishers.

Sarah Parsons Sarah Parsons Parsons, Sarah

Assent

Assent

127

129

Assent

Assent is an agreement to take part in research activities that may be given orally, in writing, or in the preferred communication medium of the participant. Assent to participate in research may be given following the provision of information about the project, or specific activity, but without the individual necessarily receiving a full disclosure about potential benefits, risks, and the procedures or activities of research participation, as would be the case for *informed consent*. In educational research, assent is most often discussed in relation to the involvement in research of children and young people (under 18 years old). A legal parent or guardian would typically be expected to provide fully informed consent before the child is approached for their assent. This entry describes the main principles and practices of assent and the different interpretations of the term in educational research.

Core Principles and Practices of Assent

Although there are differences of interpretation, there are some common core principles and practices of assent. Primarily, assent is understood to be an *agreement* to take part in a specific activity within a research project, such as an interview, group discussion, creative activities, being observed (e.g., in a classroom), or completing a questionnaire or test. The agreement to participate should be *voluntary and explicit*; nonrefusal or passive involvement are not typically accepted as indicators that a child has provided assent to take part in an activity. Participation should, therefore, be an *active choice*, which means that a child has to decide whether to participate or not. Choosing not to participate is usually called *dissent*, and children should be given clear opportunities to assent or dissent to their own research participation.

Researchers also need to respect the rights of children to be given the opportunity to assent or dissent and to respect their decision once made. To enable children to make a clear choice about research participation, they would usually be provided with information about specific project activities in an accessible way. This could be through the use of simplified text, pictures, photographs, or videos that may be accompanied by verbal explanations. This simplification of information is one of the features that distinguishes assent from informed consent.

Historical and Conceptual Development

The concept of assent emerged from developments in understanding children's status and agency as competent individuals, capable of making decisions and contributing their views. This *sociology of childhood* understands children as having unique and valuable perspectives on the world, separate and distinct from adults. This perspective has been both strengthened by and reflected in a *rights-based approach* to children's participation, stimulated primarily by the United Nations Convention on the Rights of the Child (UNCRC) in 1989. The convention is aimed at supporting children's human rights internationally and has been ratified by 195 members of the United Nations, with the United States the only member state that had not ratified it as of 2016.

Among other things, the UNCRC stipulated the right of children to be heard in all matters affecting them, with due weight given to those views according to the age and maturity of the child (Article 12). According to the United Nations Children's Fund, this means that "when adults are making decisions that affect children, children have the right to say what they think should happen and have their opinions taken into account" (United Nations Children's Fund, n.p.). Article 12 of the UNCRC recognizes that children's ability to make decisions develops with age and so the views of teenagers, for example, would be given more weight than those of a very young child. Article 13 of the UNCRC accords children the right of freedom of expression, which means that (within the law) they can receive and share information in any way they choose, including talking, drawing, and writing. The principles of Articles 12 and 13 are relevant to assent because they recognize that children, depending on their age and maturity, may not be able to provide fully informed consent to take part in research. Children can, however, provide their assent or dissent based on specific information relating to specific activities, and their choice may be communicated in different ways.

Interpretations of Assent in Educational Research

There is not a consensus about what assent means in practice or in principle in educational research. Some researchers do not think that assent is a valid concept, partly because it implies that children are not competent or capable of giving their informed consent and that adults are always needed to provide an informed view. This stance critiques assent because it undermines the agency of children to make their own decisions.

There is also recognition that adults exert power over children’s decision making in ways that make it difficult to dissent; for example, if a parent or teacher has already provided informed consent, then the child may not feel able to opt out. Assent is problematic when young people are involved in research about sensitive topics (e.g., teenage pregnancy, sexuality, illegal activities) where it could be detrimental (to the young person) to seek informed consent from parents or carers. By contrast, some researchers take a more pragmatic view, arguing that a child needs to understand and feel comfortable about what they are being asked to do, and a process of assent can enable this understanding. Within this context, there is an onus on researchers to be knowledgeable about, and sensitive to, the needs of the participants they want to include in their research. This means taking care to tailor the presentation and content of information in ways that will be accessible and meaningful for children and young people and revisiting assent throughout a project.

Sarah Parsons

See also Informed Consent; Institutional Review Boards; Qualitative Research Methods

Further Readings

Cocks, A. J. (2006). The ethical maze: Finding an inclusive path towards gaining children’s agreement to research participation. Childhood, 13(2), 247–266.

Dockett, S., & Perry, B. (2011). Researching with young children: Seeking assent. Child Indicators Research, 4(2), 231–247.

Hammersley, M. (2015). Research ethics and the concept of children’s rights. Children … Society, 29(6), 569–582.

Hurley, J. C., & Underwood, M. K. (2002). Children’s understanding of their research rights before and after debriefing: Informed assent, confidentiality, and stopping participation. Child Development, 73(1), 132–143.

Parsons, S., Sherwood, G., & Abbott, C. (2016). Informed consent with children and young people in social research: Is there scope for innovation? Children … Society, 30(2), 132–145.

Renold, E., Holland, S., Ross, N. J., & Hillman, A. (2008). Becoming participant—Problematizing informed consent in participatory research with young people in care. Qualitative Social Work, 7(4), 427–447.

United Nations Children’s Fund (n.d.). Fact sheet: A summary of the rights under the Convention on the Rights of the Child. Retrieved from http://www.unicef.org/crc/files/Rights_overview.pdf

Trena M. Paulus Trena M. Paulus Paulus, Trena M.

ATLAS.ti

ATLAS.ti

129

132

ATLAS.ti

ATLAS.ti stands for *Archiv für Technik, Lebenswelt und Alltagssprache* (Archive for Technology, Lifeworld and Everyday Language.text interpretation) and is one of the several computer-assisted qualitative data analysis software (CAQDAS or more simply QDAS) packages that can be used to manage every phase of a qualitative research study. Other QDAS packages include QSR NVivo, MAXQDA, Dedoose, HyperResearch, QDA Miner, Quirkos, and Transana.

ATLAS.ti was developed from 1989 to 1992 as an interdisciplinary research project by scholars in psychology, educational science, and computer science at the Technical University of Berlin. In 1993, Scientific Software Development GmbH released the first commercial version. In 2013, Free iPad and Android apps were released, and in late 2014 a Mac-native version was released. Version 8 is scheduled for release in late 2016. Two ATLAS.ti user conferences were held in Berlin in 2013 and 2015, and the conference proceedings are available online.

QDAS packages can assist with the management and analysis of a wide variety of qualitative data useful for educational research, such as interviews, focus groups, recordings of classroom interactions and observational field notes, web pages, documents and records, social media conversations, images, videos, Google Earth maps, and responses to open-ended survey questions. This entry describes ways in which ATLAS.ti can be used to carry out a variety of analytic strategies, provides examples of how various components of the software can be used to do so, and recommends best practices for reporting the use of ATLAS.ti in research reports.

in research reports.

Analytic Activities

QDAS packages should not be confused with data analysis software such as SPSS, STATA, or SAS, which automatically analyze the data according to statistical formulas. Rather, ATLAS.ti is a platform, or workbench, in which researchers can choose how to organize, store, and structure their unstructured or semi-structured data in a systematic way that is aligned with their methodological approach. As described by the software manual, visualization, integration, serendipity, and exploration principles underlie its design. *Visualization* tools help researchers elicit meaning from the data; all project materials can be *integrated* within the software; browsing the data with the software encourages *serendipitous* findings; and the software supports an *exploratory* yet systematic approach to analysis.

Any qualitative methodological approach can be enacted in the software, be it thematic analysis, grounded theory analysis, discourse analysis, or ethnographic approaches to name a few. Christina Silver and Ann Lewins have suggested that there are five main categories of analytic activities that can be supported by ATLAS.ti and other QDAS packages: *integrating* data sources and analytic approaches; *exploring* the content and structure of the data; *organizing* materials and ideas; *reflecting* on data, interpretations, processes, and results; and *retrieving*, reviewing, and rethinking ideas about the data.

ATLAS.ti can be used to create what Zdeneˇk Konopásek called a textual laboratory to organize, store, and manage data sources alongside other project documents such as data collection instruments, ethics board approval forms, and even the research literature. Literature reviews require, in essence, a type of qualitative data analysis, and PDFs of articles can be uploaded into ATLAS.ti and the components used to analyze them in a systematic and visible way. With version 8, bibliographic data from reference management software such as Endnote, Zotero, and Mendeley can be imported into a project and triangulated with other data sources. Organizing the data in a project file makes the data portable, and annotating the data within the ATLAS.ti project file creates a visible audit trail. Both of these features support smooth collaboration across team members.

By taking a laptop or iPad into the field, researchers can easily type up field notes, take photos, record video and audio, and import relevant PDF documents notes, take photos, record video and audio, and import relevant PDF documents into a project. The iPad data can then be uploaded to cloud-storage programs such as Dropbox and imported into the desktop version of the program for analysis. The iPad app does allow for direct coding or memoing of the files which can be useful for immediate analysis and note taking that can be more fully developed upon return from the field.

ATLAS.ti supports the transcription of audio and video recordings as well as the association between the transcripts and the audio and video sources. Transcriptions can be done in ATLAS.ti with shortcut keys or by connecting a foot pedal to facilitate typing, and the resulting transcripts can be synchronized with the recordings. In this way, when the researcher clicks into the transcript, that part of the recording will be played, thus keeping the analyst closer to the source of the data. Audio and video files can also be coded directly without transcribing.

ATLAS.ti has some automated analysis tools such as text search tools, a word frequency count tool, and an auto-coding feature which allows the researcher to quickly find and label key words of interest. The various coding features allow the researcher to create and link analytically meaningful labels to various segments of the data, after which all labeled, and thus related, portions of the data can be retrieved at once. In this way, the researchers can review all related sections of the data together as they create the interpretations to answer the research questions. Data and initial interpretations can be graphically displayed for further exploration through visualization. All analytic work can be exported into text files or spreadsheets for further work outside of the software. ATLAS.ti provides writing tools in which reflective memos and notes, interpretations of the findings, team meeting notes, and other important decisions about the study can be documented.

Components of ATLAS.ti

Effective use of ATLAS.ti requires selecting and using software components in a way that will enact the desired analytic strategy. Nicholas Woolf and Christina Silver call this process Five-Level QDA, in which individual analytic tasks are matched to the underlying components of the software. They have organized ATLAS.ti’s components into five major groups: components that support *providing* data, *segmenting* data, *conceptualizing* the segments, *writing*, and *interrogating* data and its analysis.

Providing data involves creating “primary documents” within the project file. These can be existing data sources (e.g., digital images or interview transcripts created outside of the software) or can be created internally to the software (e.g., by typing observational field notes directly into the software). Primary documents that are related in some way can be grouped and later interrogated for how the results of the analysis are distributed across any cases, participant groups, or other demographic characteristics of interest.

Segmenting data entails creating analytically meaningful units within the primary documents. These units are called “quotations” and can exist on their own (free quotations) or can be linked to other components—such as memos, codes, or other quotations. Quotations are the building blocks of the analysis, with all reports of the analysis organized by numbers of quotations, for instance, by numbers of quotations assigned to a certain code, numbers of coded quotations per primary document, or the co-occurrence of coded quotations across the data.

Conceptualizing segments refers to the process of creating analytic meaning from the data. “Codes” can be created and attached to quotations, codes that are related in some way can be organized into groups, and codes can be linked to other codes. ATLAS.ti has an “in vivo” coding feature where exact words of the participants become the code name. Codes can be organized by color or into hierarchies using prefixes or other naming conventions.

Writing is a fundamental practice of qualitative research. The “comment” component provides a space in which to capture important information about the primary document data sources, the meaning of codes, and reflections on individual quotations. A robust “memo” tool provides a flexible way to, for example, document the analytic approach being used, write up analytic insights and interpretations, capture team meeting notes, or pose questions that arise during the analysis. Memos that are related to each other in some way can be organized into groups.

The interrogating components of ATLAS.ti allow the researcher to ask questions of the data after quotations, memos, and codes have been created and/or linked. These include the ability to retrieve quotations that have been assigned to a certain code so that they can be viewed together, network views in which displays of linked components provide a visual representation of the analysis, and the co-occurrence explorer, which can retrieve quotations that have been coded with more than one code in order to display possible relationships coded with more than one code in order to display possible relationships between the codes.

Together, the components of ATLAS.ti provide a robust toolkit with which researchers can impose structure on the data in a way that is aligned with the methodological design of the study.

Best Practices in Reporting the Use of ATLAS.ti

Megan Woods, Trena Paulus, David Atkins, and Rob Macklin conducted a literature review of all peer-reviewed journal articles published from 1994 to 2013 that reported use of ATLAS.ti and QSR NVivo in order to investigate the prevalence of software use in qualitative research. Although the use of QDAS was found to be on the rise, most researchers are using it only for the data analysis phase of their studies and for traditional qualitative data sources (e.g., interviews, focus groups, documents, field notes, and open-ended survey questions).

Few researchers included details about how they used the software other than mentioning that they did so. Given the flexibility of the software, such lack of detail may perpetuate persistent misconceptions—that it can automatically analyze the data, for example, or that using QDAS inherently improves the study’s rigor. Instead, researchers should report the following information when using QDAS in their studies: (a) Given that the components of QDAS change with each new version, identify the version that was used. (b) So as not to give the impression that the software, rather than the researcher, is doing the analysis, use active voice (the research team created quotations and assigned codes to the data) rather than passive voice (ATLAS.ti was used to analyze the data). (c) Provide a brief description of what the software is, what it is used for, why it was selected, and which components were used and how. If possible, include software outputs (e.g., code lists and definitions or network view graphical representations) as part of the data display and findings in order to retain the connection between the use of the software and the final researcher interpretations.

Trena M. Paulus

Further Readings

Friese, S. (2014). Qualitative data analysis with ATLAS.ti 7 (2nd ed.). London, UK: Sage.

Gilbert, L. S., Jackson, K., & diGregorio, S. (2014). Tools for analyzing qualitative data: The history and relevance of qualitative data analysis software. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), Handbook of research on educational communications and technology (4th ed., pp. 221–236). New York, NY: Springer.

Konopasek, Z. (2008). Making thinking visible with ATLAS.ti: Computer-assisted qualitative analysis as textual practices. FORUM: Qualitative Social Research, 9(2), Art. 12.

Paulus, T., & Bennett, A. (in press). Integrating ATLAS.ti into qualitative research methods courses: Beyond data analysis. International Journal of Research and Method in Education. Retrieved from http://dx.doi.org/10.1080/1743727X.2015.1056137

Paulus, T., & Lester, J. (2016). ATLAS.ti for conversation and discourse analysis. International Journal of Social Research Methodology 19(4), 405–428.

Paulus, T., Lester, J., & Dempster, P. (2014). Digital tools for qualitative research. London, UK: Sage.

Paulus, T., Woods, M., Atkins, D., & Macklin, R. (in press). The discourse of QDAS: Reporting practices of ATLAS.ti and NVivo users with implications for best practices. International Journal of Social Research Methodology. Retrieved from http://dx.doi.org/10.1080/13645579.2015.1102454

Pope, L. (2016). On conducting a literature review with ATLAS.ti. ATLAS.ti Research Blog. Retrieved from http://atlasti.com/2016/09/01/litreview/

Silver, C., & Lewins, A. (2014). Using software in qualitative research: A step by step guide (2nd ed.). London, UK: Sage.

Woolf, N., & Silver, C. (in press). Qualitative analysis using ATLAS.ti: The five-level QDA method. Routledge.

Meng-Jung Tsai Meng-Jung Tsai Tsai, Meng-Jung

Attention

Attention

132

134

Attention

The term *attention* refers to the way in which humans allocate limited cognitive resources to information processing. *Arousal, effort, mental effort, concentration, mental involvement,* and *engagement* are the terms that are usually used for defining attention. Selective attention, sustained attention, and divided attention are the issues of greatest concern in educational settings.

Due to the different definitions and issues emphasized in different disciplines such as cognitive psychology, clinical psychology, and neuropsychology, the assessment of attention involves different approaches with different instruments. The assessment of attention can reveal the individual differences in learning concentration and control strategies; therefore, it is important for studies in the science of learning, educational counseling, and individualized learning. This entry describes the definitions of attention from different perspectives and reviews the primary assessment approaches based on these perspectives. The entry concludes with an overview of the advantages and disadvantages of these evaluation approaches as well as a list of resources on the measurement of attention.

Issues of Attention

Multidimensional perspectives of attention have been addressed in various psychological disciplines since the 1950s when studies on cognitive process began to be increasingly emphasized. The foundation of attention in cognitive psychology is the capacity model, which argues that the total amount of cognitive resources for attention is limited. In the information processing model, attention is a process of information selection and filtering between the humans' sensory registers and working memory.

Selective attention can explain how students catch the main ideas in school lectures. *Effort* or *mental effort* is a term that is often used to indicate how much attention an individual puts into a task. *Mental involvement* and *mental engagement* are terms that are sometimes used to reveal different degrees of attention paid to processing specific learning materials. From the neuroscientific perspective, attention has been regarded as an *arousal* that is spontaneously activated by environmental stimulations or intentionally controlled for achieving specific goals. The manifestations of arousal include eye blink, pupil dilation, skin conduction, and brain wave.

Sustained attention indicates how individuals keep focused on a task. It is an important indicator for discriminating attention-deficit/hyperactivity disorder in clinics and special education. *Divided attention* refers to the divided allocations of attentional resources when an individual performs multiple tasks simultaneously. It is associated with the control and management of limited resources and plays an important role in the performance of multitasking. These attentional models serve as the theoretical foundations of the development of the multimedia learning theory and the cognitive load theory, the two primary guides for the contemporary design of instructional technology and digital learning.

Measurement of Attention

Generally, three primary approaches have been used for the assessment of attention: *reported scales, performance-based tests*, and *physiological measures*. Reported scales are the questionnaires or checklists to be checked by learners themselves or by others, such as teachers or parents. The attention assessed by a self-reported scale is often referred to as *perceived attention*. Sometimes, interviews are employed as complements of reported questionnaires. Performance-based tests are the most commonly used approach in lab-based experiments and clinical practice. For example, the continued performance test for sustained attention is commonly used for diagnosing attention-deficit/hyperactivity disorder in young children. Dual-task performance is an approach usually used to assess the attention one pays to the primary task by measuring the response time or error rate of the secondary task, which in turn reveals the cognitive load of the primary task indirectly. Reaction time or error rates are the measures often reported by performance-based tests.

Physiological measures are rooted in the significant correlations between the attentional features and physiological measures tracked by specific types of equipment. For example, *eye-tracking* systems can detect and track an individual’s visual focus during a task and output various measures such as fixation-based, saccade-based, pupil-based, and eye-blinking measures. Analyses of these measures are usually used to reveal an individual’s visual attention distributions and transfers on learning materials, which may imply the individual’s mental workloads or learning motivation. An *electroencephalogram* can reveal an individual’s wake–sleep state by detecting the brain waves of the α and θ signals. It is the most reliable tool to measure sustained attention. *Functional magnetic resonance imaging*, on the other hand, can reveal the brain areas activated by specific cognitive functions.

Advantages and Disadvantages

Different advantages and disadvantages are associated with different evaluation methods for attention. Questionnaires are the most convenient instrument to use for conducting a large-scale survey. Well-developed scales or checklists are easy to use for preliminary diagnoses for attentional problems in educational settings or clinical institutes. Self-reported questionnaires are more reliable to use for adults than for young children.

Self-reported attention is sometimes referred to as perceived attention due to it being limited by self-awareness abilities. Performance-based tests are the most common tool used in laboratory-based experiments. With rigorous experimental controls, it is reliable to examine theoretical hypotheses via performance-based tests. However, the lab-based environments may sometimes restrict the generalizations of results into real and practical contexts.

Physiological measures are the most direct approach for the assessment of attention. Along with the rapid technological development in this area, an increasing amount of research has indicated that it is powerful to reveal humans’ implicit behaviors through the assessment of physiological measures. These implicit behaviors include humans’ visual attention, concentration, and metacognitive learning strategies. The traditional disadvantages of this method include the high costs of experimental equipment and could involve intrusive treatment. These problems may be changed by the rapid development of computer and image processing technology.

Finally, the three methods have different advantages and disadvantages. The selection of the methods depends on the purpose and the context of the problem to be resolved. Recently, researchers in technology-enhanced learning have begun to explore the potential of using the dynamic assessment of attention to provide personalized feedback for adapted learning.

Meng-Jung Tsai

See also Attention-Deficit/Hyperactivity Disorder; Cognitive Neuroscience; Information Processing Theory; Performance-Based Assessment; Self-Report Inventories; Working Memory

Further Readings

Mahone, E. M., & Schneider, H. E. (2012). Assessment of attention in preschoolers. Neuropsychology Review, 22, 361–383.

Oken, B. S., Salinsky, M. C., & Elsas, S. M. (2006). Vigilance, alertness, or sustained attention: Physiological basis and measurement. Clinical Neurophysiology, 117, 1885–1901.

Sternberg, R. J., & Sternberg, K. (2014). Cognitive psychology. Boston, MA: Cengage Learning.

Tsai, M. J., Huang, L. J., Hou, H. T., Hsu, C. Y., & Chiou, G. L. (2016). Visual behavior, flow and achievement in game-based learning. Computers … Education, 98, 115–129.

Patricia Teague Ashton Patricia Teague Ashton Ashton, Patricia Teague

Attention-Deficit/Hyperactivity Disorder Attention-deficit/hyperactivity disorder

134

137

Attention-Deficit/Hyperactivity Disorder

Attention-deficit/hyperactivity disorder (ADHD) is the term designated by the American Psychiatric Association in the fifth edition of its *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*) to refer to the set of three core psychological symptoms—inattention, excessive activity, and impulsivity—when those symptoms begin by age 12, persist for at least 6 months, and interfere with individuals' development and ability to perform the tasks of everyday living. This entry further describes ADHD and discusses its prevalence, the development of the understanding of ADHD and diagnostic criteria for the disorder, risk and protective factors, treatments for ADHD, measurement issues in the evaluation of ADHD, and promising advances in diagnosis and treatment of ADHD from a neuroscience perspective.

For individuals with severe symptoms, the effects of ADHD can have lifelong negative effects on all aspects of cognitive, emotional, and social development, leading to difficulties in learning to read, poor memory, academic failure and dropping out of school, problems at work, alcohol and drug abuse, disruptive relationships with parents, friends, and coworkers, and criminal behavior. According to the Centers for Disease Control and Prevention, 6.4 million children aged between 4 and 17 years, or about 11% of children in that age range, had received a diagnosis of ADHD as of 2011. Centers for Disease Control and Prevention estimates show that boys are more than twice as likely as girls to have the diagnosis. The toll that ADHD can take on personal lives as well as its costs to the economy and society make this topic particularly relevant to the issues of research, measurement, and evaluation.

Effective identification, diagnosis, intervention, and prevention of ADHD remain a significant challenge and depend on the development of greater insights into the nature and progression of ADHD. To achieve this understanding, the into the nature and progression of ADHD. To achieve this understanding, the development of reliable and valid measurement instruments and research to create powerful strategies for prevention and intervention for individuals with ADHD are needed.

The Concept and Diagnosis of ADHD

The three core symptoms of ADHD impact thinking, feelings, and behavior. Specifically, inattention refers to the inability to focus and sustain attention on relevant information. Typical indicators include making careless mistakes, frequent forgetting and losing of items, failing to complete assignments, and difficulties in organizing and planning. Hyperactivity/impulsivity refers to the inability to control one’s thoughts, emotions, and behavior. Indicators include constantly moving and running around, fidgeting and squirming, and interrupting the activities and conversations of others.

ADHD is an incurable, chronic condition that varies in severity from mild to severe. Once considered primarily a childhood disorder due largely to the negative impact the three symptoms have on school performance, ADHD emerges as early as age 3 and continues into adulthood for about 50% of those diagnosed with ADHD as children. Although estimates of ADHD vary widely depending on access to caregivers, estimates of its prevalence range from about 5% to 7% of the world population.

The conception of ADHD has evolved over time as researchers, clinicians, and educators have contributed to the development of knowledge on the topic. Problematic behaviors in children involving their inability to focus and sustain attention and to control activity level and emotional and behavioral impulses when necessary became an issue of concern during the early 1900s with the introduction of universal education. During the 1930s and 1940s, emphasis was placed on the role of brain damage as the source of hyperactivity, but the lack of reliable and valid measures of that damage led to a change in focus. In 1980, in the *DSM-3*, the American Psychiatric Association introduced the term *attention deficit disorder* with and without hyperactivity, placing the emphasis on the attention deficit and impulsivity rather than hyperactivity.

In the *DSM-5*, published in 2013, three types of ADHD are described: inattentive, hyperactive/impulsive, and combined. Two other notable changes in the *DSM-5*, the most widely used manual for diagnosing ADHD, are the switching of ADHD from the category of disruptive behavior disorders to the category of neurodevelopmental disorders, reflecting the increasing evidence from neurological studies showing differences in brain structure and functioning in individuals with ADHD compared to individuals without ADHD and greater recognition of the developmental differences in the ADHD-related behaviors that distinguish ADHD in children from ADHD in adolescents and ADHD in adults.

Risk Factors

No single, definitive cause for ADHD has been identified, although genetic influences are important. Children with ADHD are more likely to have one or both parents and siblings with the condition than are children without ADHD. No single gene for ADHD has been found, although multiple genes have been identified that may interact with each other and the environment to influence the development of ADHD. In his book *What Causes ADHD? Understanding What Goes Wrong and Why*, Joel Nigg describes multiple factors that are involved in the development of ADHD and the need to develop a greater understanding of the interplay of these factors and the multiple pathways to ADHD.

Nigg emphasizes the complexities to consider in children’s biological and psychological development as they are influenced by family, peers, and the school and community contexts over the course of development. Among the potential influences identified in research are, in particular, the mother’s behavior during pregnancy, including smoking, alcohol, and drug use, as well as stress. Exposure to toxins in the environment such as lead also places children at risk.

Children with severe ADHD are also at risk for other psychiatric problems. Although ADHD does not cause other mental disorders, the difficulties that individuals experience due to problems with inattention and hyperactivity/impulsivity can lead to learning disabilities, particularly in reading and mathematics, and to social problems that contribute to the development of anxiety and depression and in some instances to conduct disorders such as aggression and delinquency.

Protective Factors

Despite the difficulties that individuals with ADHD face, Despite the difficulties that individuals with ADHD face, numerous positive influences can lessen the negative effects of the disorder. Characteristics of the individual can influence access to support in the environment. For example, the individual’s attractiveness, intelligence, positive mood, problem-solving skills, and outgoing personality can elicit caring and support from parents, teachers, and peers.

Parents can offer crucial support for their children with ADHD. Examples include using a positive approach in helping children to create regular routines for daily activities; listening carefully and sensitively to their expressions of their needs; giving clear, brief, and reasonable directions to them; reducing distractions in the home; anticipating situations that will frustrate them; and preparing them for how they can manage those situations before they occur.

Similarly, educators can create school environments that provide accommodations for students with ADHD that enhance the focus and sustaining of attention, such as creating individually designed instruction to meet the students’ special needs, helping students keep records of their academic performance daily so they can benefit from seeing their progress over time, assisting with organization and planning and management of time, providing work spaces that minimize distractions and extended time for assessments, and most importantly, designing instruction that is interesting and appropriate for the students’ abilities. Similarly, strategies can be employed that enhance students’ ability to calm and control their impulses, such as computer-enhanced instruction through games and simulations.

Because ADHD is considered a disability, children may qualify for special accommodations under Section 504 of the Rehabilitation Act of 1973. However, they are not eligible for special education accommodations under the Individuals with Disabilities Education Act if their grades are average but may qualify if it can be shown that they have an impairment that interferes with their learning.

Treatments for ADHD

Researchers, clinicians, and educators have identified a variety of approaches for helping individuals cope with their ADHD symptoms. Given the possibility of the lifelong duration of ADHD, a comprehensive approach that involves the patient and multiple caregivers (e.g., medical doctor, psychologist, teacher, and, if appropriate, social worker) working together to identify the most appropriate goals, measurements, and treatments to create the best possible outcomes is generally recommended. More specifically, research indicates that behavioral therapy and medication prescribed carefully to fit the specific needs of the patient are most likely to result in positive effects.

Medications

The medications used in the treatment of ADHD are not cures. They are stimulants that reduce symptoms, but when they are discontinued, the symptoms return. Even though stimulant medications have been used for over 50 years to treat the symptoms of inattention and hyperactivity, their effect on children's health and well-being have still not been adequately studied over the long term. Over the short term, combined with behavioral therapy, these drugs can be helpful in enabling students to improve their academic performance and reduce their hyperactivity and impulsiveness. Typically, side effects are minor (e.g., stomach aches, low appetite, disrupted sleep); however, in some cases, the adverse effects can be life-threatening (e.g., heart and liver problems and suicidal thoughts).

Behavioral Therapy

Extensive research has demonstrated that therapies that focus on helping individuals with ADHD increase their ability to focus, sustain their attention, and control their activity and impulses have a long history of success in the management of the classroom behavior of students with ADHD. These approaches use prompts to promote positive behaviors followed by immediate rewards to reinforce those behaviors. Similar successes have been reported when these approaches have been adapted to train parents to use these strategies with their children and to help adults with ADHD develop better self-regulation.

Measurement Issues in the Evaluation and Treatment of ADHD

The identification of children and adults with ADHD is typically based on clinical judgments from observations, interviews, physical exams, psychological tests, and behavioral rating scales completed by the person with ADHD, parents, and teachers. In addition to the subjectivity biases that can limit the reliability and validity of these judgments and perhaps contribute to the overdiagnosis of and validity of these judgments and perhaps contribute to the overdiagnosis of ADHD, the susceptibility of ADHD symptoms to variations in the environment, the idiosyncrasies of individuals, fluctuations with age, and confounding with coexisting conditions complicates the problem of obtaining accurate diagnoses. In general, these diagnoses are not connected directly to specific strategies for reducing the negative effects of the three ADHD symptoms. However, psychological assessments of specific cognitive impairments related to inability to focus attention, for example, could be tied to specific treatments to address those impairments. Extensive research and development of more objective measures linked to specific treatments designed to ameliorate such deficits are needed to help individuals learn to cope effectively with their ADHD symptoms.

Recent Advances in Diagnosis and Treatment of ADHD

Advances in technologies to study the human brain, such as functional MRIs, have the potential to offer new insights into the connections between structural and functional networks in the brain and the symptoms of ADHD. These advances promise greater specificity in identifying the neurological bases of ADHD symptoms and possibilities for more accurate diagnoses and effective treatments. However, caution is warranted in assessing the significance of research findings in the early stages of this research. Studies are often based on correlational analyses and small samples that do not support causal conclusions.

Patricia Teague Ashton

See also Anxiety; Attention; Developmental Disabilities; *Diagnostic and Statistical Manual of Mental Disorders*; Individuals with Disabilities Education Act; Learning Disabilities

Further Readings

Centers for Disease Control and Prevention (2017, February 14). Children with ADHD. Retrieved from https://www.cdc.gov/ncbddd/adhd/data.html

DuPaul, G. J., & Stoner, G, (2014). ADHD in the schools. Assessment and intervention strategies (3rd ed.). New York, NY: Guilford Press.

Gualtieri, C. T., & Johnson, L. G. (2005). ADHD: Is objective diagnosis possible? Psychiatry, 2(11), 44–53.

Lange, K. W., Reichl, S., Lange, K. M., Tucha, L., & Tucha, O. (2010). The history of attention deficit hyperactivity disorder. Attention Deficit and Hyperactivity Disorders, 2(4), 241–255.

Nigg, J. (2006). What causes ADHD? Understanding what goes wrong and why. New York, NY: Guilford Press.

Reynolds, C. R., Vannest, K. J., & Harrison, J. R. (2012). The energetic brain: Understanding and managing ADHD. San Francisco, CA: Jossey-Bass.

Justin A. DeSimone Justin A. DeSimone DeSimone, Justin A.

James M. LeBreton James M. LeBreton LeBreton, James M.

Attenuation, Correction for Attenuation, correction for

137

139

Attenuation, Correction for

Charles Spearman noted that many variables, specifically those used in fields such as psychology and sociology, are measured using imperfect approximations. For example, a psychologist might be interested in understanding how cognitive ability is related to performance. The latent construct of cognitive ability could be measured in a number of different ways (e.g., Wonderlic Personnel Test; SAT). Likewise, the latent construct of performance could be measured in a number of different ways (e.g., the number of publications or patents generated by scientists, the overall GPA of undergraduate students). After selecting measurement devices and collecting data, the psychologist can correlate scores on the measure of ability (e.g., SAT) with scores on the measure of performance (e.g., GPA). However, because all measurement systems are subject to random measurement error, the correlation between observed measures will typically underestimate the “true” correlation between the latent constructs. The correction for attenuation is intended to estimate the value of this true correlation.

This entry first gives the formula for correction for attenuation, then discusses criticisms of the statistical procedure when it was first developed, and looks at the assumptions the procedure is subject to. Finally, it provides an example of the use of the procedure.

Formula

Spearman proffered a formula to estimate the true correlation as a function of the observed correlation and the reliability coefficients of each observed measure.

Following the example in the previous section, where *X* = SAT scores and *Y* = GPA scores, the correction for attenuation is given by:

$$\rho_{xy}` = r_{xy} r_{xx} \times r_{yy} ,$$

where $\rho_{xy}`$ represents the estimate of the true correlation between *X* and *Y*, r_{xy} represents the observed correlation, and r_{xx} and r_{yy} represent the observed reliabilities of *X* and *Y*, respectively. The correction for attenuation provides an estimate of the correlation between *X* and *Y* in the absence of random measurement error (i.e., if there were a one-to-one correspondence between observed test scores and latent construct scores). Thus, the correction for attenuation is often interpreted as an estimate of the correlation, not between observed measures, but between the unobserved, latent constructs.

Spearman differentiated between attenuation (random or “accidental” error) and “systematic deviations” or errors related to unmeasured variables that bias scores in a particular direction (e.g., practice effects, fatigue). Although systematic deviations may increase or decrease the magnitude of a correlation coefficient, attenuation always decreases the magnitude, and therefore, assuming that unbiased and accurate estimates are available for the observed correlation and the reliabilities of *X* and *Y*, the estimated true correlation coefficient will always be equal to or greater than the observed correlation coefficient.

Criticisms

The correction for attenuation garnered immediate criticism from Karl Pearson who chided Spearman for not presenting algebraic proof of his formula and for presenting a formula which, in cases of extremely poor measurement, could result in a correlation coefficient exceeding unity. Spearman replied by providing the algebraic proof, emphasizing that error is rarely truly random, and agreeing that science should focus on developing measurement techniques accurate enough to eliminate the need for this formula.

Mathematically, the correction for attenuation cannot yield a true correlation coefficient exceeding unity except in cases where the observed correlation exceeds the observed reliability estimates. Because a correlation coefficient cannot theoretically exceed the magnitude of the reliability of either variable (i.e., *X* or *Y*), the correction for attenuation will only yield true correlation coefficients exceeding unity when the observed correlation or observed reliabilities have been misestimated. It is noteworthy that, at the time Spearman introduced the correction for attenuation, the only techniques available for estimating reliability were correlations between parallel forms, subsequent administrations, or multiple raters. Split-half correlations and internal consistency (e.g., Cronbach’s α) coefficients did not exist at the time.

Lee Cronbach noted that different reliability coefficients estimate different aspects of a test (e.g., equivalence, stability). As a result, each reliability coefficient operationalizes error in a different way and, therefore, has different implications for use in the correction for attenuation. Cronbach also noted that the assumptions underlying the calculation of any given reliability coefficient are rarely met.

Assumptions

The correction for attenuation was derived using classical test theory and thus is subject to the same assumptions that underlie classical test theory. Specifically, true scores must be independent of errors and errors must be independent of one another. These assumptions ensure that the correlation between *X* and *Y* is not spurious (i.e., resulting from the relationship of both *X* and *Y* with a third variable) and that error is random as opposed to systematic. Because errors for *X* and *Y* are uncorrelated, the terms in the denominator (r_{xx} and r_{yy}) are considered independent. Thus, it is possible to correct for attenuation in either *X* or *Y* while ignoring attenuation caused by the other variable ($r_{xy}r_{xx}$ or $r_{xy}r_{yy}$).

The correction for attenuation is related to the Spearman-Brown prophecy formula, which estimates the expected increase in an observed reliability coefficient as a function of the number of parallel measurements added to the test. The true correlation coefficient estimated using the correction for attenuation represents the hypothetical value one might obtain if perfectly reliable measures of *X* and *Y* were available. One method of obtaining a perfectly reliable measure is to administer an infinite number of parallel measurements. In the Spearman-Brown prophesy formula, increasing the number of measurements by a factor of ∞ will increase all reliability coefficients (except .00) to 1.00. As a result, the correction for attenuation can yield an estimate of the correlation coefficient if it were computed using infinitely long measures of *X* and *Y*.

Example

In order to calculate the correction for attenuation, one must first calculate the observed correlation coefficient and estimates of reliability for the two variables (*X* and *Y*). Remember that correlations cannot theoretically exceed reliabilities. If the correlation coefficient is higher than either reliability estimate, at least one of these has been misestimated.

Observed correlation: .35

Reliability estimate for *X*: .72

Reliability estimate for *Y*: .89

Correlation corrected for attenuation in *X* only: $r_{xy}r_{xx}$=.35.72=.35.85=.41.

Correlation corrected for attenuation in *Y* only: $r_{xy}r_{yy}$=.35.89=.35.94=.37.

Correlation corrected for attenuation in *X* and *Y*: $r_{xy}r_{xx}\times r_{yy}$=.35.72×.89=.35.64=.35.80=.44.

Justin A. DeSimone and James M. LeBreton

See also Classical Test Theory; Correlation; Meta-Analysis; Reliability; Restriction of Range; Spearman-Brown Prophecy Formula; Validity Generalization

Further Readings

Cronbach, L. J. (1947). Test “reliability”: Its meaning and determination. Psychometrika, 12, 1–16.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion unreliability in validity generalization: A false prophet in a land of suspended judgment. Industrial and Organizational Psychology: Perspectives on Science and Practice, 7, 478–500.

Schmidt, F. L., & Hunter, J. E. (2015). Methods of meta-analysis: Correcting error and bias in research findings (3rd ed.). Thousand Oaks, CA: Sage.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

Benjamin D. Rosenberg Benjamin D. Rosenberg Rosenberg, Benjamin D.

Timothy C. Silva Timothy C. Silva Silva, Timothy C.

Attitude Scaling

Attitude scaling

139

144

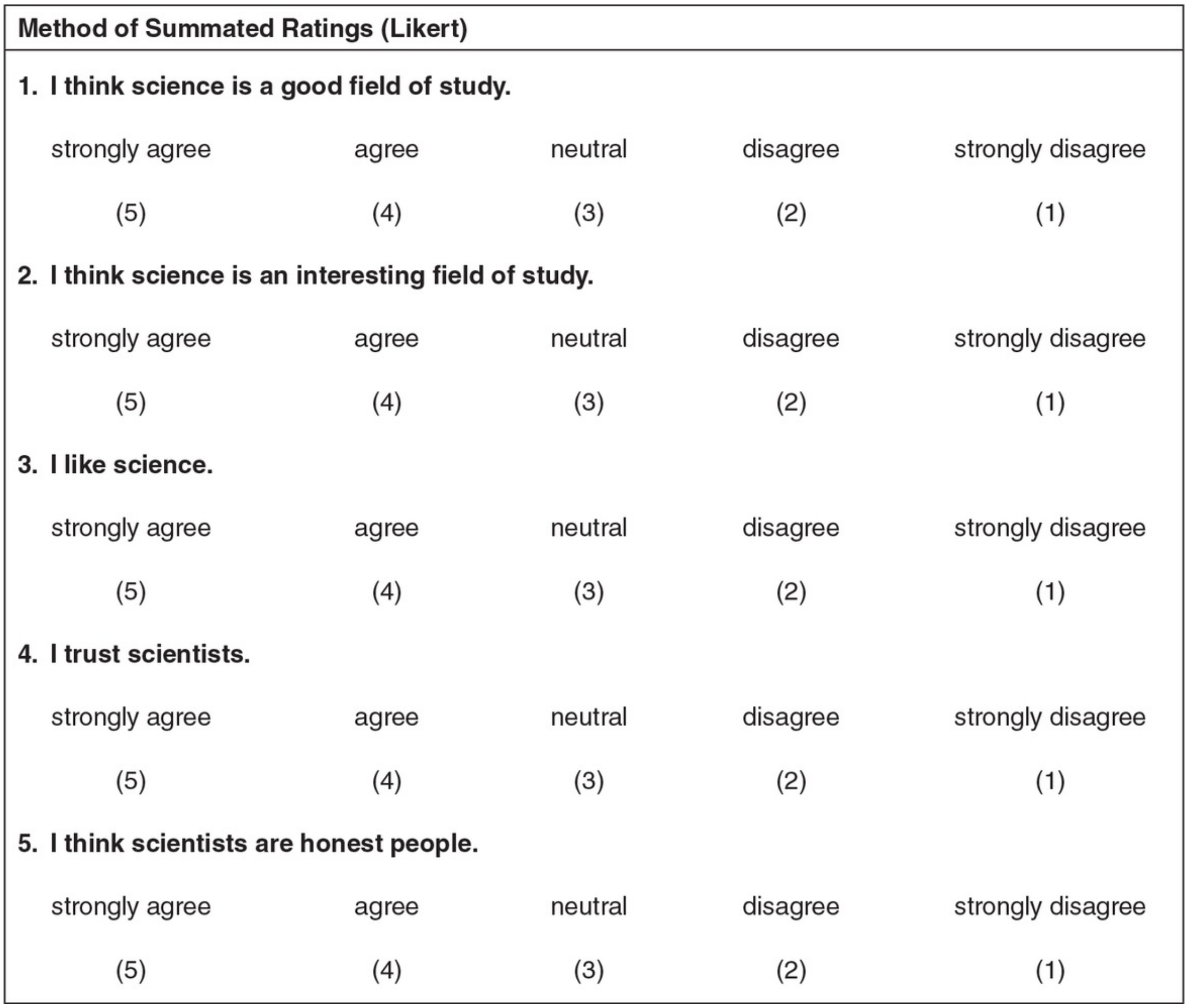
Attitude Scaling

Attitudes represent people’s overall evaluation of another person or object, which include cognitive and affective components. In general, attitudes vary in strength and lie on a continuum that ranges from unfavorable to favorable. Researchers cannot directly observe people’s attitudes and thus need to infer them by observing behavior, or by direct or indirect measurement, as through attitude scales. This entry covers the history of attitude scaling, the aspects to consider when creating methodically strong attitude scales, and future directions in the area of attitude scaling.

The concept of attitudes was introduced in social psychology and continues to play a prominent role in a wide variety of fields today (e.g., public health, communication, marketing). Early scholars such as Gordon Allport helped define the concept of an attitude, while researchers such as Louis Thurstone and Charles Osgood pioneered attitude scaling techniques that laid the foundation for their scientific study.

Attitude scales measure participants’ internal dispositions or attitudes toward a particular object or set of objects via self-report. For instance, if researchers are interested in measuring students’ attitudes toward science, they might ask participants how much they disagree or agree with a series of statements about the various fields of science (see Figure 1).

Figure 1 Example of Likert-type scale “Attitudes Toward Science”



Method of Summated Ratings (Likert)

1. I think science is a good field of study.

strongly agree	agree	neutral	disagree	strongly disagree
(5)	(4)	(3)	(2)	(1)

2. I think science is an interesting field of study.

strongly agree	agree	neutral	disagree	strongly disagree
(5)	(4)	(3)	(2)	(1)

3. I like science.

strongly agree	agree	neutral	disagree	strongly disagree
(5)	(4)	(3)	(2)	(1)

4. I trust scientists.

strongly agree	agree	neutral	disagree	strongly disagree
(5)	(4)	(3)	(2)	(1)

5. I think scientists are honest people.

strongly agree	agree	neutral	disagree	strongly disagree
(5)	(4)	(3)	(2)	(1)

Attitude Scales

Rating scales have been an important tool for measuring people’s beliefs, opinions, and attitudes in the last century of social scientific research. Since Allport proposed the concept of attitudes over 100 years ago, researchers have devised many approaches for their measurement. This section outlines the original techniques (Thurstone’s method of equal-appearing intervals and Guttman’s scalogram) and those most commonly used now (Likert’s method of summated ratings and Osgood’s semantic differential). There is not necessarily one method that is best at achieving accurate results. Instead, researchers have a variety of tools to choose and must consider the appropriateness of each type for the specific context at hand.

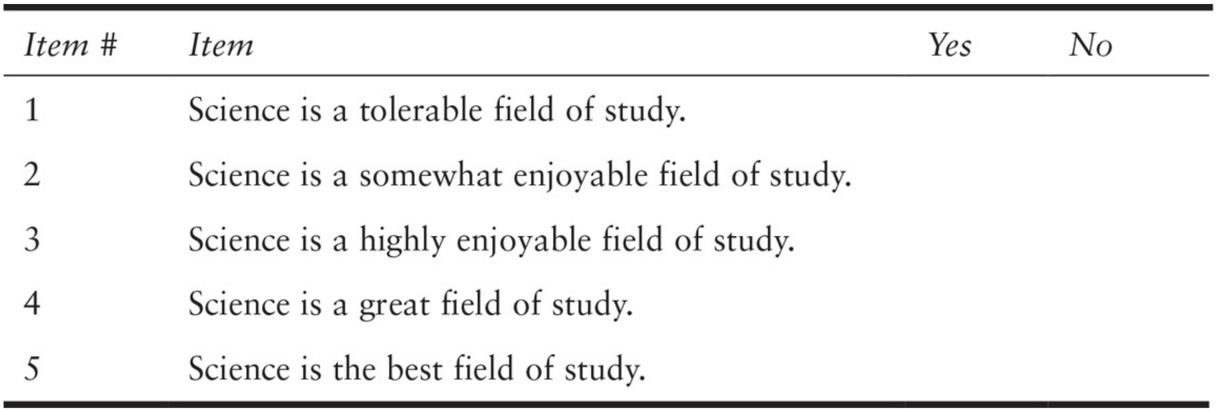
Thurstone’s Method of Equal-Appearing Intervals

Thurstone’s Method of Equal-Appearing Intervals

In 1928, Thurstone developed the first systematic way to measure attitudes. The method of *equal-appearing intervals* involves four phases of scale construction. After the researcher decides what attitude is to be measured, the first phase involves generating many possible questions to cover all aspects of the construct of interest. In the second phase, a group of judges rates the items on their favorability, which allows researchers to assess the psychometric properties of the scale. In the third phase, researchers subject the judges’ ratings to statistical analyses, using the results to choose 11–22 questions that constitute the final scale. The last phase is to administer the scale to participants who indicate whether they disagree or agree with each item. To get an overall idea of people’s attitudes, their responses to all of the items are averaged—but counterintuitively, higher averages do not necessarily indicate more favorable attitudes toward the person or object under consideration.

Guttman’s Scalogram

In 1944, Louis Guttman attempted to improve on Thurstone’s method by developing a scaling method where participants with more favorable attitudes toward an object would, in fact, have higher total scores on the scale. Scores across items can be averaged to form a cumulative score representing the favorability toward the object under investigation (see Table 1).



Item #	*Item*	*Yes*	*No*
1	Science is a tolerable field of study.		
2	Science is a somewhat enjoyable field of study.		
3	Science is a highly enjoyable field of study.		
4	Science is a great field of study.		
5	Science is the best field of study.		

Using Guttman’s method to create an attitude scale is very similar to Thurstone’s technique—investigators create a large set of items that encompass the attitude under consideration. Next, a set of judges rates the items in terms of favorability in a yes or no manner; the judges’ ratings are then tabulated hierarchically from items with the highest level of agreement to those with the lowest level of agreement. From this matrix, additional statistical analysis is conducted to finalize the scale. Lastly, the scale is administered to participants and their summed scale values represent their attitude toward the object. Due to the difficulties associated with item creation and selection, scholars today do not use either Thurstone’s or Guttman’s methods very frequently. Instead, researchers more often use the next two attitude scaling methods: Likert-type and semantic differential scales.

Likert’s Method of Summated Ratings

Rensis Likert took the next step in attitude scaling in 1932, when he developed a method that was more efficient in time and resources and more effective than both Thurstone’s and Guttman’s methods. The two previous methods required participants to choose from just two options (e.g., yes or no, agree or disagree). Likert’s new method used a multiple choice format in which people placed their response on a 5-point scale from *strongly disagree* to *strongly agree* with a *neutral or undecided* middle point. Each point along the scale would be given a value of 1 through 5, and participants’ responses would be summed or averaged to indicate their overall attitude toward the person or object under investigation, as shown in Figure 1.

Only the initial process of the Likert’s method resembles Thurstone’s and Guttman’s methods, as researchers develop a large potential set of questions. However, instead of finalizing a set of scale items that represent the attitude as a whole, in the Likert’s method, researchers select items that are moderately favorable or moderately unfavorable with regard to the attitude object. As opposed to having a group of judges rate the items, the second phase in the Likert’s method involves administering the set of items directly to respondents. A good rule of thumb is to multiply the number of questions to be administered by 10 to have an appropriate amount of respondents during this phase of the scale construction process. Once the responses have been collected, researchers apply statistical techniques such as factor analysis to the data to retain the items that will form the best final scale.

Likert’s method advanced attitude scaling from Thurstone’s and Guttman’s methods; however, it shared in some of their liabilities—namely, that these three methods are relatively time and resource consuming, and new scales must be created every time a new attitude object is to be measured.

Osgood’s Semantic Differential

In contrast to the three methods discussed to this point, in which people respond to statements about the concept under investigation, in 1952, Osgood proposed the semantic differential, in which people evaluate the person or attitude object directly using bipolar adjectives. For example, if researchers were investigating attitudes toward science, the concept of science is presented and then participants respond to a 5-point scale anchored by adjectives (e.g., good/bad, pleasant/unpleasant). These anchors can then be used for measuring different attitude objects without having to go through the time-consuming process of scale creation and validation as with the other three methods discussed. In comparison to Likert-type scales, semantic differential scales are shorter, easier to understand, can be completed more quickly, and are highly efficient in the scale creation process. Overall, this method is highly practical and efficient for researchers.

Attitude Scale Creation

When developing attitude scales, there are certain aspects to keep in mind, including question wording, response type, question ordering, and no opinion options. With a solid understanding of what constitutes a good scale, researchers can develop more accurate and efficient scales, saving both time and resources.

Wording Questions

There are two important issues that researchers should consider when deciding on how attitude questions should be worded. First, investigators must decide what they would like to know—questions meant to measure students’ evaluation of science might be very different than those assessing academics’ attitudes toward coffee. Additionally, researchers should take care to ask questions directly so that people are more likely to understand their true meaning. Similarly, researchers should use short, simple sentences that contain only one grammatical clause. When sentences contain more than one clause, these *double-barreled questions* can lead to ambiguity on the part of the respondent.

Question Ordering

A third issue that scale creators must consider is the order in which questions are presented. One important concern is that participants may be uncomfortable answering questions concerning their attitudes toward sensitive subjects (e.g., drug and alcohol use, sexual behavior, stereotypes). To account for this possibility, researchers should present the least threatening questions first and gradually work toward more sensitive material. Additionally, demographic questions (e.g., income, race/ethnicity) should be posed toward the end of the questionnaire because these too may be sensitive for participants.

Researchers must also consider the fact that when ordering questions, earlier questions and answers may affect later ones. For instance, investigators should be aware of priming effects and their ability to invoke a particular mind-set about framing and responding to questions. Counterbalancing or randomizing question order can help avoid any unintended ordering effects. When using written survey materials, the process of randomization can be time-consuming, so computer software can be highly effective at this task.

Dropout and the No-Opinion Response Format

Finally, researchers should attempt to limit participants from dropping out of the survey, as it may create problems for the generalizability of findings. When considering how to get participants to respond to all questions, researchers must consider a “no-opinion” or “don’t know” response option. Some questionnaires allow respondents to indicate a neutral response, while others force participants to indicate a preference on either side of the response continuum.

Researchers have investigated both response types for many years, and there are strong arguments for either the inclusion or exclusion of a neutral response option. Some scholars have suggested that providing a no-response option allows participants to avoid the cognitive work necessary to answer questions. Yet others indicate that having a no-response option may affect participants’ interpretation of other response options. In general, the benefits of providing such an option seem to outweigh the negative aspects, and participants seem to prefer having such an option. One concern, however, is that even though they mean different things, participants often confuse “neutral,” “not applicable,” and “no opinion” response options. Thus, when a neutral response option is given, it is advisable to make clear to participants what such a response indicates.

Additional Considerations

Additional Considerations

Scale Length

When researchers first developed attitude measures, they used elaborate methods to create scales and they believed that large question sets were needed to accurately assess attitudes. However, these methods took great time and resources to both create and administer. Over time, due to a better methodological understanding of measurement scaling, researchers have concluded that shorter scales have clear advantages. Methodologically sound condensed scales can be just as reliable and accurate as longer scaling techniques. They also have the added benefit of taking less time for respondents to complete, thus limiting fatigue and potential for participants to drop out.

Monetary Consideration

Questionnaire design and assessment may have impositions based on monetary considerations. For instance, when conducting a national telephone survey, adding one or more questions will increase the time necessary for each telephone call. Even if 1 extra minute is added to each interaction, over thousands of calls, this extra cost multiplies quickly. While in some cases using multiple questions or scales to tap people’s attitudes may be preferable, researchers sometimes may have to be content with shorter, more limited questionnaires.

Context

When people report their attitudes, the context of the situation needs to be taken into account because these reports may vary due to the context in which they were measured. For instance, if Americans were polled about their attitudes toward national security, these attitudes would likely vary if asked directly after a terrorist attack as opposed to being asked after years of relative calm. Additionally, if participants were to answer questions in the confines of a research lab, their responses may be different than in a “real-world” context. This may be due to participants in the lab wanting to be seen in a socially desirable way by the research team or perhaps the lab not providing the same real-world conditions as experienced in day-to-day life. As a result, an attitude measurement may only be useful in predicting future attitude in the environment in which it was measured.

Beyond Self-Report

Underlying all the measurement techniques discussed in this entry is a core concern—namely, that social context and social desirability can influence people’s self-reported attitudes. Indeed, many times participants may fail to recognize the impact that their surroundings and/or internal motivations have on their reports about themselves. These biases can negatively affect the quality of the data the researchers gather; they can also lead to incorrect conclusions about people’s attitudes and opinions.

In response to these realizations, scholars have developed measurement techniques for attitudes that do not rely on self-report—the so-called *implicit* measures, which assess attitudes indirectly (i.e., without directly asking people about them). These strategies for measuring attitudes disguise what attitude is actually being measured and may be effective at limiting the impact of participants wanting to be seen in a socially desirable light. Moreover, participants may fail to recognize the influence that their attitudes have on their behavior, thus giving researchers a more “real-world” friendly measurement of attitudes.

Two key implicit measures have been developed. First, researchers can unobtrusively observe people’s behavior; this technique has been widely used, including measuring helping behavior and social distance. In terms of indirectly measuring attitudes, techniques such as the implicit association test measure the strength of association between two concepts.

The implicit association test assesses the time it takes people to respond to different attitude objects in reference to negative or positive adjectives. If there is a bias toward one object over another (e.g., a preference for white faces over black faces), there will be a difference in the time it takes to respond to positive/negative words paired with the attitude object. This method is the most widely used implicit measure and has been used in a variety of domains including attitudes toward race, gender, and religion.

Future Directions

The number of online environments for completing attitudinal scales has been increasing in recent years. Sites such as Amazon’s Mechanical Turk can be increasing in recent years. Sites such as Amazon's Mechanical Turk can be beneficial for data collection and analysis in numerous ways. For one, researchers conducting studies online can use myriad formats and types of attitude scales—even beyond the traditional ones covered in this entry. For instance, recent research examined the accuracy of sliding 100-or 250-point semantic differential scales in online samples. The use of online techniques to measure attitudes also offers the ability to collect vast amounts of data quickly—compared to administering traditional pen and paper scales, the collection of online data can take mere hours for hundreds or thousands of responses, which traditionally can take months or even years. The ability to gather data quickly and efficiently in this online environment enables researchers to more quickly create and validate new attitude scales.

Benjamin D. Rosenberg and Timothy C. Silva

See also Instrumentation; Rating Scales; Self-Report Inventories; Semantic Differential Scaling; Survey Methods; Surveys

Further Readings

Allport, G. W. (1935). Attitudes. In C. M. Murchison (Ed.), Handbook of social psychology. Winchester, MA: Clark University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6, 3–5.

Eagly, A. H., & Chaiken, S. (2005). Attitude research in the 21st century: The current state of knowledge. Mahwah, NJ: Erlbaum.

Krosnick, J. A, Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), Handbook of attitudes and attitude change. Mahwah, NJ: Erlbaum.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33, 529–554.

Shlomo Hareli Shlomo Hareli Hareli, Shlomo

Attribution Theory

Attribution theory

144

147

Attribution Theory

People share a great thirst to understand the causes of situations they encounter and often attempt to explain—to themselves or others—why a specific situation occurred. *Attribution theory* explains the connection between perceived causes of situations and the psychological consequences of these perceptions. The main idea of the theory is that all causes share three basic, underlying properties: locus, controllability, and stability; these properties determine the psychological consequences of perceived causes. Perceived causes have crucial emotional and behavioral consequences, including those related to the context of achievement motivation.

Much of the existing understanding of the process explained by attribution theory comes from research conducted in the context of school achievement. Individuals’ attempt to understand the causes of their achievement in school often determines their reactions to these causes. This entry further explores the search for these causes and its psychological consequences, then looks at how the causes a person ascribes to events determines his or her psychological reality and how people use attribution theory in their dealings with others or to improve their performance.

Search for Causes of Events and Outcomes

People aspire to understand why different events and outcomes occur. This aspiration is motivated by humans’ innate desire to both understand their environment and to use this understanding to effectively manage their lives. To this end, people often engage in attempts to explain to themselves and to others why an event or outcome came about.

Because the understanding of the reasons for an outcome or event helps people manage their lives, is has important emotional and behavioral consequences. For example, a student who failed an exam may come to the conclusion that this failure was caused by insufficient efforts to study for the exam. In consequence, the student may feel guilty and decide to invest more effort when studying for a future exam. By contrast, if the student thinks that the failure was caused by a lack of ability, this student is likely to feel shame and may decide to quit studies or move to a different field of study. As this example suggests, the cause the student attributes to the outcome determines which emotions are likely to arise and what type of behavior may result from it.

More generally, causal beliefs give rise to emotional reactions and to a variety of inferences, both in the actors who attribute their good or poor performance or situation to various causes and in the involved observers of this performance. Thus, it is not only students who may react to their poor outcome as a function of what they think caused the failure but also, for example, a teacher who also holds a certain belief about what caused this performance. If the teacher thinks that low effort caused the failure, the teacher may react with anger and punish the student. If, on the other hand, the failure is attributed to low ability, then the teacher is more likely to react with pity. This teacher may also infer that the student is lazy or unintelligent, as a function of each respective attribution. Thus, the way people explain events and outcomes determines how they respond to them. This is true for all domains of life, not only for achievements.

How Causes Determine Psychological Reality

A myriad of distinct possible causes can determine a given outcome, and there is an endless number of potential outcomes and events that can occur in different contexts. This makes it rather difficult to understand why a particular cause for a specific outcome leads to a specific consequence and not to another. Why is it the case that failure in an exam attributed to low ability leads to feelings of shame in the failing student? Why does the rejection of an invitation to a romantic date, attributed to appearance, also lead to shame?

One way to resolve the complexity of the connection between causes and consequences is by searching for a possible underlying structure of the main factor of interest, in this case, causes. By finding similarities and differences between different causes, one can reveal some underlying structure of causality. The next step would be to examine if and to what extent this underlying structure can explain the connection between causes and their consequences. This will enable the narrowing down of a rather complex phenomenon to a set of simpler unifying features that define it. Bernard Weiner's attribution theory, devised in 1985, does just that.

Attribution theory reveals the underlying structure of causality by describing the properties or dimensions that define all causes. Furthermore, it describes how dimensions of causes are related to specific types of psychological consequences. Thus, attribution theory is based on the understanding that all causes can be characterized according to three basic properties: labeled locus, controllability, and stability.

Locus refers to the location of a cause, that is, whether the cause is internal or external to the actor. For example, both low effort and low ability are likely to be perceived as internal to an actor; something that is associated with the actor rather than with the situation or someone else. On the other hand, a difficult or unfair exam as a cause for failure is associated with someone else—such as a teacher—and not with the actor.

Controllability refers to the degree to which the cause is subject to volitional change, that is, the extent to which the cause is controllable or uncontrollable. Thus, low effort is within the student's control because a student can decide how much effort to invest in studying for an exam. By contrast, low ability is more likely to be perceived as an uncontrollable factor, as a person cannot control the extent to which the person is endowed with skills or abilities. An unfair or difficult exam is also likely to be perceived as being within the teacher's control.

Stability pertains to the relative endurance of a cause over time. Whereas enduring causes are seen as stable, transitory ones are seen as unstable. In our examples, low effort is likely to be perceived as unstable as on a different occasion, in principle, the student may invest more effort in studying for an exam. Alternatively, low ability is stable because basic traits and skills are perceived as being unlikely to change much or at all over time.

These characteristics of causes are independent of one another such that the fact that a specific cause is characterized by a given location on one of the dimensions does not force any specific location on another dimension. In other words, causes can represent any combination of these dimensions. Furthermore, all causes can be seen as representing a combination of different locations within each of these dimensions. Thus, low effort as a cause for failure is likely to be perceived as internal to the actor, controllable, and unstable, whereas low ability is perceived as internal, uncontrollable, and stable.

Although the dimensional placement of a cause is a subjective reality—meaning, individuals may disagree with respect to a causal interpretation—there is a great deal of consistency concerning the characteristics of particular attributions. In other words, whereas most people may perceive luck, for example, as a cause of success that is external to the person, uncontrollable, and unstable, others may perceive luck as internal, uncontrollable, and stable. That is, instead of perceiving luck as representing a set of accidental circumstances unrelated to the intentional behavior of the actor, some may perceive luck as representing the property or trait of an individual, something this person is endowed with. What is common, however, is the dimensional structure of causes, as presented earlier, and the consequences of particular attributions as a function of their causal properties, as is described next.

As the examples given earlier clearly indicate, the perceived cause of a given event or outcome determines its emotional and behavioral consequences. The link between the perceived cause and its consequences is indirect, being mediated via the perceived dimensions of the cause. In other words, the perceived properties of a cause determine its consequences.

Each causal dimension has its unique psychological significance. The locus of a cause is linked to self-esteem and related emotions such as pride. Desirable outcomes attributed to internal causes lead to greater self-esteem and pride than the same outcomes attributed to external causes. By contrast, undesirable outcomes attributed to internal causes lead to lower self-esteem than the same outcomes when attributed to an external cause. For example, success in an exam attributed to high ability or to effort leads to increased self-esteem and pride. Success due to luck or an easy exam does not lead to the same consequences. Failure due to a lack of ability or low effort will lower the achiever’s self-esteem. However, failure due to an unfair exam will not lower one’s self-esteem.

The stability of a cause determines expectations about the future as well as the emotions and behaviors related to such expectations. When a given outcome or event is attributed to a stable cause, it is expected that the event or outcome will reoccur in the future. However, when the occurrence of an event or an outcome is attributed to an unstable cause, it is not necessarily expected that it will is attributed to an unstable cause, it is not necessarily expected that it will reoccur. For example, a failure attributed to low ability will lead to expectations of similar failures in the future because low ability is a stable cause. On the other hand, failure attributed to low effort is not necessarily expected to reoccur because low effort is unstable; that is, the situation can be changed by investing more effort.

Attribution of failure to low ability will also lead to hopelessness, because nothing much can be either done or hoped for, given the stable nature of the failure's cause. A change in behavior that better suits the abilities of the achiever may therefore be the result of this attribution. On the other hand, hope is a likely emotional consequence if the failure is attributed to low effort because the fact that the cause is changeable indicates the possibility that the undesirable outcome may also change.

Controllability of a cause determines inferences of responsibility as well as the emotions and behaviors related to it. A situation or event attributed to a controllable cause leads to the inference that the person who had control over the circumstances that brought it about is also responsible for the outcome. Alternatively, a situation or event attributed to an uncontrollable cause leads to an inference that the person of relevance is not responsible for it. For example, a student who failed an exam because of low effort is likely to take responsibility for the failure, as effort can be controlled. As a result, this student will feel guilty and may decide to invest more effort in the future. The same failure attributed to low ability, however, will lead to shame because the cause of the failure—the student's innate ability—is not controllable.

Judgments and emotions elicited as a result of the behavior of others also depend on attributions about responsibility for the behavior. For example, an observer is more likely to feel pity and offer to help a person encountering an undesirable situation if it is attributed to an uncontrollable cause, such as a disease caused by a genetic defect or an accident caused by the force of nature. Yet, if the same situation is attributed to a controllable cause—such as reckless behavior—anger and avoidance are more likely reactions.

To summarize, perceived causes of events and situations determine the related psychological consequences of these events and situations. The common underlying structure of causes determines their psychological consequences.

How People Use Attribution Theory

How People Use Attribution Theory

An important derivative of research in the context of attribution theory shows that people are aware of the links between specific causes, emotions, and behaviors as described by the theory. As such, they often use this knowledge to make sense of their social surroundings and ensure that their goals are fulfilled or to improve their performance. Hence, for example, witnessing a student expressing guilt in response to failing an exam, a teacher may understand that this student didn’t invest sufficient effort in studying for the exam. This conclusion comes from the naive understanding that guilt reflects a sense of responsibility for the failure. Training people to replace undesirable attributions with desirable ones helps people improve their performance. For example, persuading students to take control over failures rather than blame them on uncontrollable causes improves their performance in school.

Shlomo Hareli

See also Educational Psychology; Emotional Intelligence; Motivation

Further Readings

Hareli, S. (2014). Making sense of the social world and influencing it by using a naïve attribution theory of emotions. Emotion Review, 6(4), 336–343. doi:10.1177/1754073914534501

Hareli, S., & Weiner, B. (2002). Social emotions and personality inferences: A scaffold for a new direction in the study of achievement motivation. Educational Psychologist, 37(3), 183–193. doi:10.1207/S15326985ep3703_4

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. Psychological Review, 92(4), 548–573. doi:10.1037/0033–295X.92.4.548

Weiner, B. (1986). An attributional theory of motivation and emotion. New York, NY: Springer.

Weiner, B. (1987). The social psychology of emotion: Applications of a naive psychology. Journal of Social and Clinical Psychology, 5(4), 405–419. doi:10.1521/jscp.1987.5.4.405

Weiner, B. (1995). Judgements of responsibility: A foundation for a theory of social conduct. New York, NY: Guilford Press.

Weiner, B., Amirkhan, J., Folkes, V. S., & Verette, J. A. (1987). An attributional analysis of excuse giving: Studies of a naive theory of emotion. Journal of Personality and Social Psychology, 52(2), 316–324. Retrieved from http://dx.doi.org/10.1037/0022–3514.52.2.316

Jordan R. Bass Jordan R. Bass Bass, Jordan R.

Auditing

Auditing

147

148

Auditing

When referring to auditing within qualitative research, numerous definitions exist. In short, auditing refers to a transparent research process where each step of inquiry is clearly presented and analyzed. Auditing is often represented through an audit trail where the data are essentially tracked from the raw form to the ultimate finished product, which could range from a narrative of rich description to a more formalized research instrument or scale. The use of the term and process of auditing have similarities to the concepts of reliability, generalizability, and validity (what Steinar Kvale calls the “scientific holy trinity”)—terms more popular in quantitative or postpositivist analysis arenas. This entry describes audit trails for qualitative inquiry and the debate over the use of auditing strategies in qualitative research.

Audit Trail

The trustworthiness of a qualitative research process is often shown through a transparent demonstration of the totality of the process of inquiry. In essence, the reader of the study can become immersed in each stage of the research and understand the decisions made at each stage. To some, this increases the quality of the work and assures the results are not a result of deception, fraud, or manipulation. Yvonna Lincoln and Egon Guba are credited with the original conceptualization of an audit trail for qualitative inquiry where a third party could theoretically follow each step of the study and recreate, or confirm, the results. Lincoln and Guba argued for six categories of data that can help inform a proper audit trail:

1. raw data,
2. data reduction and analysis notes,
3. data reconstruction and synthesis products,
4. process notes,
5. materials related to intentions, and
6. preliminary developmental information.

Marian Carcary has further clarified that an audit trail can be “intellectual” or “physical.” A physical audit trail deals with the “nuts and bolts” of the research process from the initial identification of the research problem to the resulting theory or instrument created as a result of the inquiry. The intellectual trail delves into decisions surrounding the internal thinking of the researcher throughout the process and the ways in which the researcher’s own biases and dispositions influenced the procedure. In all, an audit trail serves as a way to enhance the trustworthiness and credibility of qualitative research.

An example of both trails can be found in Carcary’s (2009) article. The author detailed her intellectual transition, describing how she questioned her traditional positivist beliefs and ultimately selected an adapted grounded theory approach. In the physical audit, the steps of the research process are clearly laid out with supporting information for each decision. For example, she describes her interview schedule:

> The semi-structured interview was the primary source of case-study evidence. Based on issues identified in the literature and in defining the research problem, an initial interview schedule was prepared. This was pre-tested in a number of pilot interviews in order to determine informants understanding of the questions and the depth of the research inquiry, and was subsequently refined. (p. 20)

Auditing Moving Forward

Considerable debate still exists over the best methods, if any, for measuring credibility, reliability, validity, and transferability in qualitative research. Many qualitative researchers have cautioned against the adoption of largely positivist ideas to “justify” or “give credibility to” qualitative work. Pierre Bourdieu has warned against a global audit culture where results and processes are scrutinized through a governance lens that ultimately influences the findings to a far greater degree than the types of audits previously discussed.

Other researchers have moved past the auditing terminology and rely on more constructivist terms such as *verification*. For example, Janice Morse and her colleagues argued for verification through (a) methodological coherence, (b) appropriate sampling, (c) collecting and analyzing data concurrently, (d) thinking theoretically, and (e) theory development. No matter what term is used, the debate about the use of auditing strategies in qualitative research likely will continue.

Jordan R. Bass

See also Grounded Theory; Mixed Methods Research; Qualitative Data Analysis; Qualitative Research Methods; Reliability; Validity

Further Readings

Carcary, M. (2009). The research audit trail—Enhancing trustworthiness in qualitative inquiry. The Electronic Journal of Business Research Methods, 7(1), 11–24.

Denzin, N. K. (2011). The politics of evidence. The SAGE handbook of qualitative research (pp. 645–658).

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry (Vol. 75). Thousand Oaks, CA: Sage.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. International Journal of Qualitative Methods, 1(2), 13–22.

Frederick Burrack Frederick Burrack Burrack, Frederick

Authentic Assessment Authentic assessment

148

151

Authentic Assessment

Authentic assessment is an approach to student assessment that involves the student deeply, is cognitively complex and intrinsically interesting, uses a format that is consistent with how ability is evaluated in the real-world, and evaluates skills and abilities that have value and meaning outside of the classroom or on the job. Educational scholar Grant Wiggins, who is credited with introducing the concept, describes authentic assessment as involving those activities or tasks that people actually do in the real-world. Indeed, *authentic* is often treated as a synonym for realistic. This entry further defines authentic assessment and discusses how it compares to traditional assessment.

Authentic assessment focuses on how students integrate and apply what they have learned through contextualized tasks. This form of assessment allows students to demonstrate learning individually or by working collaboratively with others to demonstrate competency in authentic settings. Authentic assessment usually describes classroom assessment, but the philosophy has been applied to standardized tests as well.

One goal of authentic assessment is to indicate the extent to which a student’s knowledge and skills can be applied outside of the classroom. It might also be referred to as direct assessment as opposed to traditional formats (such as multiple-choice questions) that seldom require a direct demonstration of knowledge and skills. Because authentic assessment strategies do not focus entirely on recalling facts, students are required to integrate, apply, and self-assess skills and understanding. Student understanding of disciplinary content is desired, but it is also important for students to be able to use the acquired knowledge and skills in the world beyond their classes.

Assessments have to indicate whether students can apply what they have learned

Assessments have to indicate whether students can apply what they have learned in authentic situations. When a student does well on a test of knowledge, this often infers that the student can also apply that knowledge, but that is indirect evidence. Knowledge tests can also provide evidence of knowledge about application, but again that is indirect.

Authentic assessments ask the student to use what the student has learned in a meaningful way. For example, it would not be possible to determine whether students can effectively debate a topic by listening and responding to contrasting views through multiple-choice questions or a description on a written test. Authentic assessment is designed to produce direct evidence in an authentic context. Similarly, authentic assessment can demonstrate whether students can interpret a current news story, calculate potential savings of a proposed budget, test a scientific hypothesis, play a musical instrument, converse in a foreign language, or apply other knowledge and skills they have learned.

Bruce Frey, Vicki Schmitt, and Justin Allen analyzed the concept of authentic as applied to assessment and identified nine dimensions of authenticity used in the literature. Researchers and teacher educators refer to an assessment as authentic when it has several of the following characteristics (Frey, Schmitt, … Allen, 2012, p. 5):

- the context of the assessment
    - realistic activity or context
    - the task is performance based
    - the task is cognitively complex
- the role of the student
    - a defense of the answer or product is required
    - the assessment is formative
    - students collaborate with each other or with the teacher
- the scoring
    - the scoring criteria are known or student developed
    - multiple indicators or portfolios are used for scoring
    - the performance expectation is *mastery*.

Authentic assessments nearly always are patterned after tasks that require performance of skills, supported by a foundation of required knowledge, at an achievement level at or beyond *what is expected* in the school classroom. A framework for authentic assessments begins the same way that curriculum for a program would be designed, by asking what students should be able to do as a result of what has been learned. Some examples of authentic assessments include simulations and role plays, lab experiments, budget proposals, application letters, and tasks that solve real-life problems. Students may also be asked to demonstrate learning by creating and producing a newscast, developing a museum exhibit on a specific topic, designing an efficient workflow for planning the prom, judging the efficiency of product manufacturing, carrying out pH tests of water samples, or carrying out similar tasks through which they will use acquired knowledge and skills.

Rubrics are often used to evaluate the quality of performance on tasks designated as authentic demonstrations of learning. This is consistent with performance assessments, but not all performance assessment is authentic. When the criteria for each achievement level is explicitly defined, the criteria will enable evaluation of student achievement for each learning category to be objective, consistent, and defensible. The intent of a rubric in authentic assessment is to guide students toward higher levels of achievement by providing these expectations as part of the instructions for the task.

Rubrics can also act as a guide to attain higher levels of achievement by engaging students in content and process, empowering task facilitation, contributing to synthesis of information to guide critical thinking and problem solving, and enabling the task to become a learning experience in and of itself. This also makes it likely that the assessment will be of increased interest for students, thus motivating higher levels of achievement.

Comparing Traditional to Authentic Assessment

Traditional assessment refers to forced-choice measures such as multiple-choice tests, fill in the blanks, and true-false tests for which students select an answer or recall information to complete the assessment. This type of assessment can be standardized or teacher created, administered locally or broadly. Traditional assessment often defines learning as recalling a body of knowledge and the demonstration of prescribed skills, such as working a mathematical problem without the students knowing where they could use it. These assessments are usually developed and administered to determine acquisition of knowledge as defined in a particular curriculum. Wiggins, the long-time advocate for authenticity in assessment, has emphasized, however, that traditional “inauthentic” assessment is not necessarily bad or invalid. As with all student assessment, any approach can work if there is consistency between objectives, instruction, and format.

As discussed earlier, authentic assessment requires students to synthesize and apply knowledge and skills through tasks that replicate, as close as possible, the challenges faced in real-life situations beyond the classroom. Student learning is assessed by the extent to which students demonstrate their mastery of knowledge and skill application through the particular authentic task. It is typical when authentic assessments are implemented, curriculum is designed around applied experiences with skills and knowledge taught and developed through a variety of tasks.

A teacher does not need to choose between authentic assessment and traditional assessment because they complement each other. Both types of assessment have various forms, and there is no bright line separating the two types, but traditional assessments tend to confirm recall, demonstration of skills, and connection between the two, while authentic assessment demonstrates the ways a student can apply the knowledge and skills while being assessed on attributes exhibited in the process and completion of a task.

The reason many teachers use authentic methods of assessment in addition to traditional means is because of two beliefs: That students must be prepared to do more than recall information; and the skills demonstrated in authentic assessment tasks will better prepare them for their future endeavors. Teachers also utilize authentic assessment methods because they enhance student engagement with the assessment process, engaging learners through active participation and interaction with the educational material, activities, and related community. Contemporary theories of learning suggest that when the activity in a learning environment is recognized by the students as worthwhile and meaningful, requires thoughtful creativity in solving a problem, and could have purposeful impact beyond the assignment, then the possibility of engagement with their learning is enhanced. These components are seen as strong characteristics of authentic assessment.

The judgment in scoring student learning in authentic assessment is defined through the descriptors in the scoring device based upon realistic expectations of process and product related to the designated task. Achievement expectations often allow for creativity and innovation in student response to a task, replicating or simulating the contexts in which the same proficiencies are demonstrated in one’s workplace, community, or personal life. Assessment tasks should be designed to challenge students to efficiently and effectively use a variety of designed to challenge students to efficiently and effectively use a variety of skills and draw from multidisciplinary knowledge to negotiate complex challenges. An important element of any assessment format, especially authentic assessment, is that students require appropriate opportunities to experience and practice a task, consult resources, and get feedback to refine their skills with whatever format of assessment is used. Poorly designed assessments may result in low achievement because of discomfort with the assessment format itself.

Authentic assessment tasks and the related scoring devices may require more time and effort on an instructor’s part to develop. Such assessments are not necessarily more difficult to grade, but this is dependent upon the quality of the scoring device, often being a rubric. To confirm ease of grading for authentic assessments, descriptors of traits must be sufficiently specific to differentiate qualities of achievement criteria to be judged.

Scoring authentic performance through well-designed tasks does not automatically ensure that the result validly represents learning. A measure will not be valid if it does not effectively address the learning outcomes it was designed to assess. The foundation of any good assessment builds from meaningful learning outcomes and clear expectations defining the quality of student achievement. Alignment between learning outcomes and the scoring device is essential.

Authentic assessments build upon an understanding that students construct their own meaning using information taught and gathered from other experiences with the world. It is this belief that supports assessments beyond information recall. Authentic assessment tasks allow students to demonstrate the extent to which they accurately construct meaning about what they have been taught and provide them the opportunity to engage in further construction of meaning. They also provide multiple, sometimes alternative ways for students to demonstrate what they have learned. When blended with traditional assessments, the multiple perspectives of student learning create a far more complete understanding of what students know and can do and how they can apply their acquired skills and knowledge in authentic situations.

Frederick Burrack

See also Achievement Tests; Formative Assessment; Performance-Based Assessment; Rubrics; Standardized Tests; Standards-Based Assessment; Summative Assessment

Further Readings

Bransford, J. D., Brown, A. L., & Cocking, R. (Eds.). (2000). How people learn: Brain, mind, experience and school (Expanded ed.). Washington, DC: National Academy Press.

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining authentic classroom assessment. Practical Assessment, Research … Evaluation, 17(2), 2.

Guskey, T. R. (2003). How classroom assessments improve learning. Educational Leadership, 60, 6–11.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). Classroom instruction that works: Research-based strategies for increasing student achievement. Alexandria, VA: ASCD.

Wiggins, G. (1998). Ensuring authentic performance. In Educative assessment: Designing assessments to inform and improve student performance (Chap. 2, pp. 21–42). San Francisco, CA: Jossey-Bass.

Wiggins, G. P. (1993). Assessing student performance (p. 229). San Francisco, CA: Jossey-Bass.

Gregory J. Marchant Gregory J. Marchant Marchant, Gregory J.

Authorship

Authorship

151

152

Authorship

This entry discusses the guidelines and professional norms determining who is credited as an author for a piece of academic writing. Whether the manuscript produced is a book, conference paper, technical report, or research article, the names on the final product and their order should reflect the relative contribution of those involved.

Number of Authors

Different professional societies and publications have different guidelines for who should be considered an author, but in general, a person whose involvement was "substantial" is included as an author, in that the manuscript would not have been produced without the person's contribution. For example, conducting a literature search for related materials usually would not merit authorship, whereas writing the literature review would. Proofreading a manuscript before submission would not merit authorship, whereas making revisions based on reviewer recommendations would.

Products requiring extensive time or work, such as books, longitudinal studies, and lengthy reviews, may merit more authors. Some journals charge fees for number of pages published or open-access processing that can be several thousand dollars. Sometimes a large number of co-authors are included to share not only the publication credit but also the cost; however, it is considered inappropriate to recognize a person for authorship merely to defray the cost of publication.

Order of Authors

Authorship is the public affirmation of the relative contribution of those involved in the creation of the manuscript. In the social sciences, generally, the person doing most of the writing is listed as the first author regardless of other contributions. In some fields, such as some of the natural sciences, the last author is assumed to be the most important. The student is almost always the first author on any publications resulting from a dissertation. Involvement in the project and order of authorship should be discussed and established early. However, order of authorship may be modified as the project progresses and involvement shifts.

Issues in Authorship

Although sole and first authorship are usually valued most, inappropriate practices can be a concern. Sole authorships are thought to be more likely to be fraudulent. Some demand first authorship regardless of degree of involvement. Some demand authorship for sharing a database or a minor contribution even if they are not involved with the study or manuscript. Adding a department chair or committee chair as an author for a work he or she was not directly involved with is inappropriate. Authorship is not a gift; it is a valued recognition of scholarly work.

Gregory J. Marchant

See also APA Format; Literature Review

Further Readings

American Psychological Association. (n.d.). Tips for determining authorship credit. Retrieved from http://www.apa.org/science/leadership/students/authorship-paper.aspx

Eisner, R., Vasgird, D. R., & Hyman-Browne, E. (n.d.). Responsible authorship and peer review (course module). Columbia University Office for Responsible Conduct of Research. Retrieved from http://ccnmtl.columbia.edu/projects/rcr/rcr_authorship/

The Office of Research Integrity. (2013, October). Authorship and publication (ORI introduction to RCR: Chapter 9). Retrieved from http://ori.hhs.gov/Chapter-9-Authorship-and-Publication-Introduction

Nicholas W. Gelbar Nicholas W. Gelbar Gelbar, Nicholas W.

Autism Spectrum Disorder Autism spectrum disorder

152

155

Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder, whose primary features are deficits in social communication and the presence of restricted interests and/or repetitive behaviors. In terms of social communication, individuals with ASD most commonly have difficulty with the pragmatic aspects of language, although some have difficulty using speech for communicative purposes (i.e., they are nonverbal or use few words). This entry further discusses the characteristics of individuals with ASD, the diagnostic standards for ASD and increased rates of diagnosis, the reasons why ASD is considered a complex disorder, and challenges with conducting and interpreting research on individuals with ASD.

The difficulties with the pragmatic aspects of language experienced by people with ASD involve the nonverbal cues present in language such as changes in pitch/tone that indicate emphasis or uncertainty on the part of the speaker. In addition, individuals with ASD have difficulty recognizing body language and facial expressions. These individuals may speak with a flattened aspect and tend to have difficulty understanding humor and sarcasm. They are often very literal in their use and comprehension of language. The difficulties with the pragmatic aspects of language lead to difficulties with peers, as individuals with ASD have difficulty engaging in social and play activities.

Additionally, these individuals often have a highly specialized area of interest. This restriction of interests can also interfere in their ability to develop and maintain peer relationships. Some individuals with ASD also engage in repetitive behaviors. One example is echolalia, which is the frequent repetition of vocalization for noncommunicative purposes. Some individuals will also engage in repetitive motor movements such as hand flapping, body rocking, or head banging. These verbal and motor stereotypies also interfere with these head banging. These verbal and motor stereotypies also interfere with these individuals ability to engage socially with others.

These behaviors can be detected as early as 18 months, and ASD is typically diagnosed in young children between the ages of 2 and 4 years. It affects approximately 1% of the population. However, recently more children have received the diagnosis. The screening and diagnostic standards have evolved and improved, which may explain increases in incidence in recent years. Many adolescents are being diagnosed, as they were not screened as children.

To complicate matters, the diagnostic standards for ASD have also shifted over time, which may explain increased rates of diagnosis. These increased rates may also reflect greater awareness and screening for the disorder in schools and medical settings. Previous clinical subtypes of the disorder (e.g., Asperger’s syndrome) are no longer recognized by the current nosology (system for classifying psychiatric diagnoses). There are also many traits and behaviors that are associated features with ASD but are not diagnostic symptoms. These include behavioral challenges, heightened anxiety, and deficits in executive function. As noted, some individuals with ASD are also delayed and have resulting impairments in their language abilities. In addition, individuals with ASD may have poorly developed adaptive skills such as self-care, cooking, and personal finance to manage day-to-day life.

Complexity of ASD

ASD is a complex disorder of unknown etiology. Research involving genetic and environmental causes of the disorder is still ongoing and has yet to reach any significant findings. The neuropsychology of ASD is also still evolving, and as of 2016, many brain areas had been implicated as potential sources for the unique presentation of the disorder. However, no single brain area has been able to be isolated to explain the symptom presentation across the entire spectrum.

ASD is referred to as a spectrum because individual manifestation of the symptoms varies widely. Individuals range from those who do not communicate verbally and have significant intellectual limitations to individuals who have fully developed language and intellectual abilities. These individuals may have intense interests in a narrow area of focus as well as having difficulty with initiating and maintaining social relationships. Academic difficulties, repetitive behaviors, or sensory sensitivities may or may not be present.

The developmental trajectory of individuals with ASD also varies across the life span. Generally, symptoms become less severe as individuals age. However, greater difficulties during adolescence have been noted in some samples. Also, while the symptoms may reduce over time, the functional consequences of the remaining symptoms are often greater. In other words, while the symptoms may partially remit, the individual often experiences increasing social difficulties as the social expectations become greater as they age. These difficulties can lead to issues with anxiety, depression, and even social withdrawal. In addition, the adult outcomes for individuals with ASD across the spectrum are troubling as many individuals do not live independently and are not engaged in full-time employment.

Research Challenges

There are several challenges with conducting and interpreting research for individuals with ASD. The first challenge with ASD research is defining the population, as there are several issues to contend with in this area. The first is that ASD is defined differently in the medical system and in the education system. The second is that the conception of ASD including subtypes as previously noted has changed over time. The third issue in this area is how best to provide documentation of ASD for a sample.

In terms of the first issue, the medical world relies on diagnoses, whereas the education system has criteria that define a disability relative to educational performance. As such, some research conducted in education settings using special education students will have a different subset of the population. Simply put, some children who would qualify under the medical diagnosis may not qualify under the educational diagnosis and vice versa. This means that there are differences in the populations that are studied. To further complicate matters, children with medical diagnoses of ASD may be classified under other education categories such as intellectual disability, multiply disabled, or other health impairment. In addition, some children with medical diagnoses of ASD will not qualify for special education services.

As mentioned, the second issue, comparing research on ASD over time, is due to the fact the diagnostic criteria have changed over time. While the criteria have evolved from the original concepts of Leo Kanner and Hans Asperger in the 1940s, the recent shifts are more germane to contemporary research. Under the previous nosology, autism used to be categorized as one of the pervasive developmental disorders and the term *autism* referred to more impacted individuals who demonstrated the unique signs of the disorder including language delays and repetitive behaviors (now often referred to as “classical autism”).

Individuals with at least average intellectual ability and no history of language delays were classified as having Asperger’s syndrome. These individuals also had to have evidence of a highly restricted area of interest. Mild symptom presentations were classified as pervasive developmental disorder, not otherwise specified. Other very rare subtypes such as childhood disintegrative disorder and Rett syndrome also had specific criteria. In the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*), published in 2013, the category of pervasive developmental disorder has been changed to ASD and the classification of subtypes has been altered. The diagnostic labels of Asperger’s syndrome and pervasive developmental disorder, not otherwise specified have been removed, and the rare disorders are now considered special variants of ASD. Sensory sensitivities were also added as a diagnostic symptom under the new classification.

Beyond the differences between the medical and education criteria and the evolving diagnostic standards, the final related issue involves how studies confirm the diagnosis of ASD when defining their study population. Clinical judgment is considered the best tool for making the diagnosis, but the Autism Diagnostic Observation Schedule–Second Edition is the most common tool for gathering information on the presence of the disorder. The Autism Diagnostic Observation Schedule–Second Edition is not a standardized instrument but is a structured observational tool that measures social communication and the presence of sensory interests and repetitive behaviors. There are several modules that are tailored with specific activities designed for specific ages and language abilities.

The Autism Diagnostic Interview Schedule–Revised is a structured interview for gathering information on the development of the individual. When used with the Autism Diagnostic Observation Schedule–Second Edition, this combination of assessments is considered the gold standard for documenting the presence of ASD, as this method has the most empirical support. Other rating scales are often used in the literature, but these rating scales can be best conceptualized as screeners as opposed to being diagnostic in nature.

Beyond the challenge of defining the ASD population in research, one of the further challenges for conducting and understanding research in this field is the heterogeneity of this population, as previously noted. This challenge cannot be understated, as individuals with ASD will vary from those who use augmentative communication to speak to those who were able to pursue advanced degrees.

Beyond having intellectual and communication differences, individuals with ASD will also vary in their skills in the areas of executive function, social skills, academic achievement, functional and adaptive abilities, anxiety, depression, and other psychiatric comorbidity. Further, individuals with ASD may also present with challenging behaviors such as verbal and physical aggression as well as passive noncompliance. These individuals will also have variety of strengths and interests that need to be considered when providing programming and interventions. This heterogeneity means that specifying the sample is critical in ASD research, as not all interventions will work for all individuals on the spectrum. In other words, having specific inclusion criteria (e.g., having ASD and concurrent anxiety) is important to match an intervention to the actual subpopulation of individuals with ASD on which it is expected to have an impact.

The final difficulty in conducting research on individuals with ASD is to gather large samples of these individuals. Part of this difficulty is due to the fact that there is often a need to specify a subset of the population and that ASD is a relatively low incidence disability. This can make it difficult to use traditional group-based research methods.

One way to continue to be able to use group-based methods is to limit the number of dependent variables and to tailor these variables to the intervention being conducted. Another approach is to use a class of research methods called single case designs in which one can use a small sample of individuals but utilize repeated measurement strategies. The conditions for the intervention and measurement are highly specified, and comparisons can be made across individuals or settings to demonstrate causal relationship (that the intervention is the only likely cause of the change in behavior).

Beyond the difficulties of defining and then recruiting samples, it is difficult to conduct research on individuals with ASD because they are a heterogeneous group and no one set of interventions or strategies works for all individuals on the spectrum. Nonetheless, research on individuals with ASD is growing at a rapid rate and further refinements to the disorder in terms of diagnostic and rapid rate and further refinements to the disorder in terms of diagnostic and associated features are likely to evolve.

Nicholas W. Gelbar

See also Adaptive Behavior Assessments; Applied Behavior Analysis; Attention-Deficit/Hyperactivity Disorder; *Diagnostic and Statistical Manual of Mental Disorders*; Intellectual Disability and Postsecondary Education; Single-Case Research

Further Readings

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5) (5th ed.). Arlington, VA: American Psychiatric Publishing.

Autism and Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators. (2014). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2010. Morbidity and Mortality Weekly Report. Surveillance Summaries, 63(2), 1.

Ch’ng, C., Kwok, W., Rogic, S., & Pavlidis, P. (2015). Meta-analysis of gene expression in autism spectrum disorder. Autism Research, 8(5), 593–608.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). Autism Diagnostic Observation Schedule, Second Edition (ADOS-2). Torrance, CA: Western Psychological Service.

Rutter, M., Le Couteur, A., & Lord, C. (2003). Autism Diagnostic Interview Revised (ADI-R). Los Angeles, CA: Western Psychological Services.

Alex McInturff Alex McInturff McInturff, Alex

Autocorrelation

Autocorrelation

155

157

Autocorrelation

Autocorrelation describes a set of data that is correlated with itself. When successive values ordered over time or space exhibit nonzero covariance, these data are said to be autocorrelated. Autocorrelation in time series data, often referred to as serial correlation, is frequently observed and has been widely studied and canonized. Examples are numerous: tomorrow’s temperature is often predicted by temperature today, and a county’s literacy rate next year will likely be well predicted by literacy this year.

While spatial autocorrelation remains an actively growing body of research, examples are also abundant: temperature in one county is often predicted by temperature in a neighboring county, and demographic makeup of a census block is likely similar to neighboring blocks. In both spatial and temporal data, autocorrelation has important implications for ordinary least squares regression, and other procedures that assume model errors are independent and uncorrelated. If not accounted for, the presence of autocorrelated errors can lead to misleading or invalid inference or imply model misspecification. Alternatively, autocorrelation can be marshaled for making predictions. In autocorrelated time series data, for example, future data points may be predictable because of their correlation with current and past values. This entry first explains how autocorrelation is quantified and discusses the importance of autocorrelation. It then looks at some of the nuances and implications of autocorrelation.

Quantifying Autocorrelation

Autocorrelation is easiest to demonstrate mathematically using time series data, where it is often represented in the form of a linear regression. The following equation shows the autoregressive (AR) model of autocorrelation, which shows the current observation (Y_t) as a linear function of previous observations plus a residual error:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots \beta_p Y_{t-p} + e_t.$$

Here is the familiar structure of an ordinary least squares regression with βs as constant coefficients, *Y* values at *p* previous intervals or *lags*, and a random error series e_t. This model assumes stationarity, such that the mean, variance, and covariance of the observations remain constant for all time periods. However, in many nonstationary cases, the residual error itself may be an autocorrelated series that requires more advanced AR moving average or AR integrated moving average models. First-order autocorrelation, in which an observation (Y_t) is correlated only with the observation immediately preceding it (Y_{t-1}), is commonly observed in time series data. Larger orders of autocorrelation include more lags and imply longer decays or greater inertia in temporal and spatial processes.

The most common way to measure the strength of autocorrelation is by computing the autocorrelation coefficient, symbolized as ρ. For the common case of the first-order autocorrelation, ρ is a two-variable correlation coefficient ranging between −1 and 1. Positive values of ρ are most frequently observed, indicating similarity between successive observations. Higher values of the coefficient indicate stronger dependency on previous values and milder decay. Although positive autocorrelation suggests that successive observations will move in the same direction, negative autocorrelation characterizes processes that oscillate in direction. A negative value of the autocorrelation coefficient can be exemplified by imagining the task of cutting equally sized pieces of ribbon into two unequal pieces. The smaller the first piece is cut, the larger the second piece will be. Thus, successive observations oscillate in size and are negatively correlated.

A sample first-order autocorrelation coefficient for time series data (r_t) can be calculated in similar fashion to a correlation coefficient, with *N* as the number of residuals in the time series data:

$$r_t = \frac{\sum_{t=2}^{N} (e_t)(e_{t-1})}{\sum_{t=1}^{N} e_t^2}.$$

As is clear from this equation, the total length of the time series is inconsequential, and only the *lag* between observations appears in this calculation. In addition to estimating the strength of correlation, this coefficient can be used as part of an inferential procedure to test the null hypothesis that there is no autocorrelation of residuals. In other words, the null hypothesis is that $\rho = 0$. For the first-order autocorrelation, the Durbin-Watson test is frequently employed for this purpose and generates a *p* value that can be used to determine whether the null hypothesis may be rejected or not.

A correlogram is often used to visually demonstrate and assess autocorrelation. The size of the autocorrelation coefficient is plotted on the vertical axis against increasing lag sizes on the horizontal axis. Autocorrelation coefficient values near zero on the correlogram indicate lag sizes at which correlation is weak, while large, distinct, or systematically occurring coefficient values occurring at one or more lag sizes suggest that there is dependency in the data.

Importance

Autocorrelation is a commonly occurring phenomenon, as many data change slowly over time and space. Data collection at small spatial or temporal intervals further increases the likelihood that successive observations will be correlated. Quantifying and parameterizing the strength and scope of autocorrelation as described earlier can be important in making predictions, and the aforementioned AR moving average and AR integrated moving average models excel in this regard. While the usefulness of autocorrelation in forecasting is clear, the importance of autocorrelation is more frequently framed around the problems of model inference and model specification.

Spatial and temporal data are often modeled using ordinary least squares regression or similar techniques that rely on the Gauss-Markov theorem assumption that error terms are uncorrelated. When autocorrelation is present in model residuals, this assumption is violated. Although parameter estimates remain unbiased, standard errors become biased—inflated when autocorrelation is positive and deflated when negative—and increase the chance of making Type I and Type II errors. Additionally, results of *t* and *f* tests will be invalid and confidence intervals incorrect. The Durbin-Watson test for time series data and Moran's *I* for spatial data are the most common tests researchers conduct on residuals to evaluate whether autocorrelation has the potential to invalidate inference. If autocorrelation is found, it may be corrected by transforming the data using a generalized least squares or other approach.

The approach just described assumes that the model has been correctly specified and exhibits "pure" autocorrelation. However, a misspecified model can also produce autocorrelated residuals, regardless of autocorrelation in the data. This is common when a spatially or temporally dependent variable is excluded from a regression model, as those dependencies will now occur only in the error term of the model. Spatial and temporal lags of variables should be included to appropriately model these processes. Similarly common is the use of an incorrect functional form, such as using a linear model when a quadratic form is more appropriate, which often produces autocorrelated residuals.

Autocorrelation is thus common both as a real phenomenon and an artifact of modeling. Appropriate dynamic model specification is critical, and correctly dealing with autocorrelation allows for validity of inference, more appropriate analysis, and improved precision.

Additional Concepts

Research on autocorrelation is ongoing, and many nuances cannot be covered here. However, a few additional concepts deserve mention. Much of this entry describes the first-order autocorrelation. Although it is most common, higher order autocorrelation does occur, even when adjacent values show no relationship. Higher order correlation coefficients are represented as r_k, where k is the order of autocorrelation being assessed. Taken all together, the set of (r_1, r_2, …, r_k) coefficients is the autocorrelation function.

For stationary data, autocorrelation is likely to decay quickly, maintaining a consistent mean, variance, and covariance over time. However, if data are nonstationary, the effects of the past may accumulate rather than dissipate, maintaining long "memories." Integrated processes, including random walks, never discount the effects of stochasticity and may not ever return to a mean, while fractionally integrated processes decay very slowly. As a result, such processes can produce spurious inference if nonstationarity is not accounted for.

Finally as larger and larger data sets become available, an active area of research deals with the implications of autocorrelated data on sample size. Depending on the strength of autocorrelation, more data produced via sampling processes at a higher resolution do not necessarily mean more information. A tension thus exists between sample size, resolution, and the number of independent measurements. New methods for determining effective sample size and for conducting regressions between two autocorrelated data sets continue to be developed.

Alex McInturff

See also Correlation; Pearson Correlation Coefficient; Phi Correlation Coefficient

Further Readings

Box, G. E., & Jenkins, G. M. (1976). Time series analysis: forecasting and control (Rev. ed.). Holden-Day.

Clifford, P., Richardson, S., & Hemon, D. (1989). Assessing the significance of the correlation between two spatial processes. Biometrics, 123–134.

Diniz-Filho, J. A. F., Bini, L. M., & Hawkins, B. A. (2003). Spatial autocorrelation and red herrings in geographical ecology. Global Ecology and Biogeography, 12(1), 53–64.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. Biometrika, 37(3/4), 409–428.

Enders, W. (2008). Applied econometric time series. Wiley.

Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics.

Journal of econometrics, 2(2), 111–120.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? Ecology, 74(6), 1659–1673.

Viladomat, J., Mazumder, R., McInturff, A., McCauley, D. J., & Hastie, T. (2014). Assessing the significance of global and local correlations under spatial autocorrelation: A nonparametric approach. Biometrics, 70(2), 409–418.

Mark D. Shermis Mark D. Shermis Shermis, Mark D.

Sue Lottridge Sue Lottridge Lottridge, Sue

Automated Essay Evaluation Automated essay evaluation

157

160

Automated Essay Evaluation

Automated essay evaluation (AEE) is the evaluation of written work through computer technology. AEE is an expansion of the earlier concepts of automated essay scoring (AES) and automated essay grading in that in addition to providing a numerical index of writing performance, the technology can provide qualitative feedback that can be used in formative writing applications. Although originally created for English, the technology has been expanded to other languages as well. This entry further describes AEE and how it is used in both summative and formative evaluations.

The web-based technology can be used in both summative and formative applications. AEE is used for both high-stakes tests and as part of an electronic portfolio system to facilitate the teaching of writing. AEE was originally developed by Ellis Page with the objective of making the grading of essays a bit easier by automating the evaluation of at least the mechanical aspects of writing. The hope is that, by reducing scoring costs, the technology can facilitate the writing performance items that are part of assessments created as a result of federal testing policies.

Numeric feedback, usually embedded in a rubric, is often given on a scale of 1–6 with 6 indicating *best performance* and a 1 indicating *poorest scorable performance*. In addition, the computer may be able to identify errors in grammar, mechanics, syntax, and organization and development. Some AEE programs can identify structure in narrative essays and can provide feedback along the lines of,

> The thesis of your paper appears to be that “democracy can only flourish when all citizens have the right to vote.” You appear to be making three points—democracy is one five different forms of government, voting is a key element in the formation of a democracy, and governments have limited the rights of some individuals to vote. You have a lot of information about the first point but not too much information for the second and third points.

Note that the computer does not “understand” what is written; rather, scores and feedback are based on models of human rater performance. If humans reward a certain pattern of writing, then the program will attempt to apply the same reward to similar patterns of production. Additionally, the software programs can give an assessment of how “on topic” the writer is, can generally flag essays that include bizarre or threatening content, and can flag essays that are statistically unusual (e.g., very long essays).

Two types of model construction are employed for AEE—prompt-specific models and nonprompt-specific models. Prompt specific-models are used when the scoring of content plays an important role in the evaluation of the essay. As an example of the methods used to create and validate scoring algorithms, here are the procedures that might be used to create a model for the commonly used six writing trait rubric: Six hundred essays are collected and scored by two human raters from a sample of examinees that are tested under conditions similar to the operational environment. Essay scores are typically sparse at the high score points and so training samples often have more essays at the lower score points and often very few essays at the upper score points (sometimes fewer than 20). Typically, a large proportion—66%—of the sample is used for model building, with the remainder held out as a separate cross-validation sample to evaluate model performance.

In the case of the 600 essays, 400 essays are randomly selected for model building. The essays are parsed and tagged, and variables (or consolidated meta-variables) classified by the parser are regressed against the scores of the human raters. Once the model is built and finalized, the remaining 33% (or 200 in this case) essays are scored by the model as part of a cross-validation procedure. The fit of the validation set is usually not as good as it was for the training set, as the models tend to overfit the training sample. Consequently, the regression weights associated with the model are adjusted to reflect the prediction inaccuracy on the validation set. Usually the adjustment is made by using all of the essays in the validation set. Usually the adjustment is made by using all of the essays in the data set for forming the model.

Nonprompt-specific models evaluate good writing characteristics in an essay and not the content of what is written. These models are used when similar, but not the same, questions are asked of a narrowly stratified population (e.g., candidates taking a certification exam). The goal is to determine the writing ability of the individuals and not the mastery of a content domain. Model building is the same as for prompt-specific models except that the training and validation sets draw across multiple prompts.

An extension of the nonprompt-specific model is referred to as the “generic” model that is constructed when the goal is to evaluate across multiple prompts for one or more genre of writing. Again, the goal is to provide feedback on general writing characteristics. Generic models may be configured for a specific genre or developmental level of writing, but the attempt is to provide broader coverage than for other nonprompt-specific models. They are generally used for formative feedback because their estimates tend not to be as precise as for prompt-specific models. Model building is also generated from a much broader set of prompts within each genre.

Summative and Formative Use of AEE

AEE has been evaluated in both summative and formative contexts. The most comprehensive evaluation of the technology was part of a Hewlett Foundation–sponsored demonstration that contrasted the AES performance of eight commercial vendors and one university laboratory’s performance on AES with that of human raters. That study employed eight different essay sets drawn from six states representing the Partnership for Assessment of Readiness for College and Careers and Smarter Balanced Assessment Consortium. Four of the essays were “source based.” A student was asked to read an artifact and then respond with an essay. The remaining four essays reflected prompts of a more traditional variety (i.e., narrative, descriptive, persuasive).

Over 17,000 essays were randomly divided into two sets, a training set (*n* = 13,336) in which the vendors had 1 month to model the data and a test set for which they were required to make score predictions within a 2½-day period. The training set consisted of two human rater scores, the so-called resolved score, and the text of the essay. The resolved score reflected the final score assigned by the state. The test set consisted only of the text of the essay. Six of the eight essays were transcribed from handwritten documents using one of the two transcription services. Accuracy rates of transcription were estimated to be over 98%. The challenge to the nine teams was to predict scores based on essay performance that matched the ratings assigned as the resolved score.

Performance was evaluated on five different measures of a single evidentiary criterion—agreement with human raters. One set of measures focused on agreement at the distributional level and the other set on agreement at the individual response level. The individual-response-level measures included exact agreement, exact + adjacent agreement, κ, quadratic weighted κ, and the Pearson product–moment correlation. The AES engines performed well on the distributional measures. With a high degree of consistency, all nine demonstrators were able to replicate the means and standard deviations for the scores assigned by the states.

With regard to agreement measures, there was some variability, but the AES engines performed well on three of the five measures (exact + adjacent agreement, quadratic weighted κ, and correlation). On the two measures where the performance was not as high (exact agreement and κ), there was also high variability among the performance of operational human raters. In addition, scaling artifacts attributable to the way the state scores were adjudicated may have contributed to the relative lack of precision on predicted scores.

A follow-up study was performed as a competition with data scientists using the same (but anonymized) data sets. Over 200 competitors participated over a 3-month development period for a top prize of US$60,000. The course of the competition included a development forum in which participants helped one another by disseminating the results of their programming experiments. The top team came in with a quadratic weighted κ coefficient of .78.

Formative use of AEE has been in place for several years, but comprehensive evaluations of the technology are few. Mark Shermis, Jill Burstein, and Leonard Bliss looked at the impact of the Educational Testing Service product criterion in the writing outcomes of a 10th-grade English class at an urban high school in Miami, FL. After seven writing assignments, the researchers found that students produced fewer errors in writing and showed some evidence of growth.

A study on the use of AEE in a K–12 formative program showed performance across a range of grades within and across traits in a formative system. AEE was implemented in an online practice assessment program in a Southern state, using the CRASE platform, automated scoring proprietary software by Pacific Metrics Corporation. Responses were scored on a three-trait rubric (ideas, style, and conventions) with a score range of 1–4. The responses were scored by humans, and the CRASE engine was trained on these responses. The software program performed similar to humans across the grades and traits.

Averaged across the grades, mean scores were similar for CRASE and humans, with composite scores showing slightly larger, but still similar, differences. More important, though, are agreement rates. Exact agreements are influenced by score distributions and the rubric scale. Computer and humans agreed exactly 70–80% of the time with generally lower agreement rates on the conventions trait. Correlations between the software and manual scoring were around .70 for the traits and .80 for total scores.

There is more work to be done in the area of AEE, but the technology has demonstrated the capacity to accurately score in the contexts of formative and summative assessment. Predictions seem to be better for essays than for short-form constructed responses, but even with the latter, technology improvements are being made on a regular basis. Specific challenges for the technology include recognition and better assessments of arguments, making predictions about the reasonableness of conclusions and providing more targeted feedback on the nonmechanical aspects of writing.

Mark D. Shermis and Sue Lottridge

See also Formative Evaluation; Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium; Summative Evaluation

Further Readings

Shermis, M. D. (2014). State-of-the-art automated essay scoring: A United States demonstration and competition, results, and future directions. Assessing Writing, 20, 53–76.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. Educational Assessment, 20(1), 46–65.

Shermis, M. D., & Burstein, J. (2003). Automated essay scoring: A cross-disciplinary perspective (pp. xvi–238). Mahwah, NJ: Erlbaum.

Shermis, M. D., Burstein, J., & Bliss, L. (2004). The impact of automated essay scoring on high stakes writing assessments. Paper presented at the annual meetings of the National Council on Measurement in Education, San Diego, CA.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 1–15). New York, NY: Routledge.

Shermis, M. D., Burstein, J. C., Elliot, N., Miel, S., & Foltz, P. W. (2015). Instructional applications for automated writing evaluation. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.), Handbook of writing research (2nd ed., pp. 395–409). New York, NY: The Guilford Press.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Paper presented at the Annual National Council on Measurement in Education.

Shermis, M. D., & Hamner, B. (2013). Handbook of automated essay evaluation: Current applications and new directions. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 313–346). New York, NY: Routledge.

Shermis, M. D., & Morgan, J. (2015). Using prizes to facilitate change in educational assessment. In F. Drasgow (Ed.), Technology in testing: Measurement Issues (pp. 323–338). New York, NY: Psychology Press.

Vantage Learning. (2001). A preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments. Newtown, PA: Author.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated essay scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.

Wilson, J., & Andrada, G. N. (n.d.). Using automated feedback to improve writing quality. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), Handbook of research on computational tools for real-world skill development (pp. 678–703). Hershey, PA: Learning Disabilities: A Contemporary Journal.

B

Scott Bishop Scott Bishop Bishop, Scott

b Parameter *b* parameter

161

163

b Parameter

Item response theory (IRT) uses several parameters. These parameters control the position and shape of IRT item characteristic curves (ICCs) that map the relationship between an examinee's ability, usually denoted by theta (θ), and the examinee's probability of making a particular item response (see Figure 1). The *b* parameter is present in all common IRT models. It is called the *difficulty* parameter, as it is the IRT analogy to traditional measures of item difficulty, such as *p* values. This entry first provides further context and discusses the importance of the *b* parameter. It then discusses unidimensional IRT models for dichotomous items and polytomous items, multidimensional IRT (MIRT) models, and the relationship between *p* values and IRT *b* parameters.

Context

IRT models the probability that an examinee will make a particular response to an item based on examinee's standing on the trait that a test measures. For a mathematics achievement test, IRT can model the probability that an examinee will earn a particular score on an item based on the examinee's achievement in math. For an instrument measuring extroversion, IRT can model the probability that an examinee will select the response *very accurate* for the statement *I make friends easily* based on the examinee's degree of extroversion. Virtually any type of instrument (e.g., ability and achievement tests, personality and attitude assessments, and questionnaires and surveys) and any type of item (e.g., true or false, multiple choice, Likert) can be analyzed using IRT.

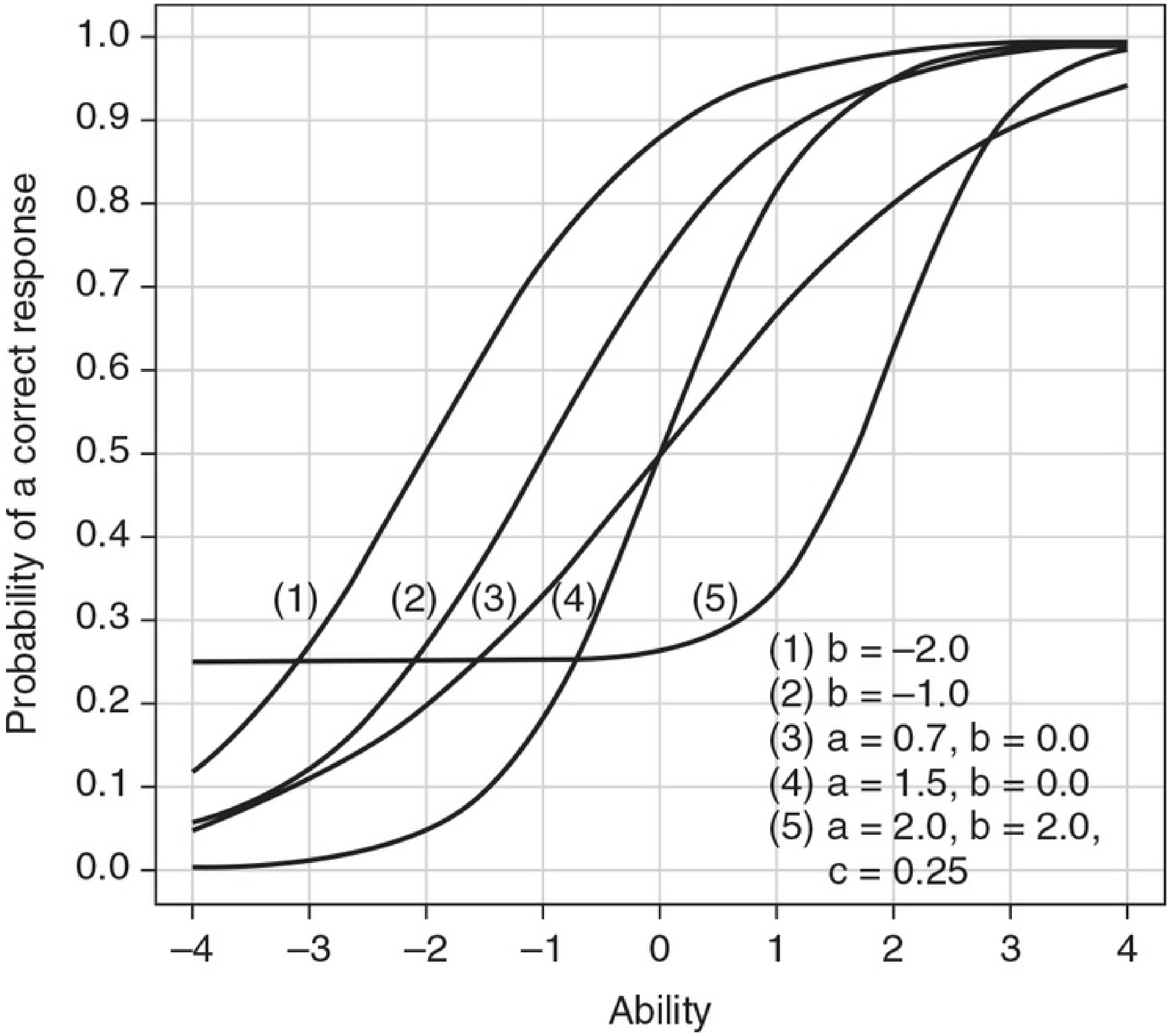
Importance of *b* Parameter

There are many practical benefits to the *b* parameter relative to classical measures of item difficulty. First, *b* parameters are not group dependent (provided the IRT model fits the data). Another advantage is that item difficulty, *b*, and examinee ability, θ, are on the same scale. This, combined with the fact that an item's maximum information occurs at or near the *b* parameter, means that inspection of the *b* parameters is helpful in the test construction process.

Unidimensional IRT Models for Dichotomous Items

The most common IRT models assume that (a) a single ability underlies the examinees' response processes and (b) items are dichotomously scored (e.g., right vs. wrong). The relationship between the probability of a correct response and examinee ability has a monotonically increasing ICC that is roughly *s* shaped, although it can be compressed and stretched at various points along the ability scale (see Figure 1).

Figure 1 Item characteristic curves for five hypothetical items



The *b* parameter describes the ICC's location on the θ scale. Specifically, it shows where there is an inflection point on the θ scale (i.e., where the concavity of the curve changes) in the ICC.

The three-parameter logistic (3PL) model has:

1. a discrimination parameter, denoted *a*, that controls the slope of the ICC;
2. a pseudo-guessing parameter, denoted *c*, that represents the lower asymptote of the ICC; and
3. a difficulty parameter, denoted *b*.

The two-parameter logistic (2PL) model does not have the pseudo-guessing parameter, *c* (or equivalently, one can consider *c* = 0). The one-parameter logistic (1PL) model also does not have the pseudo-guessing parameter, *c*, and all items are assigned a common slope, meaning items are modeled with only the *b* parameter.

In the 3PL model, the *b* parameter is located where the probability of a correct response is (1 + *c*)/2 on the θ scale. The 1PL model and 2PL model have no *c* parameter, so the *b* parameter's location on the θ scale is where the probability of a correct response is 1/2. Figure 1 illustrates ICCs for 2 1PL items, 2 2PL items, and 1 3PL item. The *b* parameters for Items 1– 4 are at −2.0, −1.0, 0.0, and 0.0, respectively. The probability of a correct response for these items at those θ values is 0.50. Item 5, the 3PL item, has a probability of a correct response of (1 + 0.25)/2 = 0.625 at a θ of 2.0. Note that items with higher *b* parameters have ICCs that are shifted to the right along the θ scale.

Unidimensional IRT Models for Polytomous Items

Use is increasing for unidimensional IRT models for examinee response processes for items that are polytomously scored. The function of the *b* parameter depends on the exact polytomous model in question. The generalized partial credit model models response categories that are adjacent (e.g., 0 vs. 1, 1 vs. 2, 2 vs. 3). The *b* parameter represents the overall difficulty of the test item, and there are separate difficulties for each response category (denoted by d_s, where *s* is a particular item score). The graded response model models multiple dichotomous outcomes, where the examinee scores in a particular response category or any higher categories versus the examinee scoring in any lower response categories (e.g., 0 vs. 1, 2 and 3; 0 and 1 vs. 2 and 3; and 0, 1, and 2 vs. 3). Both the GPCM and GRM are used to show ordered response categories. The nominal model is used with unordered response categories, and it also has a difficulty parameter.

MIRT Models

The *b* parameter also occurs in MIRT. Although there are as many *a* parameters as there are dimensions of the MIRT model, there is only one *b* parameter, and the ICC becomes a surface instead of a curve. Strictly speaking, in MIRT models, the *b* parameter is related to the multidimensional difficulty of the item, which rescales the *b* parameter by dividing it by the square root of the sum of the squared *a* parameters.

Relationship Between *p* Values and IRT *b* Parameters

Allan Birnbaum and Frederic Lord noted that under certain conditions, an approximate relationship exists between the IRT *b* parameter and classical *p* values. The relationship is only approximate for *b* parameters under the 2PL model, so very little or no guessing by examinees should be expected. The 2PL *b* parameter is approximately equal to γ/*r*, where γ is the negation of the inverse cumulative normal distribution deviate corresponding to the item’s *p* value and r_{biserial} is the item’s biserial correlation.

Unlike *p* values, where lower values indicate harder items, *b* parameters that are higher in value indicate harder items that fewer examinees endorse or answer correctly. Only examinees high in ability will have a moderate to high probability of answering these items correctly. Lower *b* values indicate easier items that many examinees endorse or answer correctly. Whenever an examinee’s ability equals an item’s *b* parameter, the examinee will have 0.50 probability of answering the item correctly for 1PL and 2PL models.

Scott Bishop

See also *a* Parameter; *c* Parameter; Item Response Theory; *p* Value

Further Readings

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29–51.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159–176.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. (Psychometrika Monograph, No. 17). Richmond, VA: Psychometric Society.

Rebecca Mazur Rebecca Mazur Mazur, Rebecca

Backward Design

Backward design

163

168

Backward Design

The term *backward design* refers to an approach to schooling, both at the system level and classroom level, predicated on a tight focus on achieving predetermined mission-related goals. Backward design for education was largely created and popularized by Grant Wiggins and Jay McTighe as an alternative to traditional design approaches wherein teachers typically began with content (such as textbooks, novels, or standards) and then created instruction around the ideas or questions that arose from selected materials. The backward design approach asks teachers and school leaders to determine skills, ideas, understandings, and dispositions most critical for students and then build learning experiences that ensure those outcomes. This entry discusses the application of backward design to curriculum design, schools and school systems, system-wide programming, and evaluation.

Although its applications are varied, backward design theory and technique is most typically used by teachers to develop lesson and unit plans and by educational leaders to improve system-or building-wide curricula. The most foundational principle of backward design is that educators must allow their work to be guided by jointly established goals (either for student understanding or for school/district improvement) and assessed using authentic performance assessments that generate acceptable evidence. The popularity of backward design increased dramatically through the first decade or so of the 21st century and it is commonly taught as part of educator preparation programs. Backward design is a valuable framework for educational evaluation, as it helps to operationalize key variables involved in school, program, or curriculum design.

Backward Curriculum Design

Backward Curriculum Design

Backward curriculum design is sometimes also termed *understanding by design* after the book of the same name in which it is outlined in detail. It is not a prescriptive system of curriculum development, but rather is intended to be a way of thinking about instruction that keeps student understanding of essential concepts and ideas at the heart of schooling. Backward curriculum design may improve instruction by giving teachers a framework that encourages a focus on student growth rather than on the process of teaching and that avoids two common pitfalls of classroom instruction: teaching that is focused on activities or busy work and teaching that is focused on covering some quantity of content, such as a chapter in a textbook.

In the broadest terms, backward curriculum design is usually approached as a three-phase process. First, a teacher (or group of teachers) identifies desired results. Second, a determination is made about what will constitute acceptable evidence of learning. Third, the learning experience is planned in detail.

Identifying Desired Results

A frequently used analogy likens backward curriculum design to vacation planning; before planning the particulars of a trip (e.g., plane tickets, hotel bookings, excursions), a traveler must first decide on a destination. Similarly, when planning a learning experience, teachers who use backward design principles decide at the outset on what results they hope a lesson or unit will achieve. In other words, they decide what understandings, knowledge, and skills students will gain or enhance by participating in the experience.

Identifying such big ideas, however, can be challenging. Practitioners of backward curriculum design often begin by sorting their materials into three categories of importance: those ideas, facts, and skills that are essential, those that are important, and those that are worthwhile. Although the definitions of these categories are fluid and must be determined by each practitioner’s assessment of learning needs of the practitioner’s students, the basic guidelines are as follows:

Essential Ideas and Skills

Those enduring, transferable ideas and skills that are critical for navigating the world. These often require significant investments of time and effort.

Important Knowledge and Skills

Discrete facts, skills, and techniques that are broadly useful in school and in life, such as solving mathematical equations or crafting thesis statements. These can usually be learned with modest amounts of instruction and effort.

Worthwhile

Key dates, facts, figures, and terminology. These are often easily memorized facts that can be accessed through textbooks or ready reference sources such as printed or online encyclopedias.

Traditionally, schooling has focused on the latter two categories and given them priority over essential ideas and skills. Backward curriculum design, conversely, is driven by the first category and uses the important and worthwhile facts and skills in the service of those that are deemed essential.

Enduring Understandings and Essential Questions

Practitioners of backward design often frame their work by coupling enduring understandings with essential questions that are implied by the big ideas and skills that have been identified. Essential questions are those that inspire students to think about the big ideas that the teacher has determined to be critical. For example, the teacher whose students are studying the great depression might have posed the essential question, “What happens when a government’s responsibilities to its individual people come into conflict with responsibilities to its industries?” Or, “Which group deserves more protection from the government: Individuals or industry?”

Essential questions are meant to provoke deeper thought, spark debate, and inspire more questions. They most often do not have “right” answers, but rather they can be approached from a variety of angles. They should intrigue students, they should be likely to have application outside the classroom, they should have relevance to core ideas inherent to the discipline, and they should help students make sense of abstract, multifaceted concepts. Furthermore, essential questions can be general, such as the two posed above, or content-specific, such as, “Were the government’s actions during the Depression fair or unfair?”

Once determined, enduring understandings (there are typically at least two) serve as the “destination” or the desired result of the lesson to be designed. Everything that transpires through the course of a lesson or unit, including lecturing, activities, and reading, is done with the intention of helping students acquire or deepen those understandings. To some new practitioners of backward design, the focus on enduring understandings can be misunderstood as an eschewal of established curriculum frameworks or standards; however, this is not the case. State or national standards help to determine what content will be taught, whereas backward design helps teachers determine how content will be taught, and what broader understandings students will gain from the study of content.

Determining Evidence of Learning

Once enduring understandings are identified (along with key skills, terms, and knowledge), practitioners of backward curriculum design must determine what evidence will be sufficient to show that students have achieved the desired results. In other words, how will the instructor know that the desired understandings have been attained? Typically, backward design calls for multiple and varied assessments to occur throughout a unit or lesson and that ignorance (or lack of learning) should be assumed until evidence proves otherwise. Common types of assessments are informal checks for understanding, traditional tests or quizzes, open-ended academic prompts that are addressed orally or in writing, and authentic performance tasks that simulate (to the extent possible) the real-world problems or situations and which result in tangible products or polished performances.

Most often, authentic performance tasks are used as a unit’s summative assessment (toward the end of the learning experience), whereas other types of assessment are used in a formative way (throughout the learning experience). Authentic performance tasks are intended to engage students in realistic questions or challenges that provide an opportunity to demonstrate the extent to which a student has grasped the intended enduring understanding. For example, students studying the Great Depression might be asked to look at a present-day issue that pits the rights of individuals against those of corporations and to make judgments about what should be done.

Planning the Learning Experience

Backward curriculum design is largely grounded in constructivist learning theory, and it asks practitioners to consider various elements of accepted pedagogical techniques when planning instruction. The following are the seven key considerations that must be present in the design of any lesson or unit:

- Do students know what they are learning and why it is important?
- Are students enticed to find the learning intriguing through some kind of initial hook?
- Throughout the lesson, are students provided with the necessary skills, knowledge, and experiences that will help them gain understanding and meet performance goals?
- Are students given opportunities throughout a learning experience to reflect on what they are learning, and revisit the big ideas of the lesson, and rethink their opinions?
- Are students given time to self-assess their learning and progress?
- Is the lesson differentiated to account for different learning needs and various levels of interest and readiness?
- Is the lesson organized in a way that makes understanding likely to happen (*vs*. being organized to ensure coverage of a topic)?

Although backward design requires teachers to spend time planning and organizing learning experiences, most practitioners see those prepared plans as flexible blueprints that are subject to change as rates of student learning become evident. Typically, instructional plans are designed with some flexibility to allow some students to move more quickly and possibly interact with more “important” and “worthwhile” knowledge and skills, whereas others remain focused on the essential components of the lesson. Determining what is essential, important, and worthwhile during the planning stage allows the instructor to make informed decisions about how to adapt the lesson while it is in progress.

Backward Design of Schools or School Systems

The same way of thinking that is used to plan individual lessons or units can be used to design system-or school-wide improvement. At base, the idea is simple and straightforward—all educators in a school or system decide together on what is most critical for students to know, understand, and be able to do by the end of their schooling and then educators work in a cohesive, coordinated way to ensure student success. However, this stands in contrast to many countries' deeply entrenched tradition of teacher isolation and autonomy, and it requires significant and sustained support on the part of school leaders. Schools that are using backward design as an improvement or reform strategy are guided by the following principles:

- Educators' job is to bring about student learning.
- All expectations for student learning should be clearly defined and regularly measured.
- When gaps are evident between what students are learning and what they are meant to be learning, it is the job of all educators to close such gaps.
- Schools must plan for reform with end results in mind.

Backward Design of System-Wide Programming

Schools using a backward design framework use their stated mission as a guide for all decisions and that mission reflects those big, cross-disciplinary ideas, understandings, and dispositions that should be reflected in every part of the curriculum. Those ideas then guide vertically coordinated teams of content-area specialists (e.g., arts, English, math) who determine which big ideas lie at the heart of their specific discipline, and thus which skills, dispositions, and understandings are the ultimate goals for student achievement. Those skills, dispositions, and understandings serve as the backbone of the curriculum, and they recur frequently as students move through their study of each discipline; content standards are also used in service to those larger goals but are not considered ends in themselves. Each course in a given discipline is designed to ensure student growth around that discipline's overarching goals; essential questions and enduring understandings that both fit with those goals, and with more content-specific goals, guide the instruction. Ten principles guide system-wide backward design curriculum work:

- A central goal of schooling is the ability learn knowledge, skills, and understandings that are highly flexible and can be adapted for use in various real-world situations.
- Students must understand the value of what they are learning so that they are motivated to undertake worthwhile challenges.
- Successful transfer of knowledge and skills requires that big ideas recur frequently throughout a curriculum.

- Enduring understandings cannot be “delivered” but must be uncovered so that student can make sense of the power of big ideas.
- All learners (children and adults) need clear guidance about what constitutes excellence in a given setting or assignment.
- All learners (children and adults) require timely, ongoing, specific feedback in order to improve.
- All learners (children and adults) require regular reflection in order to gain or deepen understanding.
- All learners (children and adults) must be encouraged to rethink and refine previously held beliefs.
- All learners (children and adults) need a safe and supportive environment in which to learn.
- Learning is most effective when it can be made personal and when it connects to or has application in the learner’s everyday life.

District and school leaders should consider a three-phase approach to system-wide backward design. In the first phase, school and district leaders identify desired results, including long-term goals for what both students and teachers should know, understand, and be able to do. In the second phase, school leaders must determine how they will gauge success, that is, how they will know the extent to which goals are being achieved. In the third phase, leaders must plan for initial and ongoing actions toward achieving goals, carefully considering what types of support and resources teachers will need as changes are implemented. Six primary responsibilities fall to school leaders: helping to craft a mission, ensuring a coherent curriculum, identifying gaps between mission and reality, personnel management and development, effective policy creation and management of resources, and guidance of school culture.

A Backward Design Approach to Evaluation

Backward design may be of particular use to evaluators, as it is a framework against which to judge the quality of educational systems or programs. Using backward design, evaluators can operationalize key variables involved in school, program, or curriculum design.

Curriculum Mapping

Regardless of a school district’s size, backward-designed evaluation of curriculum can help ensure that all students, regardless of which building they are housed in or which program of studies they choose to follow, are experiencing opportunities to engage with those ideas, dispositions, and skills that a district’s educators have identified as the most essential. Critical to successful backward design of system-wide curricula is proper curriculum mapping. The purpose of curriculum mapping in a backward-design setting is not to guide pacing or lock-step instruction of content, but rather to ensure that big ideas—those central to the school, to the discipline, and to subdisciplines—are recurring throughout the curriculum in ways that are intellectually coherent.

Backward-designed evaluation also usually looks at indications of authentic performance tasks, analytic and longitudinal rubrics, examples of work, suggested resources, formative assessment strategies, suggestions for differentiation, and troubleshooting strategies. The job of curriculum auditors or evaluation teams using a backward-design framework is to predict and identify problems with existing curricula, to collect feedback about existing curricula from stakeholders, and to identify gaps between written, taught, and tested curricula. Furthermore, evaluators may examine the extent to which a curriculum adheres to the principles of understanding and knowledge transfer on which backward design is based.

Gap Analysis

Identifying gaps between mission (or goals) and reality (or results) is a core component of backward design, and one that is useful for educational evaluators. Gaps may be evidenced by quantitative data, such as test scores, graduation rates, absenteeism, grade distributions, or other observable indicators; gaps may also be evidenced by qualitative data such as school accreditation reports, surveys of constituent groups, or targeted interviews. The identification of such gaps is predicted on the clear articulation of the desired outcomes of any program or curriculum.

Usually, gap analysis is based on the goals of backward curriculum design. For example, a school might determine that one of its primary reform goals is to allow students regular opportunities to reflect on and revise their work based on feedback from teachers and peers. In that case, an evaluator (or team of evaluators) would collect data in order to determine the extent to which formative assessments are used to provide student feedback, the frequency with which students are asked to assess peers, and the frequency with which students are given the opportunity to revise their work.

Gap analysis may also be based on programming or school-wide goals grounded in a theory of action about improvement. For example, a school may determine that critical thinking is an important outcome goal for students and that high-quality questioning strategies is a key component involved in helping students develop that skill. Data must then be collected, usually through a process of observation-based, fine-grained evidence recording, to determine the extent to which such practices are at work in the school. Gap analysis is also helpful for evaluating school policies, procedures, and physical spaces. With clearly defined goals for understanding and dispositions, evaluators can examine the extent to which resource allocation, discipline policies, homework practices, and the layout of physical spaces supports or constrains student understanding of essential concepts.

Rebecca Mazur

See also Action Research; Authentic Assessment; Constructivist Approach; Curriculum; Curriculum Mapping; Formative Assessment; Zone of Proximal Development

Further Readings

Brooks, J. G., & Brooks, M. G. (2001). In search of understanding: The case for constructivist classrooms. Upper Saddle River, NJ: Merrill/Prentice Hall.

McTighe, J., & Wiggins, G. P. (2013). Essential questions: Opening doors to student understanding. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiggins, G. P., & McTighe, J. (2005). Understanding by design (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Wiggins, G. P., & McTighe, J. (2007). Schooling by design: Mission, action, and achievement. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiggins, G. P., & McTighe, J. (2012). The understanding by design guide to advanced concepts in creating and reviewing units. Alexandria, VA: Association for Supervision and Curriculum Development.

Zmuda, A., McTighe, J., Wiggins, G., & Brown, J. L. (2007). Schooling by design: An ASCD action tool. Alexandra, VA: Association for Supervision and Curriculum Development.

Audrey Michal Audrey Michal Michal, Audrey

Priti Shah Priti Shah Shah, Priti

Bar Graphs

Bar graphs

168

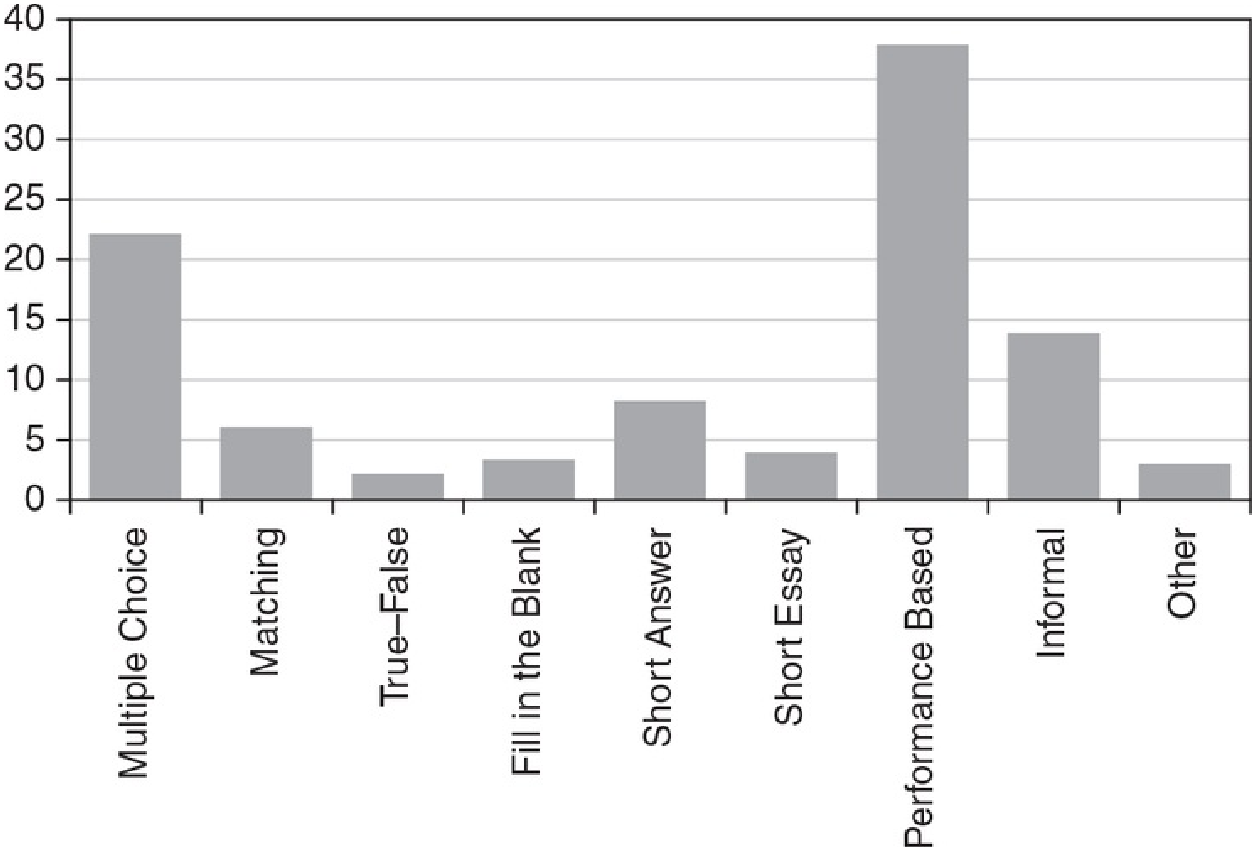
170

Bar Graphs

Bar graphs (also called *bar charts*) are a type of data visualization in which data points are represented by rectangular bars. Typically, the bars extend vertically from the bottom of the *x*-axis up to the data value, which is plotted along the *y*-axis; thus, the height of the bar (physical magnitude) is analogous to the numerical magnitude of the data point; bars may also be horizontal with length representing magnitude. Each data point is either labeled along the *x*-axis or referenced in a legend in a separate location near the graph. This entry discusses how bar graphs are used, factors that affect comprehension of bar graphs, and implications for educational research.

Bar graphs are often used to communicate scientific results, qualitative trends, and statistical analyses, such as main effects and interactions. Because data points are aligned along a common axis, bar graphs can facilitate comparison between individual data points; for instance, a user can quickly assess whether the data points are the same or different. Additionally, a user can easily compare differences between data points by judging the relative sizes of height gaps between bars. Global patterns, such as linear trends, are also salient in bar graphs. Thus, bar graphs are generally better for visualizing qualitative data patterns than exact values, which are more easily extracted from tables. In Figure 1, a simple bar chart shows the percentage of times different classroom assessment formats were used by a sample of teachers.

Figure 1 Percentage of times for different classroom assessment formats



Factors That Affect Comprehension

The visual features of a bar graph, such how the bars are organized and how far apart the bars are from one another, can affect comprehension. Bar graph comprehension is often facilitated when the bars are grouped in various ways; for instance, bars that are clustered together along the *x*-axis or that have the same color are perceived as belonging to the same group and viewers are more likely to make comparisons within rather than across such groups. Additionally, larger effects (i.e., larger height differences between bars) are more salient and easier to perceive.

Because bars are usually spatially segregated in a bar graph, it is typically easier to compare discrete values than continuous values, for which line graphs are better suited. However, Jeff Zacks, Ellen Levy, Barbara Tversky, and Diane Schiano found that discrete height judgments of bars are subject to bias from neighboring elements in bar graphs (such as the presence and height ratio of nearby bars).

Additionally, the knowledge that the user brings to the graph can affect comprehension of bar graphs. One type of knowledge that affects comprehension is graphical literacy, which involves having basic knowledge about how graphs should look and what they are used for. Having graphical literacy would include, for example, knowing that independent or categorical variables are represented along the *x*-axis and/or legend, whereas the dependent variable is represented along the *y*-axis. Graph expertise is especially helpful because it can allow the user to compare data more efficiently through the use of mental manipulation and perceptual shortcuts.

Another type of knowledge that influences graph comprehension is familiarity with the content of the graph. Although content familiarity generally enhances understanding of a graph, it can also bias how the graph is interpreted if the user expects to see a specific effect in the data due to overinterpretation, exaggeration of effects, or overlooking some effects while emphasizing others.

Bar graph comprehension is also subject to certain processing constraints and individual differences. People do not extract information from graphs all at once but rather proceed through the different regions of a graph, often returning to look at the same region several times. Additionally, because the user must keep track of multiple elements, such as which bars to compare and how variables correspond to a legend, comprehension may be constrained by visuospatial skills and working memory capacity, which vary substantially across individuals.

Implications for Educational Research and Communication

Contrary to popular belief, bar graphs are not always intuitive, and the time it takes to comprehend a graph is more akin to the time it takes to read a paragraph than to perceive an image. Making inferences about data can be especially difficult for students. Additionally, Audrey Michal, Priti Shah, David Uttal, and Steven Franconeri used eye tracking to show that young children are more likely than adults to attend to bar graphs from left to right and that this left-first strategy was associated with inefficient graph comprehension. Finally, bar graphs may display several effects simultaneously, and it is not always clear which subset of the data are relevant.

Several design principles are thought to enhance the effectiveness of a bar graph as a communicative tool. Any significant differences between data points should as a communicative tool. Any significant differences between data points should be large enough to be discriminated visually. Relevant comparisons can be emphasized with visual cues (e.g., highlighting) or grouping cues, such as spatial proximity or similar color. The use of extraneous information should generally be limited because it can interfere with magnitude judgments and overload working memory. For instance, Zacks, Levy, Tversky, and Schiano showed that adding extraneous depth cues (i.e., 3-D information) to bar graphs lowered accuracy of magnitude judgments by a small amount. Additionally, Jennifer Kaminski and Vladimir Sloutsky found that adding extraneous visual information to bars in a graph, such as countable objects, interfered with young children’s ability to compare the bars based on physical magnitude.

In educational settings, bar graphs are often used to report test scores, grades, and other evaluative information. Bar graphs are a useful format for score reporting because they can facilitate comparisons, such as between scores of different groups of students or between subscores for an individual student. However, bar graphs of score reports are subject to biases, such as overinterpretation of salient data (e.g., seemingly large effects), emphasizing relative versus absolute score differences, and oversimplifying results. These biases are particularly likely to occur when the user is unfamiliar with the content of the graph; thus, users should take caution when interpreting score reports.

Audrey Michal and Priti Shah

See also Data; Data Visualization Methods; Data-Driven Decision Making; Quantitative Literacy; Quantitative Research Methods; Scatterplots; Score Reporting Bayesian Statistics

Further Readings

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. Topics in Cognitive Science, 3(3), 446–474.

Kosslyn, S. M. (2006). Graph design for the eye and mind. New York: Oxford University Press.

Shah, P., Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. Cambridge, UK: Cambridge University Press.

David J. Weiss David J. Weiss Weiss, David J.

Basal Level and Ceiling Level

Basal level and ceiling level

170

170

Basal Level and Ceiling Level

The basal level (also called the floor level) and the ceiling level are components of the termination criteria used in Binet-type individually administered adaptive tests that are used to measure IQ in educational and other settings. These tests are administered by psychologists or trained examiners and operate from a question/item bank that is organized into mental age levels. Mental age is defined for each test item as the average chronological age at which approximately 50% of the standardization examinees correctly answered the item.

The adaptive test is begun by the examiner by selecting a mental age level based on the examiner’s knowledge of the child being tested. This information can simply be the child’s chronological age or can be based on other information such as a teacher’s statement about the child’s probable IQ (e.g., she is an “above average” student or he is “below average” or “average”). Having selected a mental age level to begin the test, the examiner administers each item at that level and scores the result as correct or incorrect.

When all items at this entry level have been administered, the score for that level is tallied. If the child has answered all items at that level correctly, the examiner has identified the *basal level* for that test and has identified the upper limit of the portion of the item bank that is too easy for the child. If the basal level has not been identified, the examiner then has the option of moving to the next higher or lower mental age level and administering the items at that level. Again, the proportion correct is tallied and a decision is made based on that information. If the basal level had not yet been identified, the test can continue with items at lower mental age levels until the child correctly answers all items at a given mental age level.

Alternatively, the examiner can first seek the *ceiling level* for that child—the mental age level at which all items are answered incorrectly, thus identifying the lower limit of the portion of the item bank that is too difficult for the child. Once both the basal level and the ceiling level have been identified, the test is terminated and the portion of the item bank that provides effective measurement for the child has been identified.

The IQ score that is computed from this procedure is based on a weighted function of the mental ages of the items answered correctly between the basal and ceiling levels. Because the items administered in the adaptive test are selected from the item bank based on the child's performance as the test is administered, the scores resulting from this type of test have greater measurement precision than those from tests in which all examinees receive the same set of items, many of which might not be appropriate for a given examinee.

David J. Weiss

See also Basal Level and Ceiling Level; Computerized Adaptive Testing; Intelligence Quotient; Intelligence Tests; Stanford–Binet Intelligence Scales

Further Readings

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. Journal of Methods and Measurement in the Social Sciences, 2(1), 1–27.

David Kahle David Kahle Kahle, David

Bayesian Statistics

Bayesian statistics

170

176

Bayesian Statistics

Bayesian statistics is a comprehensive and systematic interpretation of the field of statistics based on the quantification and manipulation of uncertainty in the form of probability distributions enabled by the Bayesian interpretation of probability laws. It is commonly considered a branch of statistics. This entry provides a basic description of Bayesian statistics. Although Bayesian statistics includes nonparametric approaches, the entry's scope is limited to the more common setting of parametric Bayesian inference.

Although the historical details surrounding the origins of Bayesian statistics are somewhat hazy, the broad strokes are well-documented. The basic theoretical machinery underpinning Bayesian statistics was established in the second half of the 18th century with the discovery of Bayes's theorem by Thomas Bayes, Richard Price, and, independently, Pierre-Simon Laplace. If beliefs concerning unknown quantities are represented with probability distributions, Bayes's theorem provides a mechanism to update the beliefs upon the arrival of new evidence, thus laying a foundation for scientific reasoning. This way of thinking became the de facto standard among statisticians from the time of its discovery until well into the 20th century under the name of *inverse probability*.

The modern term *Bayesian statistics* grew out of a great and divisive disagreement in the 20th century concerning the interpretation of probability as a contrast to the term *frequentist statistics*—statistics viewed from the perspective of the frequentist interpretation of probability. Although the frequentist methods of Ronald Fisher, Jerzy Neyman, and Egon Pearson dominated statistical thought for the majority of the 20th century, Bayesian methods did not achieve widespread use until the mid-to-late 20th century due to computational challenges of the paradigm. These challenges were significantly alleviated near the end of the 20th century due to the proliferation of the personal computer and advances in Monte Carlo algorithms.

By the late 20th century, Bayesian ideas had permeated virtually every area of statistical science. Now in the early 21st century, Bayesian methods continue to flourish and expand to larger audiences; however, frequentist methods are still more commonly used in practice and accepted by regulatory agencies. They also comprise virtually all introductory statistics education.

Introduction and Notation

The basic problem of statistical inference assumes that the observed data y_1, y_2, …, y_n constitute a random sample from a population of interest represented by a probability distribution $p(y)$, which is either a probability mass function, if Y is a discrete variable, or a probability density function, if Y is a continuous variable. Although the exact distribution of Y is unknown, $p(y)$ is typically assumed to be one of a collection of possible distributions called a *statistical model M*, and the problem is to determine which distribution in the model represents the population, the true distribution of Y. In the common setting of parametric statistics, the distributions in the model are indexed by one or more quantities called parameters, collectively denoted θ, so that $M=\{p(y|\theta)\}_{\theta\in\Theta}$, where Θ is the set of indices called the parameter space. Under the assumption that $p(y)$ is one of the distributions in M, $p(y)=p(y|\theta*)$ for some $\theta*\in\Theta$, and the inference problem is reduced to estimating θ^* using the data. The estimator is commonly denoted ; it is a function of the data so that .

The field of statistics provides many methods to construct estimators of θ, regardless of the perspective one takes on how to interpret probability. One common way is maximum likelihood estimation (MLE). An MLE, , is an estimator that maximizes the likelihood function $I(\theta|y)=p(y|\theta)$, which has the same functional form as the distribution $p(y|\theta)$ but is considered a function of θ for a fixed value of the variable $Y = y$. The *likelihood principle*, a philosophical underpinning of statistical inference associated with the 20th-century statisticians Ronald Fisher and Allan Birnbaum, states that conclusions about θ drawn from the data ought only depend only the data through I(θ|y).

Flipping a coin provides a simple example of this to clarify notation. Suppose Y is a variable denoting the result of flipping a coin with $Y = 1$ if the coin flips heads and $Y = 0$ if tails. If the coin flips heads with probability θ, $P[Y{=}1|\theta]\theta$ and $P[Y{=}0|\theta]1-\theta$, so that $p(y|\theta)=P[Y{=}y|\theta]=\theta^{y}(1-\theta)^{1-y}$ for $y = 0$ or 1, and the model is $M = \{p(y \mid \theta)\}_{\theta\in\Theta} = \{P[Y = \mathrm{y} \mid \theta]\}_{\theta\in\Theta} = \{\theta^{\mathrm{y}}\,(1-\theta)^{1-\mathrm{y}}\}_{\theta\in\Theta}$, where $\Theta=(0,1)$. If Y_1, $Y_2, \ldots, Y_n$ denote n independent flips of the coin, the joint distribution of any sequence of flips is:

$$\begin{aligned} &P\left[Y_1 = y_1, \ldots, Y_n = y_n \mid \theta\right] \\ &\quad = p\left(y_1, y_2, \ldots, y_n \mid \theta\right) \\ &\quad = \prod_{i=1}^{n} p\left(y_i \mid \theta\right) = \theta^{\sum y_i}\left(1-\theta\right)^{\sum 1-y_i}. \end{aligned}$$

For a given data set $y_1, y_2, \ldots, y_n$, this expression is the likelihood $\mathrm{I}(\theta|y_1, \ldots, y_n)$, and it can be shown that the MLE for θ is , that is, the proportion of 1's in the observed data set. The Bayesian solution to this estimation problem is different and described later in this entry.

The Bayes's Theorem and Associated Distributions

Of fundamental importance to the Bayesian approach is the idea that all uncertainty is represented by probability distributions. Thus, in the Bayesian inferential setting, because the parameter θ is an unknown quantity (or quantities), it must be assigned a probability distribution that encodes all the uncertainty about θ—what values it is likely to be and what values it is unlikely to be. Because of this, it is often said that Bayesians treat parameters as random variables. This is only true in the sense that parameters are given distributions in the same way measured variables are; however, the quantities themselves (θ) are believed to be fixed, unknown quantities. Upon seeing data, the beliefs about θ change: They are updated in the light of the new evidence, the data.

The fundamental mechanism enabling the updating process is Bayes's theorem. In brief, Bayes's theorem is a mathematical result detailing how one can reverse the order of probabilistic conditioning. The probability density version of the result is introduced here for self-containment, although the result also applies in discrete and even more general settings. If $D = \{y_1, y_2, \ldots, y_n\}$ is the data, Bayes’s theorem states:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta)d\theta}$$

$$= \frac{\mathrm{I}(\theta \mid D)p(\theta)}{\int \mathrm{I}(\theta \mid D)p(\theta)d\theta}.$$

In this equation, $p(\theta)$ is known as the *prior distribution*; it represents the beliefs about θ before the data are gathered. It is a probability density function defined on the parameter space Θ. The quantity $\mathrm{I}p(\theta|D)$ is the likelihood of the data previously described. The quantity on the left $p(\theta|D)$ is called the *posterior distribution* and represents the updated beliefs about the parameter θ after having seen the data D. The quantity in the denominator is the marginal probability of the data; it is a weighted average of the likelihood of the data over all the probability distributions in the model M and is equal to $p(D)$, while nevertheless still incorporating the modeling decisions of the functional form of $p(D|\theta)$ and the prior belief $p(\theta)$. This marginal distribution normalizes the product of the likelihood and the prior into a proper probability distribution. Considering the equation only as a function of the parameter(s) θ, the result is often written simply as:

$$p\left(\theta \mid D\right) \alpha \mathrm{I}\left(\theta \mid D\right) p\left(\theta\right).$$

The prior distribution $p(\theta)$ is often selected from a parametric family of probability distributions. When it is, it is commonly written in reference to the parameters that characterize it, $p(\theta|\eta)$; these parameters are referred to as *hyperparameters* and play an important role in Bayesian modeling.

The Bayesian approach also includes two *predictive* distributions that can be used at different stages of the inferential process to make predictions about the next observation of the variable Y. These distributions do not depend on the value of the parameter θ, as they eliminate it through the process of marginalization. The *prior predictive distribution* of the variable Y is the distribution $p(y)=\int p(y|\theta)\ p(\theta)d\theta$. It is the marginal distribution of the variable when the belief about the parameter is represented by the prior distribution $p(\theta)$, before data are collected. It can be used to predict the value of the variable using only the prior belief about the parameter.

The *posterior predictive distribution* of the variable is the distribution obtained when using the posterior distribution $p(\theta|D)$ instead of the prior distribution $p(\theta)$ to represent the belief about θ in the marginalization process: $p(y \mid D)=\int p(y|\theta)\, p(\theta|D)d\theta$. The posterior predictive distribution can be used to predict the value of a new observation Y having first had a belief about θ represented by the prior $p(\theta)$, seen data *D,* and updated one’s belief according to Bayes’s theorem.

Marginalization plays a key role throughout Bayesian statistics. In addition to the prior and posterior predictive distributions described earlier, marginalization can be used to systematically remove *nuisance parameters* from a model, parameters that affect the data but are not of interest. In the Bayesian setting, all parameters are endowed with a joint prior distribution that is updated through Bayes’s theorem to a joint posterior distribution. Marginalizing this distribution over the nuisance parameters provides a posterior distribution over only the quantities of interest, eliminating the nuisance parameters entirely. The ability of Bayesian methods to systematically and easily deal with nuisance parameters is considered a great advantage of the Bayesian paradigm.

To continue the estimation of a population proportion example from the previous section, the prior distribution is often selected from the family of β distributions, so that , where $B(\alpha,\beta)$ is the β function and the hyperparameter is the pair of β parameters α and β. The resulting posterior distribution $p(\theta|D,\eta)$ is also a β distribution; this property is described in the section on prior specification later in this entry. The prior predictive and posterior predictive distributions $p(y|\eta)$ and $p(y|\eta, D)$ are discrete distributions in the β-binomial family.

Bayesian Methods

Point Estimation

It is often desirable to have estimates of the unknown parameters θ. In a frequentist setting, maximum likelihood and the method of moments provide general strategies to construct estimators. In Bayesian statistics, after observing data, the belief about θ is completely described by the posterior distribution *p*(θ|*D*). If one is forced to reduce the posterior distribution into one parameter value, many options are available and routinely used in practice. All are summaries of the posterior distribution and referred to as Bayes's estimators. The most common is the posterior mean, but the median and mode of the posterior distribution are also routinely used. The latter method is referred to as *maximum a posteriori estimation* and is equivalent to maximum likelihood with a constant (uniform) prior.

Credible Intervals

One of the most commonly used statistical procedures is that of probability intervals, such as the confidence intervals of frequentist statistics. Confidence intervals are collections of parameter values generated from the data that contain the true population parameter with a certain probability. One of the limitations of the frequentist approach to probability sets is that the frequentist interpretation of probability does not result in a natural interpretation of the confidence set. Where one wants to say, "the probability the parameter is in the set is 1 − α," this statement is not meaningful (or useful) in the frequentist interpretation of probability, and one is forced into the interpretation "the probability the set contains the parameter is 1 − α." The Bayesian approach to probability intervals, called *credible intervals* or *credible sets* for more than one parameter, alleviates this problem, because it is meaningful to refer to the probability that a fixed yet unknown quantity lies in a certain region of space.

There are many ways to create Bayesian credible intervals, and as with estimation, all depend on the posterior distribution *p*(θ|*D*). In the case of a single parameter, the most common credible interval is the *quantile interval*: One simply forms an interval from the α/2 and 1 – α/2 quantiles of the posterior distribution *p*(θ|*D*); marginalization can be used to eliminate nuisance parameters. Although this method often allows for the easy computation of credible intervals, it does not always produce the smallest interval of a given probability 1 − α. This interval/set is called the *highest posterior density region*. A (1 − α) 100% highest posterior density region is the smallest region of the parameter space that contains 1 − α probability. Although highest posterior density regions are small, they suffer from two disadvantages: They can be difficult to compute, and they may not be connected.

Bayesian Hypothesis Testing and Bayes's Factors

In frequentist statistics, one is often tasked with assessing a null hypothesis H_0 : $\theta \notin \Theta_0$ against an alternative hypothesis H_1 : $\theta \notin \Theta_0$, and decision rules are constructed based on Type I and Type II errors. Bayesian statistics treats this problem very differently, and indeed the term *hypothesis testing* is hardly ever used in Bayesian statistics. Bayesian statistics views the choice between two hypotheses, or more generally two models, as simply another application of Bayes’s rule: Prior beliefs (often uniformly distributed) are updated, and decisions are based on posterior beliefs. Additionally, the Bayesian setting seamlessly accommodates several hypotheses/models in the same fashion. This is in stark contrast to the frequentist setting, where the selection of which hypothesis is the null, and the alternative is often chosen for mathematical convenience and yet has a dramatic effect on the result.

In the case of two competing models M_0 and M_1, where one or the other must be the case, a Bayesian analogue to frequentist hypothesis testing can be carried out with the use of Bayes’s factors. It is easily shown using Bayes’s theorem that

$$\frac{p(M_0 \mid D)}{p(M_1 \mid D)} = \frac{p(D \mid M_0)\,p(M_0)}{p(D \mid M_1)\,p(M_1)} = B\,\frac{p(M_0)}{p(M_1)}.$$

In words, the posterior odds of M_0 to M_1 is proportional to the prior odds with the constant of proportionality equal to , a quantity called the *Bayes’s factor* in favor of M_0. In basic cases, such as where the two models are representing different distributions in the same parametric family, the Bayes’s factor simply reduces to the likelihood ratio, which is widely regarded as an evidentiary measure in support of M_0. In general, it indicates the extent to which the odds of M_0 have changed in light of the data. If $B > 1$, the data advocate in favor of M_0, and if $B > 1$ in favor of M_1, and the further they are from 1 indicates stronger evidence. Various scales have been suggested as to how to interpret the magnitude of B in terms of standards of evidence; however, these are outside the scope of this entry. In general, B is considered a suitable statistic for scientific reporting, provided it does not change substantially as the prior distributions used to compute B, for $k = 0,1$, are changed. Note that in this description, M_0 and M_1 can be very different models; they can even have parameter spaces of different dimensions.

Hierarchical Models

Hierarchical Models

One of the major advantages of Bayesian statistics is that the paradigm provides a natural and flexible modeling platform: Bayesian hierarchical models. A *Bayesian hierarchical model,* also known as a *multilevel model* or *Bayesian network,* is a statistical model with hyperparameters that are endowed with prior structure of their own. A hierarchical prior structure breaks apart uncertainty in a prior distribution into its constituent components, each of which are individually much more manageable. Prior structures on prior structures, often repeated several times, can yield very flexible and expressive models, especially when data are included as part of the prior structures.

Challenges of Bayesian Statistics

Computation

One of the chief challenges in the practical application of Bayesian methods, and one of the reasons they took so long to become widely used in practice, were computational challenges presented by the paradigm. Specifically, computing the marginal integral $\int p\,(y|\theta)p(\theta)d\theta$ in the denominator of Bayes's theorem is often incredibly difficult and almost always requires numerical quadrature (integration) methods.

The exception to this rule is found with a class of priors called conjugate priors. A prior $p(\theta)$ for θ is called a *conjugate prior* if it is a member of a parametric family $\{p(\theta|\eta)\}_{\eta\in H}$, and the posterior is also a member of the same family. In other words, both the prior distribution $p(\theta) = p(\theta|\eta)$ and the posterior distribution $p(\theta|D) = p(\theta|\eta')$ are members of the same family of distributions, so that the parameter of the posterior distribution η' is some (often simple) function of the prior parameter and the observed data, $\eta' = g(\eta, D)$. Most textbooks on Bayesian statistics contain tables of such priors, which list for a given statistical model (likelihood) the conjugate family and the function g. From an applied perspective, the chief advantage of using a conjugate prior is the alleviation of the computing problem, and this is the chief reason why conjugate priors are so often used.

Apart from the conjugacy setting, there are two main ways other than numerical quadrature that Bayesian inference is carried out: variational methods and Monte

Carlo methods. Variational methods attempt to approximate the posterior distribution with a simpler distribution. For example, in many cases, the posterior distribution of the parameter converges to a multivariate normal distribution, as the sample size tends to infinity regardless of the prior used, a result known as the Bernstein–von Mises theorem. In these cases, it is often reasonable to approximate the posterior distribution with a surrogate multivariate normal distribution, which is very tractable, known as *Laplace approximation*.

More commonly used than variational methods are Monte Carlo methods, especially Markov chain Monte Carlo methods. Instead of computing the integral directly, the aim of Monte Carlo methods is to generate random samples from the posterior distribution, with the idea being that if one can sample from the distribution at will, one can compute any aspect of the distribution to an arbitrary degree of accuracy. Markov chain Monte Carlo methods are a class of sampling algorithms, including the Gibbs sampler and the Metropolis–Hastings algorithm, that involve a random walk whose stationary distribution is a probability distribution, here the posterior distribution. Markov chain Monte Carlo methods are typically the method of choice for most practicing Bayesian statisticians and are implemented as black boxes in the free software OpenBUGS, JAGS, and Stan.

Prior Specification

One of the most challenging components of a proper Bayesian analysis is the specification of the prior distribution $p(\theta)$, and there is a vast literature dedicated to this topic. There are a number of approaches taken with prior specification. The systematic quantification of belief is known as *prior elicitation* and has been a subject of statistical and psychological research since the middle of the 20th century. Because the prior tangibly affects the data analysis, one of the key considerations to take into account is how *informative* a prior is; this is loosely defined as how strong its beliefs are concerning the unknown parameter. Strongly informative priors can dominate the likelihood in Bayes’s theorem, so that the posterior reflects almost entirely prior belief regardless of the data.

There are several terms used as opposites to informative: *non-* or *uninformative, reference*, *diffuse*, and in some cases *objective*. These are generally intended to represent vague or equal belief concerning the parameters. However, basic results demonstrate that no prior is truly uninformative, and priors that appear uninformative in one parameterization may be quite informative after transforming the parameter. This consideration led to the development of the *Jeffreys prior*, a prior distribution based on Fisher information invariant under transformations and widely regarded as having little influence on the posterior.

Unfortunately, the method to construct a Jeffreys prior does not always result in a *proper prior* (one having total probability one), and consequently the posterior may also be improper, which is generally considered a bad thing. As a last alternative, upon selection of a prior family of distributions $\{p(\theta|\eta)\}_{\eta \in H}$, the *empirical Bayes's* strategy selects the prior parameter . based on the marginal distribution $p(y|\eta) = \int p(y|\theta)p(\theta|\eta)d\theta$, typically the MLE; however, this strategy is not considered "fully Bayesian."

However the prior is selected, varying the prior distribution and observing the effect on the posterior is generally considered good practice. This is known as a *sensitivity analysis*. If the posterior varies considerably when the prior is changed, further investigation is recommended.

David Kahle

See also Bayes's Theorem; Distributions; Prior Distribution

Further Readings

Berger, J. (1993). Statistical decision theory and Bayesian analysis (2nd ed.). New York, NY: Springer.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. Bayesian ideas and data analysis: An introduction for scientists and statisticians. Boca Raton, FL: CRC Press.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? Bayesian Analysis, 1(1), 1–40.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press.

Lindley, D. (2000). The philosophy of statistics. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3), 293–337.

Lindley, D. (2013). Understanding uncertainty (2nd ed.). New York, NY: Wiley.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., & Rakow, T. (2006). Uncertain judgements: Eliciting experts' probabilities. Chichester, England: Wiley.

Talbott, W. (2001). Bayesian epistemology. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Summer 2015 ed.). Retrieved from http://plato.stanford.edu/archives/sum2015/entries/epistemology-bayesian/

Joseph K. Blitzstein Joseph K. Blitzstein Blitzstein, Joseph K.

Carl N. Morris Carl N. Morris Morris, Carl N.

Bayes’s Theorem

Bayes’s theorem

176

181

Bayes’s Theorem

Bayes’s theorem is a way of estimating the likelihood of some event having occurred, or some condition being true, given some evidence that is related to the event or condition. This entry describes how Bayes’s theorem is used, discusses three forms of the theorem, and provides detailed examples of the use of the theorem.

Throughout the sciences, we are faced with questions about how likely an event of interest is, given some information. For example,

> How likely is a student to achieve at least a certain level of academic performance in the future, given the student’s past performance?
> What is the probability that a patient has a certain disease, given the patient’s diagnostic test results and background health information?
> How likely is it that a defendant is guilty of a certain crime, given all available evidence?
> What is the probability that a coin would land heads at least 60 times in 100 tosses, given that the coin is fair?
> How likely is a certain hypothesis, given the observed data?

These are all questions about *conditional probability*. Philosophical controversies have raged for centuries about exactly how to interpret probability, but conditional probability has a simple, uncontroversial definition: the probability of an event A, given an event B (with $P(B) > 0$), is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Swapping the roles of *A* and *B*, we have

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Note that *P*(*A*|*B*) is different from *P*(*B*|*A*). Confusing *P*(*A*|*B*) with *P*(*B*|*A*) is a common—and commonly devastating—blunder, sometimes called the *prosecutor's fallacy* (though not all prosecutors commit this fallacy, nor is the fallacy exclusive to prosecutors).

This definition immediately yields two useful expressions for *P*(*A* and *B*):

$$P(A|B)\,P(B) = P(A \text{ and } B) = P(B|A)\,P(A),$$

for *P*(*B*) > 0. Dividing through by *P*(*B*) gives a simple but powerful result that explains precisely how *P*(*A*|*B*) and *P*(*B*|*A*) are related. The next section gives several ways to express this relationship.

Three Forms of Bayes's Theorem

Basic Form

As explained earlier, a simple but fundamental consequence of the definition of conditional probability is the following theorem, which connects *P*(*A*|*B*) to *P*(*B*|*A*):

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Here *P*(*A*) is called the *prior* probability of *A* (it is the probability of *A* before we know whether *B* occurred), *P*(*A*|*B*) is the *posterior* probability of *A* given *B* (it is the updated probability for *A*, in light of the information that *B* occurred), and *P*(*B*) is the *marginal* or *unconditional* probability of *B*.

Remarkably, this theorem, whose proof is essentially just one line of algebra, has deep consequences throughout statistical theory and practice. Often *P*(*B*|*A*) is easier to think about or compute directly than *p*(*A*|*B*), or vice versa; Bayes’s theorem enables working with whichever of these is easier to handle and then bridging to the other. For example, in a criminal trial, we may be especially interested in the probability that the defendant is innocent given the evidence, but it may be easier at first to consider the probability of the evidence given that the defendant is innocent.

Bayes’s theorem is named after Reverend Thomas Bayes, due to his seminal paper *An Essay towards Solving a Problem in the Doctrine of Chances*, which was published posthumously in 1763 with help and edits from Bayes’s friend Richard Price. Bayes’s paper established conditional probability as a powerful framework for thinking about uncertainty and derived some important properties (including Bayes’s theorem). Some historical controversies have arisen about whether anyone discovered Bayes’s theorem earlier than Bayes, and how much of a role Price played.

The mathematician Pierre-Simon Laplace also played a crucial role in the early development of Bayes’s theorem. He rediscovered the result (apparently unaware of Bayes’s work), publishing it in a 1774 paper. A major part of Bayes’s and Price’s motivation was to provide ammunition for a theological debate with the philosopher David Hume; in contrast, Laplace showed that Bayes’s theorem could be used to tackle scientific problems and make sense of data.

Odds Form

The *odds* of an event *A* are given by odds

$$(A) = \frac{P(A)}{1 - P(A)}.$$

For example, the odds of an event *A* with probability 1/4 are (1/4)/(3/4) = 1/3; this is usually worded as “odds of 1 to 3 in favor of *A*” or “odds of 3 to 1 against *A*.” Bayes’s theorem has a very convenient formulation in terms of odds, which is especially useful when testing a hypothesis.

Let θ be a parameter of interest, and suppose there are only two possible values of θ: It is either θ_0 or θ_1. We wish to test the null hypothesis $\theta = \theta_0$ versus the alternative hypothesis $\theta = \theta_1$. We observe data $Y = y$. By Bayes’s theorem,

$$\frac{P(\theta = \theta_1 \mid y)}{P(\theta = \theta_0 \mid y)} = \frac{P(\theta = \theta_1)}{P(\theta = \theta_0)} \frac{P(Y = y \mid \theta = \theta_1)}{P(Y = y \mid \theta = \theta_0)}.$$

Note that, conveniently, $P(Y = y)$ has canceled out in this expression. There are three key ingredients in this statement:

> The ratio $P(\theta = \theta_1)/P(\theta = \theta_0)$ is the *prior odds* in favor of θ_1;
> the ratio $P(\theta = \theta_1|y)/P(\theta = \theta_0|y)$, is the *posterior odds* in favor of θ_1;
> the ratio $P(Y = y|\theta = \theta_1)/P(Y = y|\theta = \theta_0)$ is the *likelihood ratio*, which is the ratio of probabilities of the data that actually were observed, for the two different hypotheses. That is, Bayes’s theorem says how to update our prior odds to get our posterior odds for a hypothesis of interest, given the observed data.

Density Form

Much of statistical inference focuses on estimating unknown parameters or predicting future observations, given what is known (the observed data).

Let θ be the unknown parameter of interest in some model, and suppose that we observe the data $Y = y$. We wish to find the *posterior distribution*: the distribution of the unknown (θ), given the known (*y*). However, statistical models are typically formulated by giving the distribution for *Y* given θ. For example, the binomial model with parameters *n* and *p*, with *p* unknown, is given by:

$$P(Y = y \mid p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1 - p)^{n-y},$$

for $y = 0,1, \ldots, n$. But Bayes’s theorem lets us convert between these conditional probabilities: the posterior density of θ given the data *y* is:

$$g(\theta \mid y) = \frac{f(y \mid \theta) g(\theta)}{f(y)}.$$

Here $g(\theta)$ is the *prior* density for θ and $f(y)$ is the *marginal* (i.e., not conditioned on θ) density for *y*. The distribution of θ can be discrete or continuous and likewise for *Y*; we interpret “density” as a probability mass function in the discrete case and as a probability density function in the continuous case.

If θ and *Y* are both discrete, then this formulation of Bayes’s theorem is exactly the same as the basic form, just with different notation. If one or both of θ and *Y* are continuous, it is completely analogous to the basic form, with probability density functions replacing probabilities when needed.

Among the most central concepts in statistical inference is the *likelihood function*, which is the function $f(y|\theta)$, when viewed as a function of θ, with *y* fixed at the observed value of *Y*. When treating *y* as fixed, the denominator $f(y)$ acts as a constant, making $g(\theta|y)$ sum or integrate to 1. Then Bayes’s theorem has the following pithy summary: “Posterior is proportional to likelihood times prior.”

Examples

This section provides several detailed examples of applications for Bayes’s theorem.

Spam Filtering

Bayesian thinking suggests a way to build a spam filter for an e-mail system, using the fact that certain words appear much more frequently in spam (junk) e-mail than in legitimate e-mail. The goal is to determine in an automated way the probability that an e-mail message is spam, given information such as word frequencies in the e-mail.

Suppose that 80% of e-mail messages are spam. In 10% of the spam e-mails, the phrase “free money” is used, whereas this phrase is only used in 1% of nonspam e-mails. A new e-mail has just arrived, which does mention “free money.” Given this information, what is the probability that it is spam? We can answer this question readily using Bayes’s theorem.

Let *S* be the event that an e-mail is spam and *F* be the event that an e-mail has the “free money” phrase. By Bayes’s theorem,

$$P(S \mid F) = \frac{P(F \mid S)P(s)}{P(F)}.$$

The quantities $P(F|S)$ and $P(S)$ were given in the problem, so we just need $P(F)$. For this, we use the law of total probability:

$$P(F) = P(F \mid S)P(S) + P(F \mid S^c)P(S^c),$$

where S^c denotes the complement of *S* (i.e., the event that the e-mail is legitimate). Thus,

$$\begin{aligned} P(S \mid F) &= \frac{0.1 \cdot 0.8}{0.1 \cdot 0.8 + 0.01 \cdot 0.2} \\ &= \frac{80}{82} \approx 0.976. \end{aligned}$$

The strategy of using Bayes’s theorem in tandem with the law of total probability, as just shown, is extremely useful in a wide variety of problems. The example just given is simplistic, as it uses only one bit of information: whether the new e-mail contains the phrase “free money.” More realistically, suppose that we have created a list of, say, 100 key words that are much more likely to be used in spam than in nonspam. Let W_j be the event that an e-mail contains the *j*th word or phrase on the list. Let,

$$p = P(\text{spam}) \text{ and } p_j = P(W_j \mid \text{spam}),$$

where “spam” is shorthand for the event that the e-mail message is spam.

Let *K* be the observed evidence, specifying which of the key words appear and which do not appear. Then, Bayes’s theorem looks the same as before, with *K* in place of *F*:

$$P(S \mid K) = \frac{P(K \mid S)P(S)}{P(K)}.$$

However, it may be challenging to compute or estimate $P(K|S)$. Under the assumption that $W_1, \ldots, W_{100}$ are conditionally independent given that the e-mail is spam, and also conditionally independent given that it is not spam, the problem becomes much easier. A method for classifying e-mails (or other objects) based on this kind of assumption is called a *naive Bayes classifier*. For example, this conditional independence assumption implies that:

$$P(W_1, W_2, W_3^c, W_4^c, \ldots, W_{100}^c \mid spam)$$
$$= p_1 p_2 (1 - p_3)(1 - p_4) \ldots (1 - p_{100}).$$

The conditional independence assumption may be plausible or naive, but either way we can use it together with Bayes’s theorem to obtain a spam filter; later, we can directly evaluate the performance of the spam filter on a large sample of future e-mails.

Disease Testing

Suppose that a patient is being tested for a certain rare disease, afflicting 1% of the population. Let *D* be the indicator for the patient having the disease, defined to be 1 if the patient has the disease and 0 otherwise. Let *Y* be the indicator for the patient testing positive, so $Y = 1$ if the patient tests positive and $Y = 0$ otherwise. Suppose that $P(Y = 1|D = 1) = 0.95$ and $P(Y = 0|D = 0) = 0.95$. The quantity $P(Y = 1|D = 1)$ is known as the *sensitivity* or *true positive rate* of the test, and $P(Y = 0|D = 0)$ is known as the *specificity* or *true negative rate*.

The patient has just tested positive. How worried should he or she be? More precisely, what is the probability that the patient has the disease, given the positive test result? We know $P(Y = 1|D = 1)$, but want $P(D = 1|Y = 1)$; this calls for Bayes’s theorem.

Note that this problem has the same structure as the first example (disease corresponds to spam, and testing positive corresponds to “free money” being mentioned). The corresponding calculation yields:

$$\begin{aligned} P(D=1 \mid Y=1) &= \frac{P(Y=1 \mid D=1)P(D=1)}{P(Y=1)} \\ &\approx 0.16. \end{aligned}$$

So there is only a 16% chance of having the disease, given the test result. This seems surprising to many people at first because the specificity and sensitivity values make it sound as though the test is highly accurate. But Bayes’s theorem shows that there is a fundamental tradeoff between the accuracy of the test and the rarity of the disease. It is crucial to incorporate the prior information about the prevalence of the disease (the base rate) into our calculations, not just the information about the accuracy of the test.

Equivalently, we can work in terms of odds: posterior odds are prior odds times the likelihood ratio. Here, the prior odds are 99 to 1 against having the disease, but the likelihood ratio is 0.95/0.05 = 19 to 1 in favor of having the disease, so the posterior odds are 19/99 in favor of having the disease. This corresponds to a probability of 19/(19 + 99) ≈ 0.16, in agreement with the previous calculation.

Normal–Normal Model

The *normal–normal model* is a widely used statistical model, in which both the data and the mean parameter for the data follow normal distributions.

Assume a scalar target μ is given that a priori follows a normal distribution,

$$\mu \sim N\left(\mu_0, \tau^2\right),$$

with μ_0 and τ both being known. Suppose that an unbiased estimate *y* of μ becomes available, normally distributed with mean μ and known variance $V = \tau^2$, that is,

$$y \mid \mu \sim N(\mu, V).$$

Then Bayes’s theorem provides this normal distribution for μ given *y*:

$$\mu \mid y \sim N\left((1-B)y + B\mu_0, V(1-B)\right),$$

with

$$B \equiv V / \left(V + \tau^2\right).$$

The *shrinkage factor B* determines by how much the expected value of μ, given the data *y*, shrinks *y* toward the prior mean μ_0. The following example illustrates how this result about Bayes’s theorem in the normal–normal model context can be applied.

Evaluating Educational Testing

The Educational Testing Service once conducted experiments in several schools to see how effective coaching for SAT tests might be. Students in the school that showed the greatest gain averaged points higher on their SAT, which had a standard deviation of 15 points. Because this extreme school has the largest estimate, we would expect that its true value μ is likely to be less than 28. We will use the normal–normal Bayesian model again to estimate this.

We need to establish a base rate for μ, and our choice will be based on seven other schools that also evaluated their coaching effects. These other schools had an average effect of 6 points and a (between groups) standard deviation of τ = 11 points. We can summarize the normal–normal Bayesian model as follows:

$$\bar{y} \mid \mu \sim N\left(\mu, V = 15^2\right),$$

and

$$\mu \sim N\left(6, \tau^2 = 11^2\right).$$

The object is to determine the distribution of μ, the mean improvement at the extreme school, conditional on its observed . The variance *V* is 15^2 and the shrinkage factor $B = V/(V + 11^2) = 0.65$. Bayes's theorem (as given in the example introducing the normal–normal model) says that, given ,

$$\mu \mid (\bar{y} = 28) \sim N\left(0.35 \cdots 28 + 0.65 \cdots 6, 15^2 \cdots 0.35\right).$$

So conditioning on sample and using base rate information, the true mean μ at School A has expectation 13.7 SAT points with a standard deviation of 8.9 SAT points. With this, and because μ has a normal distribution, we have that μ lies in the interval 13.7 ± 1.96…8.9, that is, $-3.7 < \mu < 31.1$, with probability .95.

The extreme school had the largest effect, SAT points for the sample of students tested there. Even so, the base rate information suggests that its true coaching effect μ, if evaluated with many more of its students, probably lies much closer to the average of the other schools than to its own average, .

Conclusion

Along with the uses described in this entry, widespread Bayesian applications now exist that involve much more complicated data structures, enabled by high-speed computers and an ever-increasing array of efficient Monte Carlo techniques used to fit correspondingly complicated models. These advances emphasize the use of Bayesian hierarchical modeling, akin to but beyond the preceding normal–normal examples here. Readers especially interested in how these more advanced models apply to educational data are referred to *Hierarchical Linear Models: Applications and Data Analysis Methods*, by Stephen W. Raudenbush and Anthony S. Bryk.

Once the data (*y*) have been observed, as eventually happens with any application to real data, and given a prior distribution *g*(θ), Bayes's theorem allows statisticians to update their uncertainty about the unknown parameters (θ), given the known observed data (*y*). Intuitively, this is more meaningful than the frequency approach, which averages over values of *y* that might have occurred for the given data set but didn't.

Unfortunately, this advantage of the Bayesian approach is lessened because it requires a prior distribution *g*(θ) for θ, and how to make that choice has been the principal source of a long-standing philosophical controversy among statisticians. One widely used option is to choose a prior distribution *g* that provides little information, relative to the information in the given data. Then Bayes's theorem can be used to calculate a procedure for any *y*, so that the data, not the prior, dominate in determining the inference.

The Bayes's frequency controversy has diminished over the years. Now many thoughtful data analysts are able to develop approaches for their data from both perspectives. Indeed, there is little inherent conflict between Bayesian and frequentist approaches: Using Bayesian thinking to develop a procedure does not preclude using frequentist thinking to evaluate the procedure in repeated sampling.

Joseph K. Blitzstein and Carl N. Morris

See also Bayesian Statistics; Posterior Distribution; Prior Distribution

Further Readings

Bayes, T. (1683–1775). An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society, 53(1763), 370–418.

Blitzstein, J., & Hwang, J. (2015). Introduction to probability. Boca Raton, FL: CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis. Boca Raton, FL: CRC Press.

Hoff, P. (2009). A first course in Bayesian statistical methods. New York, NY: Springer-Verlag.

Morris, C. N. (1987). Comment. Journal of the American Statistical Association, 82(397), 131–133.

Raudenbush, S., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods: Vol. 1. Thousand Oaks, CA: Sage.

Stigler, S. (1983). Who discovered Bayes's theorem? The American Statistician, 37(4), 290–296.

Shriniwas Chinta Shriniwas Chinta Chinta, Shriniwas

Bayley Scales of Infant and Toddler Development Bayley scales of infant and toddler development

181

184

Bayley Scales of Infant and Toddler Development

Development is an umbrella term that encompasses language, cognitive, and motor as well as behavioral, social–emotional, and mental health domains. Screening, the process of testing infants and children to identify those needing further evaluation, is best conducted with standardized tests, which have a known rate of detection when administered correctly. The Bayley Scales of Infant and Toddler Development is one such direct assessment developmental screening measure. This entry describes developmental screening tests and then looks at the development and revision of the Bayley Scales of Infant and Toddler Development and the components, administration, scoring, and properties of the test.

The American Academy of Pediatrics recommends that all infants and young children be formally screened for developmental delay at periodic intervals and if concerns are raised by a parent or provider during routine developmental surveillance. It is estimated that 16% of children have a developmental and/or behavioral disorder. However, only 30% are identified before school entrance. Children with a disorder that is detected after school entrance miss the opportunity to participate in early intervention services. The primary goal of developmental surveillance and screening is early detection of developmental delays. Early detection by primary care providers results in early referral for diagnostic evaluation and early treatment, providing children with medical and ancillary support that is necessary to meet their full developmental potential.

Numerous tests exist that can be useful in screening delays in the five or so developmental domains: cognitive, gross and fine motor, speech and language, adaptive, and psychosocial. These clinician-administered, direct assessment screening tests have the benefit of direct assessment of skills and typically are used by pediatric health-care providers who have a particular interest in developmental problems. They may be used as the only screening test to complement the results of parent-report instruments or to explore an area of concern in greater depth (e.g., gross motor skills). Many health-care providers who use screening tests find that it enhances their relationship with the family and child and provides valuable information to make appropriate referrals. The Bayley Scales of Infant and Toddler Development is designed to assess the developmental functioning of infants and young children 1–42 months of age. It is used to identify suspected developmental delays in children and to provide information to plan and develop appropriate interventions.

Development and Revision of Bayley Scales

The Bayley Scales of Infant and Toddler Development is a standard series of measurements originally developed by psychologist Nancy Bayley used primarily to assess the developmental status of infants and toddlers, aged 1–42 months. This developmental measure consists of a series of play tasks and takes between 45 and 60 minutes to administer. Raw scores for successfully completed items are converted to scale scores and to composite scores. These scores are used to determine the child’s performance compared with norms taken from typically developing children of their age (in months).

Both the Bayley Scales of Infant Development (BSID) and the BSID, Second Edition (BSID-II) have been used in the assessment of severely delayed individuals who are outside the age range for which the test was standardized. The Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III), published in 2006, is a revision of the BSID-II.

The Bayley-III now includes growth scores that can be calculated to monitor the individual’s progress over time. The Bayley-III also can be used to obtain an estimate of developmental level when more age-appropriate measures cannot be used for older children or individuals with severe delays, such as those with profound mental retardation. The Bayley-III maintains the same types of tasks as those in previous editions, promoting task involvement through play-based activities for individuals with limited ability. The most significant revision to the Bayley-III is the development of five distinct scales (as compared to three scales in the BSID-II) to be consistent with the areas of appropriate developmental assessment for children from birth to age 3 years. Although the BSID-II provided Mental, Motor, and Behavior Scales, the Bayley-III revision includes Cognitive, Language, Motor, Social–Emotional, and Adaptive Behavior Scales.

The Bayley-III was standardized on a normative sample of 1,700 children aged between 16 days and 43 months and 15 days living in the United States in 2004. Stratification was based on age, gender, parent education level, ethnic background, and geographical area. Normative data for the Social–Emotional and Adaptive Behavior Scales followed the same stratification pattern but were derived from smaller groups (456 and 1,350 children, respectively).

The Bayley-III is a technically sound instrument, with strong internal consistency, as well as test–retest stability. The test was revised with the goal to update normative data, strengthen the psychometric quality, and improve clinical utility. The test is also revised to simplify administration procedures and instructions by reorganizing the manual. It now includes updated item administration by making the instructions more play based, reducing the effect of receptive and expressive language on cognitive items, and allowing caregiver involvement providing administration procedures are followed. The test has updated stimulus materials to allow selection of materials that appeal to the child and to make materials more appealing but at the same time maintains basic qualities of the Bayley Scales. The Bayley-III shows scores that are consistent with other ability tests that have been revised in recent years and shows expected levels in various clinical groups.

Description

The Bayley-III is composed of five subscales:

1. Cognitive subscale assesses play skills, information processing (attention to novelty, habituation, memory, and problem-solving), counting, and number skills.
2. Language Scale assesses communication skills including language and gestures. It contains two subsets:
   1. Receptive Language subscale
   2. Expressive Language subscale
3. Motor Scale is divided into two subsets:
   1. Fine Motor subscale

2. Gross Motor subscale

Two scales, the Social–Emotional Scale and the Adaptive Behavior Scale from the Social–Emotional and Adaptive Behavior Questionnaire, are completed by the parent or primary caregiver. The Social–Emotional Scale assesses emotional and social functioning as well as sensory processing. It is based on the *Greenspan Social–Emotional Growth Chart: A Screening Questionnaire for Infants and Young Children* (Greenspan, 2004). The Adaptive Behavior Scale assesses the attainment of practical skills necessary for a child to function independently and to meet environmental demands. It is based on the Adaptive Behavior Assessment System–Second Edition. The only modification to the Greenspan and Adaptive Behavior Assessment System–Second Edition in the Bayley-III is the use of scaled scores in addition to the originally provided cut scores, so that these measures may be more easily compared to the other Bayley-III subtest scores.

Test Administration and Scoring

Administration of each scale is started at a predetermined item based on the child’s age. A child must achieve a score of 1 on the first three consecutive items at the predetermined start point to achieve the basal score. If not, administration begins at the start point for the next youngest age level (reversal rule). The reversal rule continues to apply until the child has achieved the first three consecutive items beginning at the determined start point. To complete testing and achieve the ceiling, a child must score 0 on 5 consecutive items. After having received these five consecutive 0 scores, no further items are administered (discontinue rule).

The Bayley-III provides norm-referenced scores. When scoring, each of the five subscales is given a raw score based on the number of items the child has achieved in addition to the number of items preceding the basal that were not administered. Higher scores indicate more mature development.

From these raw scores, scaled scores can be calculated for the Cognitive Scale and the two combined Language Scales and Motor Scales. These scores can then be used to determine composite scores, percentile ranks and confidence intervals, developmental age equivalents, and growth scores. Scores for the Cognitive, Language, and Motor Scales are provided in 10-day increments for children aged 16 days to 5 months and 15 days and in 1-month intervals for children over 5 months and 15 days. Scaled scores for the Social–Emotional Scale are reported according to the stages of social–emotional development. Scaled scores for the Adaptive Behavior Scale are reported in 1-month intervals for 0–11 months, 2-month interval for 12–23 months, and 3-month intervals for 24–42 months.

Total administration time ranges from 50 minutes for children younger than 12 months up to 90 minutes for children 13 months and older. The Bayley-III is intended to be administered by individuals who have training and experience in the administration and interpretation of comprehensive developmental assessments. Those administering the Bayley-III should have completed some formal graduate or professional training in individual assessment.

According to the technical manual for the Bayley-III, diagnosing developmental delay can be based on any one of the several criteria: 25% delay in functioning when compared to same age peers, 1.5 standard deviation units below the mean of the reference standard, and performance of a certain number of months below the child’s chronological age. It cautions against the use of age equivalent scores as they are commonly misinterpreted and have psychometric limitations. It also states that scores on the Bayley-III should never be used as the sole criteria for diagnostic classification. It should also not be used to diagnose a specific disorder in any one area. Rather, poor performance in any particular area should be used as a measure to make recommendations or referrals for appropriate services.

Test Properties

Reliability

The Bayley-III has established reliability with internal consistency and shows reliability coefficients for the subscales and composite scores that range from 0.86 to 0.93. Reliability coefficients for the special groups assessed are similar to or higher than those of the normative sample, indicating that the Bayley-III is equally reliable for children with clinical diagnoses or risk factors as for the general population.

Test–retest reliability of the Cognitive, Language, and Motor Scales assessed on 197 children aged over 2–15 days shows correlation scores that range from 0.67 to 0.94 for the different subtests depending on the children’s age. Test–retest to 0.94 for the different subtests depending upon the children´s ages. Test–retest reliability for the Adaptive Behavior Scale was calculated by asking 207 parents to rate their child twice over 2 days to 5 weeks. Reliability coefficients ranged from 0.71 to 0.92. Scores from the Greenspan Social–Emotional Growth Chart, which makes up the Social–Emotional Scale, indicate strong internal consistency with coefficients ranging from 0.76 to 0.94.

Validity

The Bayley-III has established convergent and divergent validity after correlating with other relevant instruments. It has shown good correlation with the Wechsler Preschool and Primary Scale of Intelligence, Third Edition (intelligence correlation score between 0.52 and 0.83), the Preschool Language Scale, Fourth Edition (language correlation score between 0.50 and 0.71), and Peabody Developmental Motor Scales, Second edition (motor skills correlation score between 0.55 and 0.59).

The validity of the Bayley-III in children with specific conditions or risk factors was also examined. These "special groups" included children with down syndrome, pervasive developmental disorder, cerebral palsy, specific or suspected language impairment, asphyxiation at birth, prenatal alcohol exposure; those who were small for gestational age, premature, or low birth weight; and other children at risk for developmental delay. Results indicate that the Bayley-III is sensitive to differences in performance of typical children and children at risk for developmental delay. There is no specific information provided regarding the validity of the Bayley-III Social–Emotional Scale, and there is moderate to low validity correlation of Bayley-III Adaptive Behavior Scale with other similar scales.

Shriniwas Chinta

See also Adaptive Behavior Assessments; Childhood; Cognitive Development, Theory of; Standardized Tests

Further Readings

Bayley, N. (2006). Bayley Scales of Infant and Toddler Development (3rd ed.). Administration Manual. San Antonio, TX: Harcourt Assessment.

Bayley, N. (2006). Bayley Scales of Infant and Toddler Development (3rd ed.). Technical Manual. San Antonio, TX: Harcourt Assessment.

Greenspan, S. I. (2004). Greenspan social–emotional growth chart: A screening questionnaire for infants and young children. San Antonio, TX: Harcourt Assessment.

Harrison, P. L., & Oakland, T. (2003). Adaptive behavior assessment system (2nd ed.). San Antonio, TX: The Psychological Corporation.

Kimberly Ethridge Kimberly Ethridge Ethridge, Kimberly

Kimberly Capp Kimberly Capp Capp, Kimberly

Anthony Odland Anthony Odland Odland, Anthony

Beck Depression Inventory Beck depression inventory

184

187

Beck Depression Inventory

The Beck Depression Inventory (BDI) was published by psychiatrist Aaron Beck in 1961, with the aim of better assessing depression severity and characterizing symptomatology. The author recognized the multidimensional nature of depression and need to quantify symptomatology for screening purposes. Questions center on the patient's thoughts, feelings, and how the patient thinks and views the world and self. For example, question content reflects cognitive distortions, negative thoughts, low self-esteem, and suicidal ideation as well as somatic/affective components (e.g., sleep or appetite disturbances and fatigue).

The strength and popularity of Beck's original inventory and subsequent revisions published in 1978 and 1996 in part reflects its ease of use, strong psychometric characteristics, and ecological validity. This entry discusses the characteristics of the original BDI and subsequent revisions, then looks at its validity and reliability, the normative sample used in developing it, and its clinical use, administration, and limitations.

Versions of the BDI

BDI-I

The original BDI measure consisted of 21 multiple-choice questions asking the patient to rate their feelings over the past week. Questions were ranked on a Likert-type severity scale from levels 0 to 3 (3 representing more intense or

Likert-type severity scale from levels 0 to 3 (3 representing more intense or severe feelings). Respondents would be instructed to circle the number corresponding to the statement that was most accurate, and the responses were summed to yield a total score. Higher total scores indicated a more severe number of depressive symptoms.

Beck developed standard ranges and cutoffs for the scores, so that a clinical impression could be easily assessed from the total sum. Descriptors and ranges were minimal (0–9), mild (10–18), moderate (19–29), and severe (30–63).

BDI-IA

The BDI-IA was published in 1978 as an amended (revised) version of Beck’s original questionnaire. Improvements included rewording and restructuring some items to remove the (a) and (b) choices to make the choices clearer for patients. In the original questionnaire, examinees were asked to answer questions based on their mood over the preceding week. This time frame was lengthened to 2 weeks on the BDI-IA so as to allow for a wider range of possible life events and emotions that might be tabulated.

Despite increased ease of administration and use in this version, the BDI-IA only addressed six of the nine *Diagnostic and Statistical Manual of Mental Disorders, Third edition* (*DSM-III*) symptom criteria for major depressive disorder. This flaw prompted the second revision of the BDI.

BDI-II

The second revision of the BDI occurred with the advent of the *DSM-IV* in 1996. This version, the most recent as of 2017, is one of the most widely used depression screening measures. The BDI-II retains a 21-question format, although 18 items were reworded to reflect new diagnostic criteria accompanying implementation of the *DSM-IV*. Questions pertaining to suicide, interest in sex, and feelings of punishment were not revised. Items referring to sleep and appetite were reworded to account for both increases and decreases in these domains as *DSM-IV* allowed for either direction of the disturbance to count as symptom criteria. Items assessing body image, hypochondriasis, and difficulty working were removed, as they no longer reflected diagnostic criteria.

Identical to the BDI-IA, the BDI-II asks individuals to choose their responses based on their thoughts or feelings over the most recent 2-week span. Respondents circle the number of the statement that most closely matched their feelings or thoughts, with 0 being the *least severe feeling* and 3 being the *most severe feeling*. All 21 items are tabulated, with higher total sums indicating a more severe number of depressive symptoms. BDI-II descriptors are minimal (0–13), mild (14–19), moderate (20–28), and severe (29–63).

Validity and Reliability

The BDI-II has strong internal validity, external reliability, and high test–retest reliability. Twenty-one items are highly intercorrelated, demonstrating strong internal reliability with a correlation of .92 in outpatients and .93 in college students. Test–retest reliability is strong (.93; $p < .001$), when the questionnaire is readministered 1 week after the first administration. Content validity of the BDI-II is higher than the BDI-I or BDI-A, which is thought to reflect updates to item content to more closely align with the *DSM-IV* diagnostic criteria. The BDI-II has been shown to be able to consistently differentiate between depressed and nondepressed patients when administered. BDI-II scores were on average 3 points higher than the BDI-IA scores.

Factor analysis of the BDI-II by Beck revealed two main types of factors of depression: somatic-affective and cognitive factors. While additional research has suggested that there may be additional factors or variations in factors that are indicated on the BDI, Beck continued to use the two-factor approach in revising the original BDI for subsequent versions. Questions pertaining to the somatic or physical components of depression as well as the cognitive or thought disturbance aspects of the disorder are thoroughly addressed throughout the 21-item questionnaire.

The BDI-II is highly correlated with the Beck Anxiety Inventory, a screening measure geared toward physiological symptoms of anxiety that was also developed by Beck. For this reason, pairing of these two questionnaires may be beneficial in screening both for symptoms of depression and for symptoms of anxiety.

Normative Sample

The normative sample for the BDI-II was made up of 500 psychiatric outpatients (63% female) from both rural and suburban areas from across the United States. Participants had been diagnosed with depression using either *DSM-III-R* or *DSM-IV* criteria. The mean age of the normative sample was about 37 years with an age range of 13–86 years. The racial makeup of the sample was lacking in diversity (91% White, 4% African American, 4% Asian American, and 1% Hispanic). Another, smaller sample of 120 Canadian college students served as a normal comparative group. Age range and racial data on this sample are not reported in the literature.

Clinical Use

The BDI-II is one of the most commonly used screening measures for depression because its inception into neuropsychological batteries in the early 1960s. It is suitable for clients who have reading and comprehension abilities of at least a fifth-grade level and who understand standard written English. There is also a Spanish version of the BDI-II that can be used in appropriate populations.

Because the BDI-II is meant to be used as a primary screening measure, it can be given to any individual who is experiencing symptoms that are similar to, or diagnostic of, depression, regardless of whether or not there is a prior or current diagnosis of depression. It is a useful tool in clinical practice to help characterize associated somatic and cognitive disturbances.

Administration

Ease and speed of administration makes this an ideal way to screen for depressive symptomatology and other mood and related cognitive disturbances. It is used in private practice, hospital settings, and other clinical situations, in which a quick and reliable measure is needed.

The BDI-II is a self-report questionnaire and is filled out independently of the examiner’s assistance. Instructions are written at the top of the page and are also meant to be read aloud by the examiner to ensure comprehension. Instructions ask the examinee to carefully consider their feelings over the last 2 weeks. In order to obtain a total score, examinees are required to answer each item on the double-sided questionnaire. Examinees are instructed to choose a higher number (indicating greater severity) if they are torn between two response choices on an item. Administration is untimed but typically takes approximately 5 minutes item. Administration is untimed but typically takes approximately 5 minutes.

Limitations

The BDI-II may not be an effective screening measure for elderly populations due to the mixed age range of the normative sample as well as potential variations in symptom manifestation across the life span. Additional research may be required to determine the clinical utility of the BDI-II in older adults, although it should be noted that the Geriatric Depression Rating Scale is the most widely used depression screening measure for the older adult population currently. Furthermore, the BDI-II is only normed for ages 13 and older and should not be used for children or young adolescents of any race or ethnicity, unless there is a clinically defensible reason for doing so.

Based on the homogeneity of races in the normative sample (i.e., 91% White), there is little research on the efficacy of the BDI-II in different populations and ethnic minorities. Future research is warranted in this area as well to better improve screening measures for depression in racial and ethnic minority populations. There is a Spanish translation of the Beck, although there are currently no other translations available. The Spanish BDI-II is available for appropriate populations, but lack of alternate translations should be considered when giving this form to nonnative English speakers. Cultural variations in how examinees experience depressive symptoms suggest that translation without adjustment to item content might not be adequate for generalizability.

Clinicians and researchers should also consider how the BDI-II aligns with the *DSM-5* published in 2013. Clinicians and researchers must determine whether any such differences are relevant to the purpose of a given administration (e.g., diagnosis vs. symptom characterization). Significant discrepancies may require selection of an alternative questionnaire or supplementation of additional content that is noted with appropriate documentation and consideration for potential compromises to validity.

Clinicians should not use the BDI-II as the primary measure of diagnosis, as it is not meant to serve alone as a diagnostic tool. The BDI-II should rather be used as a screening measure to inform treatment, guide interventions and assessments, and help the examinee gain insight into the type and severity of symptoms experienced.

Kimberly Ethridge, Kimberly Capp, and Anthony Odland

See also Anxiety; *Diagnostic and Statistical Manual of Mental Disorders*; Minnesota Multiphasic Personality Inventory; Psychometrics; Screening Tests; Reliability; Test–Retest Reliability; Validity

Further Readings

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders: DSM-5. Washington, DC: Author.

Beck, A. T., Rial, W. Y., & Rickels, K. (1974). Short Form of Depression Inventory: Cross-Validation. Psychological Reports.

Beck, A. T., & Steer, R. A. (1984). Internal consistencies of the original and revised Beck Depression Inventory. Journal of Clinical Psychology, 40(6), 1365–1367.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory–II. San Antonio, TX: Psychological Corporation.

Brown, M., Kaplan, C., & Jason, L. (2012). Factor analysis of the Beck Depression Inventory–II with patients with chronic fatigue syndrome. Journal of Health Psychology, 17(6), 799–808.

Conoley, C. W. (1987). Review of the Beck Depression Inventory (revised edition). In J. J. Kramer & J. C. Conoley (Eds.), Mental measurements yearbook, 11th edition (pp. 78–79). Lincoln: University of Nebraska Press.

Erbauch, J. (1961). An inventory for measuring depression. Archives of General Psychiatry, 562, 53–63.

Sharp, L. K., & Lipsky, M. S. (2002). Screening for depression across the lifespan: A review of measures for use in primary care settings. American

Family Physician, 66(6), 1001–1008.

J. E. R. Staddon J. E. R. Staddon Staddon, J. E. R.

Behaviorism

Behaviorism

187

191

Behaviorism

Behaviorism is a movement in psychology that focuses on the study of behaviors that can be objectively measured by a third party. Some behaviorists give little or no consideration to internal or mental events that cannot be measured, although others acknowledge the importance of internal events. This entry discusses the emergence of behaviorism, then describes methodological behaviorism and radical behaviorism, and then describes how these two strands have evolved.

Emergence of Behaviorism

Behaviorism was presented to the modern world by Johns Hopkins psychology professor John Broadus Watson (1878–1958) in an influential 1913 article *Psychology as the Behaviorist Views It*. Watson's behaviorism is based on two claims: First, that individuals' observations about their actions, motives, and mental processes are scientifically irrelevant. Second—and it almost follows from the first assumption—that the data of a scientific psychology must come from things that can be measured, and measured not by subject, but by a third party. As for theory, Watson didn't even mention it: "prediction and control" of behavior was his aim. And he recognized "no dividing line between man and brute [i.e., nonhuman animals]" (1913/1948, p. 457).

None of this was entirely new; other scientists had also rejected human consciousness as a means of explaining behavior. The process by which we see, recognize, and interpret the visual world is also hidden from consciousness. The German physicist, philosopher, and physician Hermann von Helmholtz (1821–1894) pointed out that perception operates by a sort of "unconscious inference."

In 1934, an inventor, American ophthalmologist Adelbert Ames Jr. (1880–1955), built a special kind of room to illustrate the process of perceptual inference. Viewed through a peephole (i.e., from a fixed point of view), it looks like a regular room, with right-angle corners, and so on. But when a girl walks from one side of the room to the other, the girl seems to grow magically larger. The perception is wrong of course. The girl size has not changed. The reason the girl appears to grow is that the brain assumes—without the viewer's awareness—that the angles are all right angles and the floor is level, when neither is true.

Perception involves unconsciously using very partial data to call up a complete picture of whatever the individual (unconsciously) infers he or she is seeing. Visual illusions such as the Ames room show how this process can misfire. Other examples of unconscious processes include the "tip of the tongue" phenomenon: knowing the name of the old movie star on the screen but being unable to bring it to mind until suddenly it appears. Novelists frequently say that after a certain point, their characters seem to "write themselves." Mathematicians often say that proofs and theorems simply appear in their minds without any awareness of the complex calculations that must have been made to generate them.

If not conscious, these automatic processes must then be unconscious, yet Watson attacked the very idea of the unconscious. Behavior may be the product of unconscious processes, but what *are* they? On this, Watson's behaviorism was silent.

Methodological Behaviorism and Radical Behaviorism

Watson and other researchers of the early 20th century used rats and other animals to study learning. The dominant behaviorists of the time were Clark L. Hull (1884–1952) at Yale University; Edward Chace Tolman (1886–1959), a cognitive behaviorist at University of California, Berkeley; and, to a lesser degree, Edwin Guthrie (1886–1959) at the University of Washington. The dominant movement, Hullian, and then neo-Hullian, behaviorism, was relabeled by B. F. Skinner (1904–1990), as *methodological behaviorism*. Skinner contrasted methodological behaviorism to his own proposal, termed *radical behaviorism* and described in his 1938 book, *The Behavior of Organisms*, and many later works.

Methodological and radical behaviorism differ in several ways. The neo-Hullians were devoutly theoretical. They wanted to explain the process of learning, which seemed to require the between-subject method. To compare the effects of different experiences, a researcher cannot simply give the same animal the two experiences in succession because the animal is changed by the first experience. It’s no longer “naive,” so it may behave differently after Experience B if Experience A came first than it would have if Experience B came first. Given two identical animals, the researcher could give one the first experience, A, and the other the second, B, and look at the differences in behavior that result. But because no two animals are exactly the same, the researcher must settle for two groups, to which animals are randomly assigned: the *experimental* group, which gets the treatment being studied (A), and the *control* group, which gets no treatment (B). The average response of the groups must then be compared using a method called *null hypothesis statistical test*.

Skinner’s method was quite different. He invented a simple method using the Skinner box, a device used for the animal to give a measureable response, such as pressing a lever or pecking a colored disk, that can be rewarded automatically in the presence of controllable stimuli, such as lights, colors, or patterns. The method generates quantitative data, initially in the form of a *cumulative record*, which is a graph that shows real time on the *x*-axis and cumulated responses on the *y*-axis. Skinner also discovered that animals—in his case, pigeons as well as rats—yield stable and reversible adaptations to a variety of *reinforcement schedules*, which are rules saying what the animal must do to get a bit of food—make 10 lever presses or wait 30 seconds, for example. Because the pattern of behavior produced by a given schedule is stable and can usually be recovered even after intervening experience with a different schedule, Skinner’s method of what he termed *operant conditioning* allows the study of individual animals, which can be exposed successively, ABAB and so on, with the assurance that each exposure to A, say, will give the same result.

Both these approaches are flawed. The neo-Hullians developed theories based on the average behavior of groups and assumed that what was true of the group was true as well of the individual. Many theories attempted to explain the smooth learning curve typically found when a group of rats learns to choose the left versus the right arm of a T-maze. Yet each rat may in fact learn instantly, just with different delays for different rats. The average is smooth, but the individual is not. Indeed, Skinner famously published a cumulative record of a rat learning to bar press that shows just such sudden learning. But the neo-Hullians were undeterred and continued to deal entirely in group data.

undeterred and continued to deal entirely in group data.

Reliance on statistical testing of theoretical models has tended to deemphasize quantitative predictions in favor of simple binary tests. Theorists are often satisfied to show that A is greater (or less) than B, even if the actual quantitative difference may be very small. This is logically defensible, but in practice means that the theory being tested is weak, that it presents only a partial picture of the phenomenon under test.

Finally, a serious problem that affects many areas of social and biological science is the null hypothesis statistical test method. In recent years, problems with this method have been revealed as researchers have found that many experimental results in several areas, from social psychology to drug studies, have been impossible to replicate. An experimental result is accepted as fact if the chance of getting it by accident is less than 5%. The computation is based on assumptions about probability distributions that are often questionable. The single-subject method avoids the problems of the null hypothesis statistical test method but must cope with the fundamental irreversibility of the organism’s state. The pigeon may behave in the same way on the second exposure to a given reinforcement schedule, but it is not the same pigeon. The only solution to this problem is *theory*. A theory about how exposure to one condition will affect behavior in another can be tested with individual subjects. The researcher may have an idea about how a sequence of conditions, say AB, will affect the organism’s behavior in a new Condition C. For example, suppose the pigeon is trained with two choices: peck left or peck right. It is easy to show, with no statistics required, that if the pigeon is paid off for L for a few days then for R for a few days, then given nothing, it will try both L and R for a while before finally quitting. Conversely, another pigeon, equally naive at the beginning of the experiment, but rewarded only for pecking R, when reward ceases will peck L hardly at all. There are several theories that might explain this and other *transfer effects*.

Skinner, exponent of the single-subject method, ruled out theory, however. In an influential 1950 article entitled *Are Theories of Learning Necessary?* he answered emphatically “No!” and theories of learning languished among his followers. Watson also devalued theory, claiming that the objective of psychology should be to “predict and control” behavior rather than to understand it, even though the theory of evolution by natural selection shows that prediction is often impossible.

Behaviorism began to fragment in the 1930s and 1940s. The neo-Hullians, soon to become *associative learners*, were methodological behaviorists. They believe that psychology must restrict itself to third-party measureable data and not rely on private experience, that is, on introspection. Experimental psychologists and most neuroscientists accept methodological behaviorism. But after about 1960, the methodological behaviorists began to call themselves cognitive psychologists and ceased to identify with behaviorism. The essentials of methodological behaviorism have been absorbed by empirical psychology of all types.

But radical behaviorism, based on Skinner’s work, remains as a separate and vigorous movement. The reasons are partly practical. Skinner’s emphasis on contingencies of reinforcement as the drivers of all operant (instrumental) behavior has allowed the development of effective techniques for managing autism and some other forms of mental illness. In *Verbal Behavior*, published in 1957, Skinner followed the same strategy with language as with the operant behavior of animals. He identified concepts such as *mand* and *tact* that he believed provided a way to understand how language is used, rather than what it is.

In animal learning, many Skinnerian terms already had widely used equivalents; for example, operant behavior was referred to as instrumental behavior, conditioned reinforcement as secondary reinforcement, and contingency as dependency. The concepts in *Verbal Behavior* appeared to many critics as much the same, a cumbersome reworking of traditional notions: *mand* to mean command and *tact* to mean describe or name. But Skinner was trying to understand the function of language in a way congenial to evolutionary psychology, later popularized by Richard Dawkins and many others. From an evolutionary point of view, language exists to control the behavior of other people. Skinner tried to apply what he knew of controlling the operant behavior of animals to the interaction between a human speaker and listener.

Linguists are interested in the structure of language, not its use as a tool of control. In a well-known 1959 critical review of *Verbal Behavior*, mathematical linguist Noam Chomsky discussed how behaviorism and reinforced learning cannot explain phenomena such as how children can combine words into sentences they haven’t already heard. Although Skinner retained loyal followers, Chomsky’s review effectively marginalized radical behaviorism.

Evolution of Behaviorism

In the 1960s, behaviorism was supplanted as the dominant movement in experimental psychology by the so-called cognitive revolution. Skinner's proscription of theory and the absorption of methodological behaviorism into general empirical psychology had left radical behaviorism no place to go. But some theory-friendly offshoots soon emerged. Skinner had always argued against the idea of *internal state*, the process that intervenes between stimulus and response. But he was not totally consistent about this. In the 1948 William James Lectures on which *Verbal Behavior* is based, Skinner referred to *latent* (verbal) *responses*. Because these by definition cannot be measured, they are clearly internal in some sense. Skinner had also argued that the *operant*, his behavioral unit, is defined by classes: a stimulus class and a response class defined by their orderly functional relation. Invoking the logic of historical systems, J. E. R. Staddon extended Skinner's definition to an organism's *history*, calling his modified view *theoretical behaviorism*. A class of past histories that are equivalent in terms of the organism's future behavior is termed an internal *state* but without any physical or physiological implications. Rather than having to list the effects of all possible histories on future behavior, they can be grouped into equivalence classes, states. A particular state is then a theory that describes the common effect of a set of histories.

A simple example of such a state is *hunger*. Many histories lead to a state of hunger, such as food deprivation, certain drugs, disease, and exercise. But all lead to much the same future behavior: seeking food and being rewarded by getting food. A state need not be motivational. Consider, for example, *habituation*, which is an almost universal learning effect. As a "neutral" stimulus is repeated, its effect diminishes—the dog pricks its ears and turns in response to a novel sound, but after a few repetitions—fewer the more closely spaced they are—the sound is ignored, the response extinguishes. Then, after some time with no sound, the response may recover again. But all extinctions are not equal: Habituation will take longer to dissipate if repetitions are spaced farther apart. A simple model with not one but two memory stores, one that dissipates rapidly and the other slowly, can capture this effect. It allows us to predict how long the animal will take to recover, to dishabituate, given any history, any sequence of stimuli. In such a model, at least two numbers—*state variables*—are needed to characterize *rate sensitivity* as this is called. Simply knowing that the response has extinguished, that its "strength" is zero, is not enough.

Regarding the rejection of introspection as an explanation for behavior, Skinner’s view still prevails among radical behaviorists. Skinner denies introspection but does permit “internal stimulation,” which he invokes to explain “feeling” and “thinking.” For example, he writes in *About Behaviorism* that when we answer the question “What are you thinking? … it is … likely that we are describing private conditions associated with public behavior but not necessarily generated by it” (Skinner, 1976, pp. 30–31). What this seems to mean is that “we” are describing some internal state (but the word “state” is avoided). Skinner’s alternative is “internal stimulation,” although he does not specify what is stimulated by what. This poses the problem of how to deal scientifically with an internal stimulus that cannot be seen, measured, or postulated as part of a theory.

Methodological behaviorists, now become cognitive scientists, have not entirely avoided mentalistic explanations for behavior. Experimental psychologist David Premack (1925–2015), who performed research on both animals and human infants, began his career with a hypothesis about reinforcement of behavior (that a more frequent activity might reinforce a less frequent). But then, studying the behavior of monkeys and human infants, he proposed something called a *theory of mind* as an explanation for discriminations involving a third party. For example, a 3-year-old child is shown a Crayola box and, asked what it contains, answers “crayons.” But then the child is shown that it really contains candles. Enter “Snoopy,” a third party: “What does Snoopy think is in the Crayola box?” “Candles” says the 3-year-old. “Crayons” says a 5-year-old, with a developed theory of mind, apparently aware that Snoopy will not know the right answer. The different behavior of the 3-year-old and the 5-year-old can be explained in a variety of ways, some “cognitive” and others not. Research continues.

J. E. R. Staddon

See also ABA Designs; Applied Behavior Analysis; Cognitive Neuroscience; Replication

Further Readings

Pashler, H., & Wagenmakers, E.-J. (2012). Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives in Psychological Science, 7(6), 528–530.

Rachlin, H. (1991). Introduction to modern behaviorism (3rd ed.). New York, NY: Freeman.

Sidman, M. (1960). Tactics of scientific research: Evaluating experimental data in psychology. New York, NY: Basic Books.

Skinner, B. F. (1956 May). A case history in scientific method. American Psychologist, 11(5), 221–233.

Skinner, B. F. (1948). Verbal behavior. William James lectures, Harvard University. Retrieved from http://store.behavior.org/resources/595.pdf

Skinner, B. F. (1976). About behaviorism. New York, NY: Random House.

Staddon, J. E. R. (2014). The new behaviorism (2nd ed.). Philadelphia, PA: Psychology Press.

Staddon, J. E. R. (2016). Adaptive behavior and learning (2nd ed.). New York, NY: Cambridge University Press.

Watson, J. B. (1927, September). The myth of the unconscious: A behavioristic explanation. Harper’s Magazine, pp. 502–508.

Watson, J. B. (1948). Psychology as the behaviorist views it. In W. Dennis (Ed.), Readings in the history of psychology (pp. 457–471). New York, NY: Appleton. (Original work published 1913)

Dean R. Gerstein Dean R. Gerstein Gerstein, Dean R.

Belmont Report Belmont report

191

193

Belmont Report

The 1978 Belmont Report is a 5,000-word essay by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research that outlines basic ethical principles for the protection of human subjects in research projects. The report, titled *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, has standardized the basis for decision making by institutional review boards in the United States and influenced similar bodies around the world. This entry discusses what led up to the report, the report's development, and the principles and guidelines found in the report.

History

The National Research Act of 1974 (U.S. Public Law 93-348) established and authorized the secretary of health, education, and welfare to appoint the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, which was initially expected to complete its work in 2 years. The 11 members of the commission were charged with identifying basic ethical principles and guidelines for such research, considering the boundaries between research and the routine medical practice, the nature and definition of informed consent, the role of risk–benefit criteria in assessing human subjects research, and appropriate guidelines for selecting participants in such research.

The structure of institutional review boards was already in place at universities and biomedical organizations in the United States due to the 1966 Surgeon General's Directives on Human Experimentation, which mandated prior review by institutional committees of all research involving human subjects that was supported by the federal Public Health Service. The public disclosure and termination of the long-running Tuskegee syphilis experiment in 1972, which examined the course of syphilis in nearly 400 Black men without telling them of their diagnosis or providing penicillin, highlighted the need for stronger protections for human research subjects and precipitated the 1974 National Research Act.

The commission members were mostly university faculty in law, medicine, philosophy, and behavioral and life sciences. The group held many public meetings and a 4-day closed retreat in 1976 at the Belmont (MD) Conference Center, where the structure and core ideas of the report were developed. The report was issued in 1978 and published in the Federal Register in 1979.

Contents

The report is in three parts, beginning with distinguishing between research and practice; then outlining three fundamental ethical principles regarding the treatment of human subjects in research—respect for persons, beneficence, and justice; and finally elaborating on how the principles may be implemented. Research is defined as a departure from the practice of standard or accepted clinical therapy, designed to develop or contribute to generalizable knowledge, and usually described in a protocol that defines specific goals and procedures. Departures may be as simple as comparing the results of alternative prescriptions to different standard treatments. The report does not consider research in nonclinical fields, where this definition loses precision.

Each principle outlined in the report has two dimensions that are not entirely compatible. The principle of respect for persons focuses on self-determination or personal autonomy. The principle demands *both* that persons deemed routinely capable of self-determination enter research voluntarily and with adequate information, while those with diminished capacity for self-determination due to diverse conditions such as immaturity, disability, illness, or imprisonment be specially protected by bringing in third parties as decision makers.

The principle of beneficence is *not* to harm research subjects but to *minimize* harms while *maximizing* benefits. The admonition against direct harm is softened by permitting risk of harm, and benefits may be only to the greater good through enhanced societal knowledge.

The principle of justice refers to the fair distribution of burdens and benefits of research and a principle of equality but not absolute equality. Justice demands that the relative few who may be selected to carry the burdens of research (risks of harm) not be different as a demographic class from those who might benefit from the results. The authors cite as historical inequities the use of poor patients and of prisoners in experiments to develop therapies affordable mainly by wealthier patients or the free populace.

The authors recognize that these principles present challenges when balancing conflicting claims and making difficult choices. To assist, they elaborate on how the three principles may be implemented and certain issues resolved. The report discusses applying the principles through informed consent, risk/benefit assessment, and the selection of subjects of research.

Informed consent is the means through which respect for persons is implemented before and during research participation. Advance information about the research should include key points such as its purpose, procedures, risks, and benefits. The researcher must assure that the participant understands the information offered. Questions may be asked and must be answered truthfully. Some kinds of information may be withheld at the outset if the information would threaten the validity of the research but must be disclosed afterward. Agreement must be made free of overt or subtle coercion or undue influence (excessive or improper rewards). If participants have reduced capacity for comprehension or vulnerability to pressure, both the participant and a protective third party must give informed consent.

The principle of beneficence “requires that we protect against risk of harm to subjects and also that we be concerned about the loss of the substantial benefits that might be gained from research” (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979, n.p.). Review committees need to conduct a systematic, nonarbitrary analysis of the risks of harm as against the probability of benefits from the research. Finally, in accord with the principle of justice, there should be fair procedures and outcomes in the recruitment and selection of research subjects.

Dean R. Gerstein

See also 45 CFR Part 46; Human Subjects Protections; Human Subjects Research, Definition of; Institutional Review Boards; Nuremberg Code

Further Readings

Further Readings

The Advisory Committee on Human Radiation Experiments. (1995). The Development of Human Subject Research Policy at DHEW. In The Final Report. Retrieved May 20, 2016, from https://bioethicsarchive.georgetown.edu/achre/final/chap3_2.html

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Retrieved May 15, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

William, H. S. (1966, February). Surgeon General, Public Health Service to the Heads of the Institutions Conducting Research with Public Health Service Grants, February 8, 1966 (Clinical research and investigation involving human beings) (ACHRE No. HHS-090794-A). Retrieved May 20, 2016, from http://history.nih.gov/research/downloads/Surgeongeneraldirective1966.pdf. Reprinted with addenda in Surgeon-General's Directives on Human Experimentation. (1967). American Psychologist, 22(5), 350–355. Retrieved from http://psycnet.apa.org/doi/10.1037/h0024885

Gail Tiemann Gail Tiemann Tiemann, Gail

Benchmark Benchmark

193

193

Benchmark

A benchmark describes what a student should know and be able to do in a particular content area, grade level, or developmental level at a specified point in time. Generally, benchmarks represent shorter-term goals along a path toward mastery of content standards, learning objectives, or other longer-term educational outcomes.

Benchmarks can be used as a way to monitor student progress. At the individual level, monitoring students at various benchmarks can help students, educators, and parents make adjustments in order to help students stay on or get back on track. At an educational program level, monitoring aggregate student performance at various benchmarks can help organizations provide assistance to educators or schools in order to support student achievement.

Benchmarks can also show how much students have grown as they continue down the path toward mastery. For example, if a student is not meeting the standards at a particular benchmark but improves skills in order to meet the standards at the next benchmark, the student can be commended for showing good progress. Benchmarks can also help organizations determine program-wide progress toward goals and objectives.

To illustrate the use of benchmarks, imagine a fifth-grade student at the beginning of a school year. As the student proceeds through the math curriculum, the student’s teacher evaluates the progress of the class every 9 weeks using short assessments aligned to the content standards. After interpreting a series of assessment score reports, as well as examples of the student’s work, the teacher notices that the student is struggling to add and subtract fractions, a skill that should be mastered by that point in the school year. Noticing that a few other students were struggling in the same area, the teacher

Noticing that a few other students were struggling in the same area, the teacher revisits adding and subtracting fractions with a subset of the class. With the extra help, the student does better on the next benchmark assessment, showing positive growth on the standards related to fractions. This positive growth is shared with the student and the student’s parents at the next conference.

Continuing with the example, leadership in the student’s district reviews the benchmark assessment results for fifth-grade math, districtwide, looking for patterns. Scores are analyzed at classroom and school levels, between schools, and even disaggregated by student characteristics. In particular, results are evaluated in the context of the district’s annual goals, which included closing of achievement gaps between student groups. Noting the intermediate progress made in fifth-grade math so far this year by various student groups, the district reports the results to the local school board along with reports of other efforts to address student equity districtwide.

Gail Tiemann

See also Achievement Tests; Classroom Assessment; Formative Assessment; Progress Monitoring; Tests

Further Readings

Perie, M., Marion, S., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessment. Educational Measurement: Issues and Practice, 28(3), 5–13.

Hao Helen Zhang Hao Helen Zhang Zhang, Hao Helen

Bernoulli Distribution Bernoulli distribution

194

195

Bernoulli Distribution

The Bernoulli distribution is the range of probabilities for two possible outcomes. It is a central statistical concept. This entry describes the Bernoulli distribution and Bernoulli random variables and explains the relationship between the Bernoulli distribution and the binomial distribution.

Suppose a random experiment has two possible outcomes, either success or failure, where the probability of success is p and probability of failure is $q = 1 - p$. Such an experiment is called a *Bernoulli experiment* or *Bernoulli trial*. For a Bernoulli experiment, define a real-valued random variable X which takes two values as: $X = 1$ if success and $X = 0$ if failure. Such a random variable X is called a *Bernoulli random variable*. The probability distribution of X is given by $Pr(X = 1) = p$ and $Pr(X = 0) = 1 - p$. This distribution is called a *Bernoulli distribution*, denoted by Bernoulli(p). It is named after Jacob Bernoulli, a Swiss mathematician of the 17th century.

The term *success* here means the outcome meets some special condition, and it is not based on a moral judgment. The following are some examples of Bernoulli random variables.

> Toss a coin once. Two possible outcomes are “heads” and “tails.” Suppose heads happens with probability p, while tails happens with probability $1 - p$. Let X be a random variable such that $X = 1$ if the outcome is heads, and $X = 0$ if the outcome is tails. Then X is a Bernoulli random variable and its distribution is Bernoulli(p). When a fair coin is tossed, we have $p = q = 0.5$.
> Roll a die once. Let X be a random variable which takes two values: $X = 1$ if the Number 3 occurs, and $X = 0$ otherwise. Then X is a Bernoulli random variable. If the die is balanced, then the probability distribution of X is

Bernoulli(1/6).
In clinical trials, let *X* represent a patient’s status after a certain treatment as, *X* = 1 if a patient survives, and *X* = 0 otherwise. Then *X* is a Bernoulli random variable.

Statistical Properties

Assume a random variable *X* follows a Bernoulli(*p*) distribution. Its probability mass function is given by *P*(*X* = 1) = *p* and *P*(*X* = 0) = 1 − *p*. Equivalently, it is expressed as

P(*X* = *x*)

= *p* ^ *x*(1 − p)^{1 −*x*}, \quad*x* = 0,1.

Its expectation is *E* (*X*) = *p*, variance is Var (*X*) = *p*(1 − *p*), and skewness is \frac{1–2*p*}{\sqrt{pq}}. The moment generating function is

M _ *X*(*t*) = *E*(*e* ^ {*Xt*}) = 1 − *p* + *pe* ^ *t*.

The characteristic function is

\phi _ *X*(*t*) = 1 − *p* + *pe* ^ {*it*}.

The family of Bernoulli distributions \{Bernoulli (*p*), 0\le p\le 1 \} is an exponential family.

Estimation of *p*

Suppose we take a random sample of size *n*, *X*_1, \cdots, *X*_*n*, from Bernoulli(*p*). Then an estimator for *p* is given by the sample mean:

\bar{*X*} = \frac{1}{n} \ sum _{*i* = 1} ^ *nX* _*i*, i.e.,

\hat{*p*} = \bar{*X*}.

Because the sample mean is unbiased for the population mean *p*, we have *E*(\hat{*p*}) = *p*. By the law of large numbers, \hat{*p*} is also a consistent estimator for p. In other words, the sample proportion of successes from n experiments can consistently estimate the success probability p. The estimator $\hat{p}$ is also the maximum likelihood estimator.

Bernoulli Distribution Versus Binomial Distribution

If $X_1,\cdots, X_n$ are independent random variables, all following Bernoulli(p), then their sum $Y = \sum_{i = 1}^nX_i$ follows a binomial distribution, denoted as Binomial(n, p). The probability mass function of Y is

$$P(Y = y) = {n \choose y}p^y(1-p)^y, \quad y = 0,1, \ldots, n.$$

In other words, a sum of identical and independent Bernoulli(p) random variables is a Binomial(n, p) random variable. And a Bernoulli distribution is a special case of the binomial distribution, where $n = 1$. For example, if a coin is tossed n times, with probability p of getting a heads, then the total number of heads follows a Binomial(n, p).

Bernoulli Process

A Bernoulli process is a sequence of independent identically distributed Bernoulli trials. Formally, a {\it Bernoulli process} is a finite or infinite sequence of independent random variables X_1, X_2, X_3, $\ldots$, where each X_i is a Bernoulli trial with success probability p.

For a Bernoulli process, because the trials are independent, the process is memoryless. In other words, when p is known, past outcomes do not provide any information on future outcomes.

Binary Logistic Regression

Regression analysis is a statistical technique for estimating the relationship between a dependent variable (response) and one or more independent variables (predictors). The goal of regression analysis is to estimate the conditional expectation of the dependent variable given the independent variables, which is called the regression function. When the response variable takes only two values, either 0 or 1, *binary logistic regression* is a major regression tool for estimating the probability of the response variable based on independent variables.

Generating Random Numbers From Bernoulli(*p*)

In R software, the function *rbinom*() can be used to generate random numbers from the Binomial(*n, p*) distribution. For example, rbinom(100, 1, 0.3) generates a random sample of size 100 from the Bernoulli(*p* = .3) distribution. The functions *dbinom*(), *pbinom*(), and *qbinom*() can be used to compute the density function, distribution function, and quantile function for the binomial distribution, respectively.

Hao Helen Zhang

See also Binomial Test; Maximum Likelihood Estimation

Further Readings

Bertsekas, D., & John, N., Tsitsiklis, J. (2002). Introduction to probability. Belmont, MA: Athena Scientific.

Evans, M., Hastings, N., & Peacock, B. (2000). Bernoulli distribution. Chapter 4 in Statistical Distributions (3rd ed., pp. 31–33). New York, NY: Wiley.

Johnson N. L., Kotz, S., & Kemp, A. (1993). Univariate discrete distributions (2nd ed.). New York, NY: Wiley.

McCullagh, P., & Nelder, J. (1989). Generalized linear models (2nd ed.). Boca Raton, FL: Chapman and Hall.

Papoulis, A. (1989). Bernoulli trials. Probability, random variables, and stochastic processes (2nd ed.). New York, NY: McGraw-Hill.

Heather H. Woodley Heather H. Woodley Woodley, Heather H.

Bilingual Education, Research on Bilingual education, research on

195

198

Bilingual Education, Research on

This entry first discusses the contexts of research on bilingual education, the development of bilingual education in the United States, and early research on bilingual education. It then looks at various types of research on bilingual education and research findings on the impact and effectiveness of bilingual education. Finally, the entry describes shifts in how bilingual education is conceptualized and in the ways it is researched.

Research on bilingual education, like bilingual education itself, is shaped by sociopolitical contexts, language ideologies, and communities in action. Diverse paradigms and perspectives found in qualitative, quantitative, and mixed-methods forms are used in research on bilingual education, which is any school setting where students' instruction and assessment takes place in more than one language. Research on bilingual education may highlight one or more educational programs that fall under the umbrella of bilingual education including dual language (one way and two way), heritage language, transitional bilingual, polydirectional bilingual, developmental bilingual, and maintenance bilingual education programs.

The majority of studies on bilingual education take place in U.S. schools, from preschool to Grade 12, as students learn in Spanish and English. Other studies look at U.S. bilingual education in languages including Mandarin, French, Haitian-Creole, Russian, Arabic, Korean, Yiddish, Hebrew, American Sign Language, or indigenous languages along with English. Policies on bilingual education can be vastly differ school to school and state to state, and research often focuses on individual schools, districts, or states for this reason.

Research in bilingual education also takes place in contexts beyond U.S. schools. Canadian French–English schools and multilingual European schools, such as

Canadian French–English schools and multilingual European schools, such as those in the Basque country of Spain and in Alsace on the France–Germany border, have been the focus of research for their content and language integrated learning and developmental bilingual programs. Bilingual education in the form of heritage language or revitalization programs with indigenous languages and a colonial language (English, Spanish, or French) have been the subjects of research from the Maori schools in New Zealand to Mayan language education in Guatemalan schools.

Foundation of Bilingual Education

The historical events surrounding bilingual education set the stage for research about it and for it. In the midst of the civil rights movement, Title VII of the Elementary and Secondary Education Act of 1968, known as the Bilingual Education Act, was the first piece of federal legislation that recognized the academic needs of emergent bilinguals in schools. Prior to this recognition, students were being educated bilingually across the United States in Spanish–English and German–English classrooms. However, with World War II came the banning of German–English schooling, and except for a few strong programs in Miami and New York City, Spanish–English programs were under attack, poorly financed, and rare. Research on the history of bilingual education focuses on early community efforts, seminal court cases such as *Lau v. Nichols* (1974), the impacts and intersections of immigrant action groups, and the events of 1968 as they shaped schooling and the lives of linguistically diverse students.

Early research conceptualized differences within bilingual education, especially considering the diverse learning contexts, schools, teachers, sociopolitical environments, and linguistic experiences for students. In 1974, Wallace Lambert described two types of bilingualism within schools—subtractive and additive. Subtractive bilingualism refers to educational approaches in which children’s home language use or ability diminishes as they learn the dominant language of school. In opposition to this is additive bilingualism, occurring when a new language is added to the children’s home language, which is maintained and even strengthened. Later research in bilingual education extends these ideas to reflect the reality for some emergent bilingual students of the in-between or border spaces. This can be an instance where a young person is neither monolingual nor biliterate and brings into the classroom complex language practices.

In the late 1970s and early 1980s, Stephen Krashen developed concepts in

In the late 1970s and early 1980s, Stephen Krashen developed concepts in second-language acquisition that have been integral in shaping the groundwork of research in the fields of second-language acquisition and bilingual education. His concepts of the input hypothesis, affective filter, and the natural order hypothesis laid the foundation for how teachers, teacher–educators, linguists, and educational researchers would continue to conceptualize language and bilingual learning and teaching for decades to follow. Krashen also outlined essential components for success in bilingual education including content teaching in the home language, literacy development in the home language, and comprehensible input in English.

In 1979, James Cummins introduced the concepts of basic interpersonal communication skills and cognitive academic language proficiency (CALP) into the conversation and research on bilingual education. This paradigm of categorizing and understanding language learning has shaped many bilingual teacher education programs, and thus bilingual classrooms and students. By outlining specific benchmarks for learning both basic interpersonal communication skills and cognitive academic language proficiency in a new language, this research has also been the base for bilingual programming that maintains home language use for longer periods of time. Cummins’s research continues to build on theories of language and power, promoting equity and social justice in and through bilingual education.

The National Association for Bilingual Education, founded in 1975, and other national and local organizations promote and support research on bilingual education through events and diverse opportunities. Here, communities of scholars, educators, families, and community activists are able to share their research, collaborate with colleagues, create spaces in the field of bilingual education, and build bridges to other fields.

Evaluating Impact and Effectiveness

Much of the research on bilingual education focuses on different programs’ impacts on educational outcomes for emergent bilingual students. Related research looks at bilingual education’s impact on, or intersection with, additional factors including students’ social–emotional learning, teachers’ and students’ language use, relationships between language and culture, and family inclusion in schools. Other studies focus on an evaluation of bilingual pedagogy and assessments across multiple school districts, encompassing thousands of students, often as an experimental design. Numerous other studies focus on a students, often as an experimental design. Numerous other studies focus on a single program in one school, often zeroing in on a particular classroom or even an individual teacher.

Most studies that seek to evaluate bilingual education or compare it with other forms of second-language learning have shown positive effects of bilingual approaches to teaching and learning. The conclusion of “positive effects” has a different definition and takes on a different meaning from study to study. Some of this research highlights the cognitive impact of bilingualism for young people, while others look at gains in language learning, students’ linguistic complexity, or abilities in academic tasks. An additional body of work in this research on bilingual education sheds light on the positive impact of bilingual education on student identity, social–emotional well-being, and communities, including families’ and students’ cultural, linguistic, or religious communities.

Experimental studies place different types of bilingual education programs next to other approaches to second-language learning or, in other studies, compare one type of bilingual education to another. For example, some studies measure the effects of dual-language programs in comparison to transitional bilingual education. In these studies, which are often longitudinal studies, data are mostly drawn from student work products and assessments.

A 1997 study by Wayne Thomas and Virginia Collier, followed by a 2002 study by the same researchers, concluded that bilingual programs (specifically developmental bilingual or two-way bilingual immersion) that were strong in design and implementation had significant positive effects on students’ academic achievement including English literacy, language, and content area classes. In this research, academic and linguistic outcomes for emergent bilinguals throughout five school districts were measured and analyzed in a variety of learning settings. These reports contain research that is continuously used in support of bilingual education, specifically dual-language or maintenance bilingual education, throughout the United States.

Meta-analyses analyzing numerous studies on the effectiveness of bilingual education also contribute to the body of research on bilingual education. The conclusions of multiple large-scale studies show small but favorable impacts of bilingual education on students’ academic achievement. Some researchers have noted the importance of research design in the field of bilingual education, concluding that the more effective the experimental design, the more positive were the impacts of bilingual education.

Although research findings on bilingual education are generally supportive of multilingual pedagogical practices, some research does seek to challenge these practices. Researcher Christine Rossell has asserted that bilingual education is the least effective approach to educate immigrant children. However, many researchers have disputed this claim.

Shifts in Bilingual Education Conceptualization and Research

In a 2009 study, bilingual education as conceptualized by Ofelia García challenges traditional ideas of language learning in which bilinguals were thought to have two balanced language systems, supporting the separation of languages in schools and the notion that one language plus a second-language equals two separate languages. García calls for a reconceptualization of bilingual education to reflect bilinguals' fluid language practices. Her perspective emphasizes dynamic bilingualism and pedagogy reflective of students' multiple language practices in the classroom. Critical researchers in bilingual education are using this heteroglossic framework as a foundation to challenge power structures, oppression, and inequity in the schooling of emergent bilinguals.

Recent research on bilingual education often takes into account sociopolitical context, including the backdrops of high-stakes testing, the Common Core State Standards, and language policies, as well as rising tides of anti-immigration sentiments and neoliberalism. Participatory action research, in which research is done in collaboration with those affected by the issues being studied, has involved bilingual voices, putting the lived experiences and advocacy of emergent bilinguals in the foreground of the research.

Research in bilingual education faces new directions and new challenges as the field evolves. There has been a push to bridge theory and practice and also to bridge fields. This includes more research exploring intersections of bilingual education and special education, along with the development of anti-racist bilingual education and emphasis on community empowerment. Also, with increased attention in the education field to early childhood and initiatives such as the introduction of universal preschool in New York City, there is more demand for and more activity in research exploring bilingual education in early childhood education.

Heather H. Woodley

See also Cross-Cultural Research; Cultural Competence; Culturally Responsive Evaluation; English Language Proficiency Assessment; Second Language Learners, Assessment of

Further Readings

Baker, C. (2011). Foundations of bilingual education and bilingualism (5th ed.). Clevedon, UK: Multilingual Matters.

Cummins, J. (2000). Language, power and pedagogy: Bilingual children in the crossfire. Clevedon, UK: Multilingual Matters.

García, O. (2009). Bilingual education in the 21st century. A global perspective. Malden, MA: Wiley-Blackwell.

García, O., Zakharia, Z., & Otcu, B. (Eds.). (2013). Bilingual community education and multilingualism: Beyond heritage languages in a global city. Bristol, UK: Multilingual Matters.

Krashen, S. D. (1996). Under attack: The case against bilingual education. Culver City, CA: Language Education Associates.

Krashen, S., & McField, G. (2005, November/December). What works? Reviewing the latest evidence on bilingual education. Language Learner, 1(2), 7–10, 34.

Reyes, S. A., & Kleyn, T. (2010). Teaching in two languages: A guide for K–12 educators. Thousand Oaks, CA: Corwin Press.

Rossell, C. H., & Kuder, J. (2005). Meta-murky: A rebuttal to recent meta-analyses of bilingual education. In J. Söhn (Ed.), The effectiveness of bilingual school programs for immigrant children (pp. 43–76). Berlin, Germany: Programme on Intercultural Conflicts and Societal Integration (AKI) at the Social Science Research Center Berlin (WZB). Retrieved from https://www.bu.edu/polisci/files/2009/09/Meta-Murky-A-Rebuttal-to-Recent-Meta-Analyses-of-Bilingual-Education.pdf

Thomas, W., & Collier V. (2002). A national study of school effectiveness for language minority students' long term academic achievement. Santa Cruz, CA: Center for Research on Education, Diversity … Excellence (CREDE).

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. Review of Educational Research, 55, 269–317.

Michele F. Zimowski Michele F. Zimowski Zimowski, Michele F.

BILOG-MG

BILOG-MG

198

202

BILOG-MG

BILOG-MG is a software program for the development, analysis, scoring, and maintenance of educational and other measurement instruments within the statistical framework of item response theory (IRT). As a tool for applying IRT to practical testing problems, the program is concerned with estimating the characteristics of the items in an instrument (the item parameters) and the standing or position of respondents on the underlying attribute or latent trait the items are intended to measure (the person parameters or scale scores). The program is specifically designed for the analysis of item responses classified into two categories (i.e., dichotomously scored or binary items) and offers a wide range of options for fitting IRT models to item response data of that type. This entry describes the program’s capabilities, the models and estimation procedures it implements, and the types of applications it accommodates.

Overview of the Program’s Features and Capabilities

Housed within a Windows graphical point-and-click interface, BILOG-MG is designed for the IRT analysis of instruments comprising dichotomously scored sets or subsets of items intended to measure a single underlying attribute or latent dimension. As an extension of the BILOG program of Robert J. Mislevy and R. Darrell Bock to multiple groups of respondents, the program accommodates a broad range of practical applications that involve one or more than one group of respondents and one or more than one test form (version) of an instrument. The program offers an array of options for estimating the parameters of the items in an instrument, the scale scores of persons completing it, and the latent distributions of the groups or populations represented in the data. It also provides numerous indices and plots to inform and guide the development of instruments with good measurement properties.

Models for Dichotomously Scored Items

As a program specifically designed for the IRT analysis of dichotomously scored items, BILOG-MG relies on binary logistic functions to model the relationship between the characteristics of an item and the probability that a person with a given level of the underlying trait (typically denoted as θ) will respond to the item in one of two predefined categories. The categories may represent correct and incorrect responses to multiple-choice problems on a test of educational achievement, the presence or absence of symptoms recorded on a checklist of characteristics associated with a particular medical condition, or some other binary classification of the responses to the items in an instrument. The latent trait measured with the items may be verbal proficiency, spatial ability, generalized anxiety, or any number of other underlying attributes that an individual may possess. In educational applications, the underlying trait often represents some form of cognitive proficiency measured by correct and incorrect responses to a set of multiple-choice or short-answer questions. The discussion that follows frames the program’s features and models in those terms, referring to the underlying trait as proficiency and denoting the probability that a person with proficiency θ will respond to item *j* with a correct response ($xj = 1$) as $P(xj = 1\theta = Pj\theta)$.

BILOG-MG implements three binary logistic functions for the IRT analysis of dichotomously scored items, the one-, two-, and three-parameter models. The names indicate the number of item parameters in each model. The two-parameter model, for example, expresses the probability of a correct response to item *j* as a function of a person’s proficiency and two parameters specific to item *j* that must be estimated from the data:

$$Pj\theta = 1(1 + e - aj\theta - bj).$$

The *aj* parameter represents the slope or discrimination power of the item. It indicates the extent to which an item discriminates among individuals with higher and lower levels of proficiency. Items with higher values of *a* are more effective in differentiating among individuals than items with lower values of *a*. The *bj* parameter represents the difficulty or threshold parameter of the item. It indicates the position or location of the item on the θ scale of proficiency. Items with higher positions on the scale are more difficult than items with lower positions on the scale. In the two-parameter model (as well as in the one-parameter model), the threshold parameter of an item is located at the point on the scale where a person with that scale score has a .5 probability of answering the item correctly.

The simplest of the three models, the one-parameter model, also known as the Rasch model, assumes that the items are equally discriminating. In other words, the *a* parameter is the same for all items and a person’s probability of answering an item correctly simply depends on the difficulty level of that item and person’s level of proficiency. The least restrictive of the models, the three-parameter model, adds a parameter to the two-parameter model to take into account the effect of guessing on responses to an item. It is commonly used in the analysis of multiple-choice items where a respondent may answer an item correctly simply by chance.

Models for Multiple-Group and Multiple-Form Applications

By necessity or by design, measurement instruments often consist of more than a single test form. In educational applications, the forms might represent different versions of an instrument developed over time to prevent overexposure of the item content or to satisfy item disclosure requirements or correspond to age-or grade-specific versions of an instrument developed to monitor the educational achievement of children as they progress through school. In these cases, and whenever an instrument consists of more than one form, the forms must be equated for the scores to have the same meaning across forms. In IRT, it means placing the item parameter estimates from each form on a common scale. Various procedures that involve converting estimates from separate IRT analyses of the forms are used for that purpose.

Beyond simply carrying out a separate calibration of each form, BILOG-MG performs equivalent and nonequivalent groups equating in a single IRT analysis of the data from all test forms. When the groups completing each form are random samples of respondents from the same population (equivalent groups), it treats the forms as if they were one test administered to a single population and performs a conventional IRT analysis. When the groups completing each form are composed of respondents from different populations or from different subgroups within a population (nonequivalent groups), the program places the subgroups within a population (nonequivalent groups), the program places the items on a common scale with a multiple-group model that takes into account differences among the latent distributions of the groups as it estimates the parameters of the items. The estimation procedure allows for the simultaneous estimation of the latent distributions and the item parameters makes the program suitable for a wide range of practical applications that involve more than one group of respondents and one or more than one test form.

In estimating the item parameters of the one-, two-, and three-parameter models, the multiple-group models assume that the item response function for any given item is the same across all groups of respondents, except in applications of differential item functioning and item parameter drift over time. The differential item functioning and item parameter drift models allow the difficulty of the items to vary from group to group or from occasion to occasion to test for and identify Item × Subgroup interactions (differential item functioning) and Item × Time of Testing interactions (item parameter drift).

Estimation of the Item Parameters and Latent Distributions of Proficiency

To obtain estimates of the item parameters and the latent distributions of proficiency, BILOG-MG relies on the marginal maximum likelihood method proposed by Bock and Murray Aitkin and its extension to multiple groups of respondents detailed by Bock and Michele Zimowski. The procedure provides for the simultaneous estimation of the item parameters and the latent distribution or distributions of proficiency when there are multiple groups of respondents. Except in special situations, the marginal maximum likelihood procedure assumes that the response to a particular item is independent of the responses to other items in the test for all persons with the same level of proficiency (i.e., the assumption of conditional independence). The procedure also assumes that respondents in each group are drawn from some population in which the latent distribution of proficiency has a specified shape.

To start the estimation procedure, the user must specify the shape of the latent distribution of each group represented in the item response data. BILOG-MG offers several options for that purpose, including the program default of a normal distribution. When the assumption of a normal distribution seems untenable, the user has the option of specifying the shape of the distribution, keeping it fixed at its initial specification, or estimating it directly from the patterns of correct and incorrect responses along with the item parameters in the iterative estimation procedure.

In applications consisting of a single group of respondents, the program sets the mean and standard deviation of the latent distribution to zero and one to resolve the indeterminacy in the origin and unit of the latent distribution of proficiency. In applications involving more than one group of respondents, the user may choose to resolve the indeterminacy by setting the mean and standard deviation of the combined distributions of all groups to zero and one or by selecting one group as the reference group and setting the mean and standard deviation of its distribution to zero and one. Depending on the option selected, the means and standard deviations of the groups are set relative to the reference group or relative to the mean and standard deviation of the combined groups.

The estimation procedure generates marginal maximum likelihood estimates of the item parameters, except when the user chooses to impose prior distributions on the item parameters, in which case it generates marginal maximum a posteriori estimates. Depending on the model selected, the program generates estimates of the slope (a), threshold (b), intercept ($-a \times b$), lower asymptote (guessing parameter), and a one-factor item loading for each item included in the analysis, along with the respective standard errors. It also provides estimates of the means and standard deviations of the latent distributions of proficiency of the groups.

The program generates several indices for assessing the fit of a model to the item response data. When all or nearly all response patterns are present in the data, the program computes a likelihood ratio chi-square statistic for testing the overall Goodness-of-Fit of the model to the data. When that statistic is not available, the change in the negative of the marginal log likelihood between the one-and two-parameter models and the two-and three-parameter models can be used to assess whether adding parameters to the item response model improves fit. When a test consists of more than 20 items, the program generates an approximate chi-square test of item fit for each item in a test.

Estimating Scale Scores and Evaluating the Functioning of the Instrument

BILOG-MG computes three types of IRT scale scores or estimates of

BILOG-MG computes three types of IRT scale scores or estimates of proficiency—maximum likelihood, Bayes’s or expected a posteriori, and Bayes’s modal or maximum a posteriori estimates. The user has the option of generating the estimates in the scale of the item parameters or rescaling them to another metric with a linear transformation or with respect to the location and scale of the scale score estimates in the sample. If expected a posteriori estimates are selected, the user also has the option of specifying the prior distribution to be used in their estimation and of rescaling the estimates with respect to the location and scale of the latent distribution.

For evaluating properties of the scale scores, the program computes the first four moments of the scale score distribution and an estimate of empirical reliability based on the IRT scale score variance and mean square error. For evaluating the properties of the individual items and the instrument as a whole, it plots item, test, and test-form information curves and computes theoretical reliabilities based on the item parameters, assuming normal latent distributions of proficiency.

Availability of the Program

BILOG-MG may be purchased from the website of Scientific Software International. SSI distributes the program electronically. For those who simply wish to examine the program, a free trial version is available for inspection for up to 15 days after the program is downloaded.

Michele F. Zimowski

See also Conditional Independence; Differential Item Functioning; Equating; Item Response Theory; Marginal Maximum Likelihood Estimation; Prior Distribution; Rasch Model

Further Readings

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6,

431–444.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 433–448). New York, NY: Springer.

du Toit, M. (Ed.). (2003). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, IL: Scientific Software International.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359–381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177–195.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57–75.

Websites

Scientific Software International: http://www.ssicentral.com/

John M. Ferron John M. Ferron Ferron, John M.

Seang-Hwane Joo Seang-Hwane Joo Joo, Seang-Hwane

Binomial Test

Binomial test

202

205

Binomial Test

Binomial experiments consist of a series of two or more independent trials, where each trial in the series results in one of two outcomes: a success or a failure. The purpose of the binomial test is to determine for such experiments whether the number of observed successes warrants rejection of an assumed probability of success, π. For example, a gambler may posit that the probability of getting a head in a flip of a coin is .5. A binomial test could be used to determine whether the number of heads observed in a series of independent flips warrants rejection of that hypothesis. This entry describes educational research applications, states the hypothesis and assumptions, defines and illustrates the exact probability computations, defines and illustrates the normal theory approximation, and discusses the consequences of violating the independence assumption.

Educational Research Applications

An educational researcher may believe that the probability of a child answering multiple-choice questions (with four options) correctly on a chemistry pretest is .25. That hypothesis could be tested using the binomial test, which would consider the number of successes (i.e., correctly answered questions) in a series of independently administered pretest questions. Another educational researcher may want to test the hypothesis that there is a .50 probability that an elementary school principal will support a newly proposed district policy. If so, principals could be independently sampled and interviewed to determine whether they supported the policy. A binomial test could be used to determine whether the number of successes (i.e., observed supporters) was (or was not) sufficient to reject the hypothesis that the probability of support was .50.

As a final example, consider an educational researcher who is interested in whether an intervention would increase the prosocial behavior of children with behavioral and emotional disturbances. The researcher could hypothesize that the probability of observing an increase in prosocial behavior for a child was .50. A binomial test could be used to determine whether the number of successes (i.e., number of children with observed improvements) in an independent sample was sufficient to reject the null hypothesis that the probability was .50. This final application of the binomial test would often be referred to as a sign test because it is based on counting up the number of positive and negative signed differences.

Hypothesis and Assumptions

The binomial test allows us to test the null hypothesis that the probability of success (π) is equal to some researcher specified value *a*. For a nondirectional test, the null hypothesis is H_0: $\pi = a$, and for a directional test, the null hypothesis is either H_0: $\pi \leq a$ or H_0: $\pi \geq a$. The probability calculations, which are based on the binomial distribution, assume:

1. Observations are sampled from a binary population (i.e., there are only two possible values for each observation, a success or a failure).
2. Each observation is independent implying that it is not affected by any of the other observations.
3. The probability of a success is fixed for the population.

Binomial Probabilities

For a binomial experiment where the probability of success in any one trial is π, the probability that there will be *r* successes in *n* trials is computed as:

$$P(r) = {}_{n}\mathrm{C}_{r} \times p^{r} \times (1-p)^{(n-r)} \,,$$

where is the number of combinations of *n* things taking *r* at a time where .

Illustration

Suppose a science education researcher is anticipating a student will not have the prerequisite knowledge to answer chemistry questions prior to instruction and thus expects that on a multiple-choice pretest the item responses from the student would be random guesses. For a pretest that consists of 5 multiple-choice items, each with 5 options, the researcher is hypothesizing that the probability of successfully answering a question is .20. To test the null hypothesis that H_0: $\pi = .20$ at an α level of .05, the researcher could first compute the probability that a student answers 0, 1, 2, 3, 4, or 5 questions successfully using the binomial probability formula.

$$\begin{aligned} p(0) &= {}_nC_r \times p^r \times (1-p)^{(n-r)} \\ &= 5!/(0!\times 5!)\times .20^0 \times .80^5 \\ &= .32768. \end{aligned}$$

$$\begin{aligned} p(1) &= {}_nC_r \times p^r * (1-p)^{(n-r)} \\ &= 5!/(1!\times 4!)\times .20^1 \times .80^4 = .4096. \end{aligned}$$

$$\begin{aligned} p(2) &= {}_nC_r \times p^r \times (1-p)^{(n-r)} \\ &= 5!/(2!\times 3!)\times .20^2 \times .80^3 = .2048. \end{aligned}$$

$$\begin{aligned} p(3) &= {}_nC_r \times p^r \times (1-p)^{(n-r)} \\ &= 5!/(3!\times 2!)\times .20^3 \times .80^2 = .0512. \end{aligned}$$

$$\begin{aligned} p(4) &= {}_nC_r \times p^r \times (1-p)^{(n-r)} \\ &= 5!/(4!\times 1!)\times .20^4 \times .80^1 = .0064. \end{aligned}$$

$$\begin{aligned} p(5) &= {}_nC_r \times p^r \times (1-p)^{(n-r)} \\ &= 5!/(5!\times 0!)\times .20^5 \times .80^0 = .00032. \end{aligned}$$

Note that if we sum these probabilities, we get 1.0 because the student had to answer 0, 1, 2, 3, 4, or 5 questions correctly. Next, to conduct the binomial test, we add the probability corresponding to the observed number of successes to all probabilities that correspond to numbers of successes that are as far or farther from what was hypothesized. If the student answered 4 items successfully, we would compute the probability of successfully answering four or more questions as .0064 + .00032 = .00672. Because this probability is less than our α of .05, we would reject the null hypothesis that π = .20, which in this context suggests that the student did not randomly guess but had some of the needed prerequisite knowledge.

Normal Distribution Approximation

When the number of trials is relatively small ($n < 25$), it is feasible to evaluate the exact binomial probability for each successful occasion. However, when there is a large number of trials, it could be tedious for researchers to compute the exact binomial probabilities for every possible number of success. With larger n, especially when the probability of success is close to 0.5, the binomial test could be alternatively conducted using the normal approximation approach. The normal probability approximation approach can be applied by simply evaluating the mean and standard deviation of the binomial distribution and then substituting these values into the Z score transformation formula. The computation of normal variate Z for the normal approximation is as follows:

$$Z = X - npnp\left(1 - p\right),$$

where X is the number of success, np is the mean of the binomial probability distribution, and $np\ (1-p)$ is the standard deviation of the binomial probability distribution.

Because the binomial distribution is discrete and the normal distribution is a continuous distribution, a continuity adjustment can be applied as well.

$$Z = Xa - npnp\left(1 - p\right),$$

where Xa is adjusted number of success for the discrete number of success X, such that $Xa = X - 0.5$ for a lower bound or $Xa = X + 0.5$ for an upper bound.

Note that with a large number of trials ($n > 25$), the Z is assumed to distribute as a normal distribution with 0 mean and 1 standard deviation (i.e., standard normal distribution). Once the Z score of the normal approximation for the binomial probability is computed, then the probability of *X* successes out of *n* trials can be calculated using the standard normal distribution probability. An illustration of the binomial test with the normal probability approximation is given next.

Continuing from the previous example, suppose that a pretest consists of 20 multiple-choice items, instead of 5 items, and each item has five options. The researcher anticipates that for students who did not have the prerequisite knowledge to answer the pretest, the probability of successfully answering a question correctly is still 0.2. To test the null hypothesis that H_0: π = .20 at an α level of .05, the researcher could compute the probability that the student answers 0, 1, 2, 3, and up to 20 questions successfully with the hypothesized probability 0.2 (or guessing probability). This computation with the binomial probability distribution can be time-consuming and tedious, but the normal approximation for the binomial distribution can be readily applied. If the student answered 8 items correctly, then the researcher would compute the probability of successfully answering 8 or more questions as follows.

$$Z = 8 - (20)(0.2)(20)(0.2)(0.8)$$

$$= 2.24 \text{ without correction for the continuity}$$

or

$$Z = 7.5 - (20)(0.2)(20)(0.2)(0.8)$$

$$= 1.96 \text{ with correction for the continuity.}$$

Note that *Xa* is *X* − 0.5 because the number of success is at least 8 for the discrete probability (binomial distribution), and thus the continuous probability (normal distribution) would include all values that would round to a value of 8 or higher. Then,

$$p(X > 8) =$$

$$p(Z > 2.24)$$

$$= .0127 \text{ without correction for the continuity.}$$

or

$$\begin{aligned} &p(X > 7.5) \\ &= p(Z > 1.96) \\ &= .0252 \text{ with correction for the continuity.} \end{aligned}$$

Because the probability is less than our α of .05, we would reject the null hypothesis that $\pi = .20$, which suggests that the student did not randomly guess but had prerequisite knowledge.

Consequences of Violations of the Independence Assumption

As noted previously, the binomial test is based on the assumption of independence. This entry next reviews a couple applications of the binomial test in educational research where independence could be questioned, and as a consequence, the validity of the binomial test could be challenged. In the context of analyzing single-case studies, researchers may estimate a trend line during the baseline phase and then compare the observations in the treatment phase to an extension of the baseline trend line. The binomial test was considered as a method of testing whether the proportion of observations in the intervention phase that exceeded the extrapolated baseline trend was greater than .50, which in turn would indicate a treatment effect. John Crosbie used simulations to show that the Type I error rate of the binomial test was substantially affected by autocorrelation (nonzero serial correlation), and thus the binomial test was not valid for this application.

Anthony Onwuegbuzie, Joel Levin, and John Ferron considered contexts where researchers examine differences between groups on a series of measures. A binomial test was considered as a method for testing whether the number of signed mean differences (e.g., $M_{\text{Tx}(i)} - M_{\text{Control}(i)}$ for variable i) was sufficient to reject the null hypothesis that the probability of the mean difference being positive was .50. They showed that when the variables being examined were correlated, the signed differences were not independent and that the binomial test failed to control the Type I error rate unless corrections were made for the dependency. In short, these studies show that when the trials in the binomial experiment are not independent of each other, the statistical validity of the binomial test is compromised.

John M. Ferron and Seang-Hwane Joo

See also Autocorrelation; Bernoulli Distribution; Maximum Likelihood Estimation; Normal Distribution; Single-Case Research; Type I Error

Further Readings

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. Behavioral Assessment, 9, 141–150.

Onwuegbuzie, A. J., Levin, J. R., & Ferron, J. M. (2011). A binomial test of group differences with correlated outcome measures. Journal of Experimental Education, 79, 127–142.

Ware, W. B., Ferron, J. M., & Miller, B. M. (2013). Introductory statistics: A conceptual approach using R. New York, NY: Routledge.

Ben P. Hunter Ben P. Hunter Hunter, Ben P.

Ryan W. Schroeder Ryan W. Schroeder Schroeder, Ryan W.

Kelli L. Netson Kelli L. Netson Netson, Kelli L.

Bipolar Disorder Bipolar disorder

Bipolar Disorder

Bipolar disorders (formerly known as manic depressive illnesses) are a set of mood disorders in which patients experience phases or cycles of mood symptoms that create clinically significant impairment in daily functioning. Bipolar disorders are distinguished from unipolar depression by the inclusion of cycles that consist of unusually high, overly joyful, expansive, or irritable moods, denoted as manic episodes.

According to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, bipolar disorders were classified as bipolar I disorder, bipolar II disorder, cyclothymic disorder, substance/medication-induced bipolar disorder, and bipolar disorder due to another medical condition. The shared link of all of these disorders is the presence of episodes that include abnormally severe elevated and depressed moods. The primary difference between these specific disorders is the duration of episodes, timing, course, and etiology. The disorder, in all forms, is estimated to affect as much as 3.9% of the U.S. population.

In the early 20th century, German psychiatrist Emil Kraepelin studied the course of bipolar disorder and provide detailed descriptions of the condition in adults. Since that time, descriptions and conceptualizations have remained relatively consistent, and the focus of the disorder continues to be on adult-onset cases. Although the average age of onset for bipolar disorder is 18 years, symptom onset can vary and there has recently been an increased focus on adolescent-and childhood-onset forms.

Prevalence rates among adolescents are fairly similar to those observed among adults (1–2%), whereas prevalence rates among children are not well established. Despite increased attention to early-onset cases, common diagnostic criteria continue to focus on symptoms in adults. As a result, the diagnostic criteria are often challenging to apply to children. Strict adherence to the adult-based set of diagnostic criteria in children may miss some young individuals with bipolar disorder who have developmentally different symptoms, whereas extremely liberal application of the criteria may result in over diagnosing children with typical mood swings or other behavioral difficulties.

Due to episodic mood, behavior, energy, and sleep disturbances, there may be considerable social and educational challenges for children and adolescents who develop bipolar disorder with more significant impact with earlier onset. Due in part to the historic view of bipolar disorder as an adult disorder, treatment for bipolar disorders in youths has typically mirrored treatment for adults. Initial treatment is focused toward mood stabilization and behavior management. In addition, though, treatment often incorporates use of collaborative interventions that optimize family and educational strengths.

Students who have been diagnosed with bipolar disorder may qualify for special instruction and academic accommodations from an individualized education program or Section 504 plan, which allow for accommodations under the “emotional disability” or “other health-impaired” exceptionalities. With adequate supports at home and at school and from medical providers, many young people diagnosed with bipolar disorder may develop appropriate strategies to lead productive and educationally successful lives.

Ben P. Hunter, Ryan W. Schroeder, and Kelli L. Netson

See also *Diagnostic and Statistical Manual of Mental Disorders*; Individualized Education Program; Self-Regulation

Further Readings

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorder (5th ed.). Washington, DC: Author.

Goodwin, F. K., & Jamison, K. R. (2007). Manic-depressive illness: Bipolar disorders and recurrent depression (2nd ed.). New York, NY: Oxford

University Press.

Grier, E. C., Wilkins, M. L., & Pender, C. A. S. (2007). Bipolar disorder: Educational implications for secondary student. Principal Leadership, 7(8), 12–15.

Rachel Darley Gary Rachel Darley Gary Gary, Rachel Darley

Bloom’s Taxonomy

Bloom’s taxonomy

206

210

Bloom’s Taxonomy

Bloom’s taxonomy is a multitiered model of classifying expected or intended educational learning objectives according to cognitive levels of complexity and mastery. Initially developed during the 1950s and later named after the American educational psychologist Benjamin S. Bloom, the model is concerned with the cognitive or thinking domain of learning. This entry describes both the original Bloom’s taxonomy and the revised version, which also classifies learning objectives by the types of knowledge used in thinking. The entry discusses specific changes made in terminology, structure, and emphasis in the revised version of the taxonomy and discusses the practical application of the taxonomy to the educational setting.

Bloom sought to provide a logical, progressive model that identified and classified all cognitive educational outcomes from simple to complex. Recognized by educational researchers and practitioners alike as an effective empirical model for measuring the cognitive or thinking domain of learning, Bloom’s taxonomy endures as a widely applied and taught framework across PreK–12 and higher education contexts.

Original Bloom’s Taxonomy

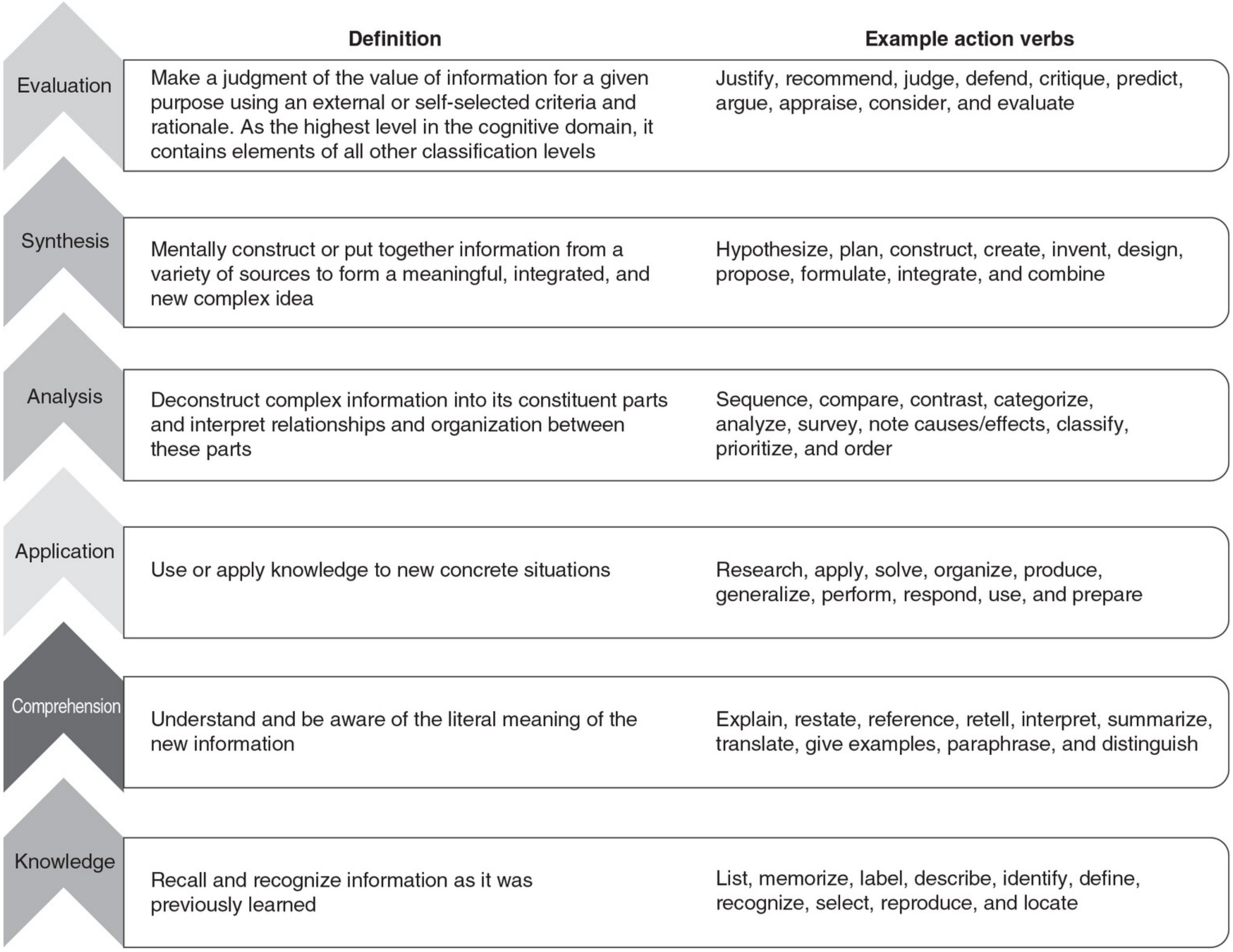
Bloom initiated the idea of creating a theoretical model of learning that sought to identify and classify all educational objectives during discussions that took place at the 1948 Convention of the American Psychological Association. Intended for university academics, Bloom hoped that in doing this work he could aid academics in reducing duplicate or redundant test items measuring the same educational learning objectives. Eight years later, Bloom and his colleagues followed through on his initial idea, identifying three domains of educational learning:

Cognitive: knowledge or thinking
Affective: attitude or self
Psychomotor: manual or physical skills

First published in 1956 under the title *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*, the cognitive domain of learning was later renamed Bloom’s taxonomy after Bloom as the model’s primary developer. In education, the prime focus has been on Bloom’s cognitive domain of learning.

Bloom’s taxonomy is a hierarchical, six-tiered model of classifying thinking based on specific cognitive levels of complexity, starting from the simplest to most complex. The original six classification levels of the taxonomy are (1) *knowledge*, (2) *comprehension*, (3) *application*, (4) *analysis*, (5) *synthesis*, and (6) *evaluation*. The taxonomy was created to assist both educational researchers and practitioners to understand the fundamental, step-by-step process in which people develop and attain new knowledge and intellectual skills. In other words, the lower classification levels of the taxonomy must be understood and mastered before progressing to the next. Figure 1 presents the taxonomy with definitions and example action verbs for each classification level.

Figure 1 Bloom’s taxonomy of educational learning objectives



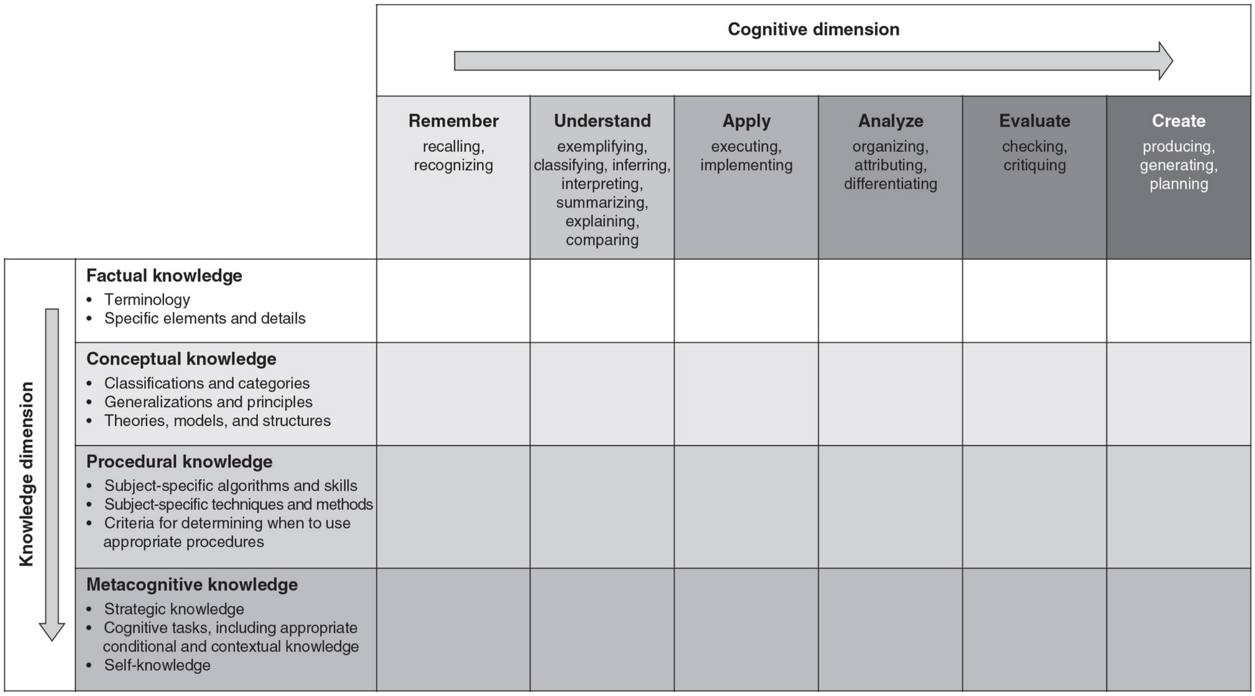
	Definition	Example action verbs
Evaluation	Make a judgment of the value of information for a given purpose using an external or self-selected criteria and rationale. As the highest level in the cognitive domain, it contains elements of all other classification levels	Justify, recommend, judge, defend, critique, predict, argue, appraise, consider, and evaluate
Synthesis	Mentally construct or put together information from a variety of sources to form a meaningful, integrated, and new complex idea	Hypothesize, plan, construct, create, invent, design, propose, formulate, integrate, and combine
Analysis	Deconstruct complex information into its constituent parts and interpret relationships and organization between these parts	Sequence, compare, contrast, categorize, analyze, survey, note causes/effects, classify, prioritize, and order
Application	Use or apply knowledge to new concrete situations	Research, apply, solve, organize, produce, generalize, perform, respond, use, and prepare
Comprehension	Understand and be aware of the literal meaning of the new information	Explain, restate, reference, retell, interpret, summarize, translate, give examples, paraphrase, and distinguish
Knowledge	Recall and recognize information as it was previously learned	List, memorize, label, describe, identify, define, recognize, select, reproduce, and locate

Bloom's levels define the steps in development of thought, and each level increases in cognitive difficulty. As such, educators often interpret the levels of the taxonomy as climbing a staircase of cognitive complexity, from lower to higher ordered thinking.

Revised Bloom's Taxonomy

In 1995, a former student of Bloom, Lorin W. Anderson, and David R. Krathwohl, a member of the academic team that developed the original Bloom's taxonomy, assembled and led a team of cognitive psychologists, teacher educators, curriculum specialists, and educational researchers in revising the taxonomy to more accurately represent 21st century teaching and learning. In 2001, Anderson and Krathwohl published their work titled *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Intentionally designed to assist educators in understanding and implementing standards-based curricula, the revised Bloom’s taxonomy presents a two-dimensional model focused on both cognitive and knowledge processes. Figure 2 presents the revised Bloom’s taxonomy in its most frequently depicted table or matrix form and includes the subcategories of levels for both cognitive and knowledge processes.

Figure 2 Revised Bloom’s taxonomy of educational learning objectives



		Cognitive dimension					
		Remember recalling, recognizing	**Understand** exemplifying, classifying, inferring, interpreting, summarizing, explaining, comparing	**Apply** executing, implementing	**Analyze** organizing, attributing, differentiating	**Evaluate** checking, critiquing	**Create** producing, generating, planning
Knowledge dimension	**Factual knowledge** • Terminology • Specific elements and details						
	Conceptual knowledge • Classifications and categories • Generalizations and principles • Theories, models, and structures						
	Procedural knowledge • Subject-specific algorithms and skills • Subject-specific techniques and methods • Criteria for determining when to use appropriate procedures						
	Metacognitive knowledge • Strategic knowledge • Cognitive tasks, including appropriate conditional and contextual knowledge • Self-knowledge						

Source: Adapted from Tables 3.1 (p. 28), 3.2 (p. 29), and 3.3 (p. 31) and Figure 3.1 (p. 32), Anderson … Krathwohl (2001).

Specific Changes

When deciding on what changes should be made to the original taxonomy, Anderson and Krathwohl’s team not only considered their own expertise but also considered the critiques and concerns expressed about the model by Bloom himself. The revised Bloom’s taxonomy includes changes made to the terminology, structure, and emphasis as compared to the original taxonomy.

Terminology

Perhaps the most obvious differences between the two models are the changes in terminology. To reflect a more active form of thinking, the revised Bloom’s taxonomy changed the names of the original six cognitive classification levels from noun to verb forms. Moreover, the lowest level of the original taxonomy, *knowledge*, was renamed to *remember* as were *comprehension* and *synthesis* renamed to *understand* and *create*.

Structure

Structure

In addition to making changes in terminology, Anderson and Krathwohl made alterations to its structure. The decision to switch the placement of the highest two levels of the taxonomy, *evaluation* (*evaluate*) and *synthesis* (*create*), represents the authors’ assertion that learners’ cognitive ability to evaluate came before their ability to synthesize or create. Additionally, they developed a separate *knowledge dimension* of the taxonomy that defined four classification levels of knowledge used in cognition. The levels of the *knowledge dimension* are (1) *factual*, (2) *conceptual*, (3) *procedural*, and (4) *metacognitive*. As such, the revised Bloom’s structure is a two-dimensional model and is typically depicted in matrix form that identifies the types of knowledge to be learned (*knowledge dimension*) and the processes used to learn (*cognitive dimension*) these types of knowledge. Furthermore, the classification levels of both dimensions are also subdivided into either three or four categories in the *knowledge dimension* and three to eight categories in the *cognitive dimension*.

Emphasis

Finally, the revised taxonomy is intended for a much broader audience than the narrow higher education purpose of the original. Rather, these changes emphasize the revised model as a more useful and authentic tool to guide all educators in curriculum and assessment development and instructional planning and delivery. Additional emphasis placed on description and explanation of the dimensions’ subcategories in the revised taxonomy aim to provide more coherence to the model.

Application of Bloom’s Taxonomy to the Educational Setting

In both its original and revised forms, Bloom’s taxonomy is used by educators across all levels and subjects to describe the degree to which they want students to know, understand, and use concepts. Bloom’s taxonomy provides educators with a common vocabulary for developing comprehensive lists of educational learning objectives for classroom instruction representative of the breadth and depth of all cognitive and knowledge processes. In doing so, Bloom’s taxonomy supports the alignment of student learning objectives with curriculum, instruction, and assessment. Applying the model can involve use of classification level verbs to plan and structure questions as part of daily lesson planning to promote higher order thinking in students. The model also can be used to create a table of specifications to design assessments in order to ensure a representative sample of assessment items across all levels of the thinking and knowledge dimensions.

Rachel Darley Gary

See also Backward Design; Critical Thinking; Goals and Objectives; Learning Theories; Metacognition; Zone of Proximal Development

Further Readings

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives. New York, NY: Longman.

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, NY: David McKay Company.

Hoy, A. W., Stinnett, A. M., Fernie, D. E., O’Sullivan, M., & Gabel, S. E. (2002). Revising Bloom’s taxonomy. Theory Into Practice, 41(4), 212–218.

Gail Tiemann Gail Tiemann Tiemann, Gail

Body of Work Method Body of work method

210

212

Body of Work Method

Body of work is a methodology used for standard setting. Broadly, standard setting is a process used to determine minimally acceptable scores of an assessment. Originally developed by Stuart Kahl, Timothy Crockett, Charles DePascale, and Sally Rindfleisch, the body of work method is generally used for setting standards for assessments that include, but are not limited to, constructed response tasks. Because examples of real student work are the heart of the body of work method, the process is considered an examinee-centered method rather than a test-centered method, which would focus more on the test items themselves. This entry discusses the work leading up to standard setting using the body of work method, what happens during the process, and the evidence that is collected to support the results generated by this method.

During the body of work process, panels of individuals with deep knowledge of the target content area convene to review examples of student work that have been previously scored. After training, the task of each panelist is to match the characteristics of that work to performance levels and extended descriptions of the student knowledge and skills required at each performance level. As panelists iteratively review the student work, sorting the work into performance levels, the scores that divide the performance levels, called cut scores, are determined.

Precursors to Standard Setting

Before standard setting, performance-level descriptors are written, describing what students should know and be able to do at different levels. A performance level could be labeled by a number or a phrase, such as Level 5, Level 4, or Level 3 or advanced, proficient, or needs improvement. The performance levels and descriptors themselves may be written by the governing body of the assessment (e.g., state education agency) along with the assessment developers and then reviewed and revised by content experts. One of the levels may be deemed as the minimum “passing” level. For example, a Level 3 might be the level that a student must achieve to be “meeting the standards” in a particular content area.

Also prior to standard setting, the assessment is administered to students and the results analyzed and scored according to the guidelines established by the test developers. For the body of work method, actual examples of student responses at all possible score points are pulled from the entire population of completed tests. The samples of student work selected for standard setting may be double scored to ensure that the standard-setting event is based on reliable test results.

Preparation for Standard Setting

The general steps that lead up to a standard setting event involve recruiting subject matter experts to serve as panelists, arranging for panel facilitators, and preparing the materials needed during the meeting itself. Standard setting panelists should be subject matter experts in the content covered by the assessment and have experience with student work in the content area. Generally, panelists are selected to represent a wide variety of experiences, population characteristics, and geographical regions.

The facilitators have experience working with groups and are specifically trained in the body of work method. The facilitator’s role is to guide panelists through the standard setting tasks, adhering precisely to the body of work procedure without directly or indirectly influencing the standard setting results in any manner. Additionally, the body of work method is material intensive. The iterative rounds of student work review require examples at all possible test score points. Examples of student work are organized into folders for each standard setting round: training, range-finding, and pinpointing.

During Standard Setting

Panelist Training

For optimal results, panelists must understand their responsibilities during each round of the standard setting process. Thoroughly reviewing examples of student work and sorting the work into performance categories is challenging yet important work, and each panelist should feel comfortable with the steps. Thus, several training activities are completed before the actual standard setting process begins.

First, it is common to ask participants to, on their own, respond to the same tasks that were required of the examinee on the assessment. The purpose of this activity is to familiarize the participant with the assessment tasks, the student performance required by the tasks, and the general difficulty of the tasks.

An additional step for panelist training involves deep review of the performance-level descriptors. Panelists may also consider what separates a student performing near the bottom of a performance level from a student performing near the middle. Considering the knowledge and skills of a student who is "just barely" in a category helps panelists focus on student performance that is near the cut score.

Finally, to prepare for the actual standard setting rounds, panelists complete a practice round. With body of work, panelists review a small group of student work samples (five–eight sets) representing a range of possible scores. Scores on each sample are hidden from the panelists but known to facilitators. Panelists proceed through the training examples, reviewing student responses and comparing each to statements found in the performance-level descriptors. After review, panelists place each response set into a performance category, anonymously marking their judgment on a rating sheet.

Once the rating sheets have been completed, facilitators compile the ratings for each response set and display the collective ratings so that panelists can see how they generally agreed or disagreed with each other. Panelists then discuss their ratings as a group, noting the particular characteristics of the task, the student responses, and/or the performance-level descriptors that contributed to their rating. Panelists may change their ratings based on discussion with peers; however, consensus is not a requirement of the body of work process.

Range-Finding

Once the training activities and practice round conclude, panelists begin the first body of work round known as the range-finding round. The purpose of the range-finding round is to make a first “rough cut” of the dividing point between performance levels.

Panelists begin by reviewing folders of student work that represent a range of possible assessment scores. Although the scores are still unknown to the panelists, the sets of student work are ordered within a range-finding folder by score. Once panelists have sorted the student work into different performance levels, facilitators record the ratings given to each student response set. Panelists may then discuss ratings as a group, reasoning for differences or similarities. Panelists may choose to change their own ratings based on discussion but do not have to do so.

At this point in the process, the cut score is the point where specific student response sets are clearly separated into different performance levels. However, if panelists disagree and there is overlap in the ratings of a particular response set, the overlap occurring near the cut score.

Pinpointing

Pinpointing folders allow panelists to view several examples of student work at particular score points in order to focus on a more precise location of a cut score. Pinpointing folders consist of several examples (about four to five) at every score point possible on the assessment. Before the standard setting meeting, the folders are prepopulated with example student responses to both constructed response and selected response items (where appropriate); however, the assessment scores remain hidden from panelists.

After reviewing the range-finding results and examining the overlap between the panelist ratings, facilitators select pinpointing folders that are near the cut scores. For example, if student work with scores between 16 and 18 was rated by panelists as both Level 4 and Level 3, but not Level 2, then pinpointing folders containing more examples of student work at scores of 16, 17, and 18 would be chosen for further review. The panelists then sort each work example into performance Level 4 or 3, again based on the statements in the performance-level descriptors.

Calculating the Cut Score

Once panelists have independently recorded their ratings of student work samples, facilitators take the data from the rating sheets and begin to calculate the cut scores. One method used to calculate cut scores is logistic regression. For the underlying variable test score, the cut score would be placed where the probability of a test score being assigned to a particular performance level is .5. An alternative method is to calculate a median score from panelist ratings on each side of two adjacent performance categories.

Evidence to Support Results

During and after standard setting, evidence should be gathered to support the reliability and generalizability of the body of work results. For example, panelists could be asked via survey to evaluate the overall standard setting process along indicators such as (a) clarity of instruction on the process, (b) level of understanding of the process, and (c) confidence in ratings and final cut scores. Additionally, standard errors of the cut scores that describe the variability in the cut scores among the participating panelists can be calculated. Interpretation of this standard error helps determine the extent to which panelists' placement of cut scores was consistent with each other. Additional evidence could be collected from replications of the body of work process in other locations or with other panelists, though these methods are logistically more complex.

Once the standard setting event concludes, the compiled cut score calculations and generalizability evidence are presented to the assessment's governing body for review, possible adjustment, and final approval.

Gail Tiemann

See also Achievement Tests; Angoff Method; Ebel Method; Psychometrics; Standard Setting; Tests

Further Readings

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1994, June). Using actual student work to determine cut scores for proficiency levels: New methods for new tests. Paper presented at the National Conference on Large-Scale Assessment, Albuquerque, NM.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1995, June). Setting standards for performance levels using the occupational tests. Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed.). New York, NY: Routledge.

Edward L. Boone Edward L. Boone Boone, Edward L.

Bonferroni Procedure Bonferroni procedure

212

216

Bonferroni Procedure

Researchers interested in determining differences between the means of groups often employ analysis of variance (ANOVA) as a first test to determine whether differences exist. When the ANOVA test indicates that differences do exist among the group means, the next step involved is identifying which group means differ. This is typically done using either contrast tests or post hoc tests on the mean. One popular post hoc test to determine which group means differ is the Bonferroni procedure, named for Carlo E. Bonferroni, the Italian statistician who popularized the approach in the early 20th century. This entry further describes the procedure and how it is used to deal with experiment-wise error rates and for multiple comparisons.

The Bonferroni procedure is a very important tool in statistical inference and is typically used in multiple comparison situations, which is applied to many areas where multiple tests need to be conducted while preserving an overall family-wise error rate (FWR). One of the beautiful aspects of the Bonferroni procedure is its simplicity and that the resulting multiple comparisons are easy to compute. The procedure is quite common in microarray and genomics studies where there are often hundreds, if not thousands, of comparisons to be made. In those cases, the Bonferroni procedure requires the difference between the means to be quite large to be declared significant.

On the spectrum of multiple comparison procedures, the Bonferroni procedure is considered a conservative approach. Procedures such as Fisher’s least significant difference, Tukey’s honestly significant difference, and Student–Newman–Keuls test are considered more liberal and Scheffé’s method is considered more conservative. The conservative nature of the Bonferroni procedure is one of its assets, where if something is declared significant using the Bonferroni procedure, then one can be sure that the specified Type I error rate is truly preserved.

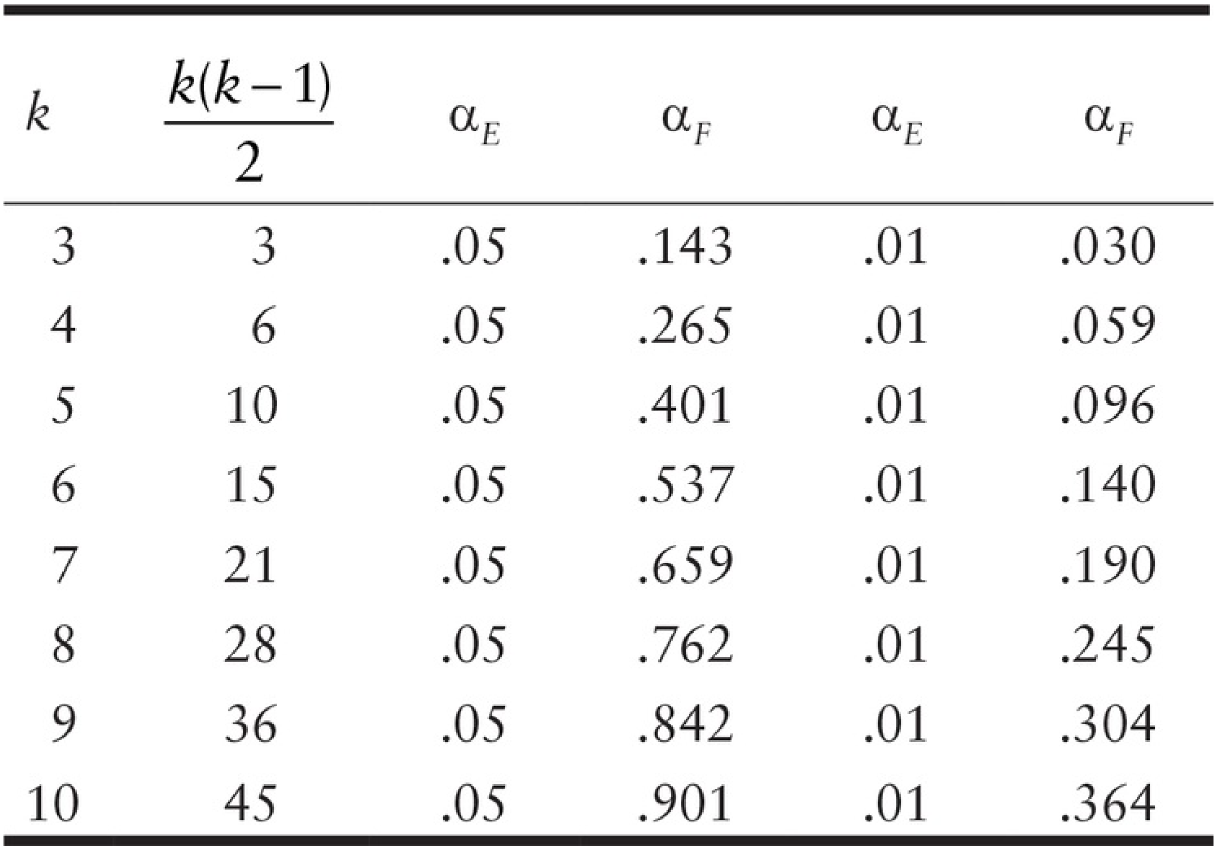
Experiment-Wise Error Rates

The main issue with comparing a large number of group means is being able to control the Type I error rate, in this case, falsely claiming two group means are different. Suppose a researcher is interested in *k* groups means: μ_1, μ_2,…, μ_k. To examine every possible difference between two means would result in separate tests. If *k* is large, this could be a large number of tests each with its own Type I error rate of α. Using probability, we can find a lower bound for the Type I error rate for all comparisons, which is much higher than α. The Type I error rate across a family of comparisons or hypotheses is called the FWR, denoted α_E, and reflects the probability that one makes at least one Type I error among all comparisons. The individual Type I error rate is often called the *experiment-wise error rate*, denoted α_E, and is the probability of making a Type I error when considering a single comparison of two means. Using probability, we can obtain a lower bound for the experiment-wise error rate based on the number of comparisons and the comparison-wise error rate and assuming all Type I errors are independent.

$$
\begin{aligned}
\alpha_F &= P(\text{at least one Type 1 error}) \\
&= 1 - P(\text{no Type 1 errors}) \\
&= 1 - P(\text{no Type 1 error on a single} \\
&\qquad \text{comparison})^{\frac{k(k-1)}{2}} \\
&= 1 - \begin{bmatrix} 1 - P(\text{Type 1 error on a single} \\ \text{comparison)} \end{bmatrix}^{\frac{k(k-1)}{2}} \\
&= 1 - (1 - \alpha_E)^{\frac{k(k-1)}{2}}.
\end{aligned}
$$

Before going forward, it should be mentioned that the formula just shown works for any situation where the number of tests/comparisons, denoted *h* can be computed. Simply replace the with *h*. This can be useful for situations where one may not be interested in all possible paired comparisons. For example, one may be simply interested in whether a specific treatment-level group differs from the control group.

Table 1 shows the FWR, α_F, for various number of groups to be compared with α_E = .05 and .01. Notice that even for a few number of groups to be compared that the number of actual paired comparisons, , is quite high. For example, consider the case when *k* = 5, the number of paired comparisons is 10 and the FWR is α_F = .401 which means there is a 40.1% chance that a Type I error will be made on at least one of the paired comparisons. The problem gets worse for larger number of groups, say *k* = 10, the number of paired comparisons is 45 and the experiment-wise error rate is α_F =.901, which means that there is a 90.1% chance that a Type I error will be made on at least one of the paired comparisons. Many researchers would consider the FWR to be unacceptably high in these cases.



k	$\frac{k(k-1)}{2}$	α_E	α_F	α_E	α_F
3	3	.05	.143	.01	.030
4	6	.05	.265	.01	.059
5	10	.05	.401	.01	.096
6	15	.05	.537	.01	.140
7	21	.05	.659	.01	.190
8	28	.05	.762	.01	.245
9	36	.05	.842	.01	.304
10	45	.05	.901	.01	.364

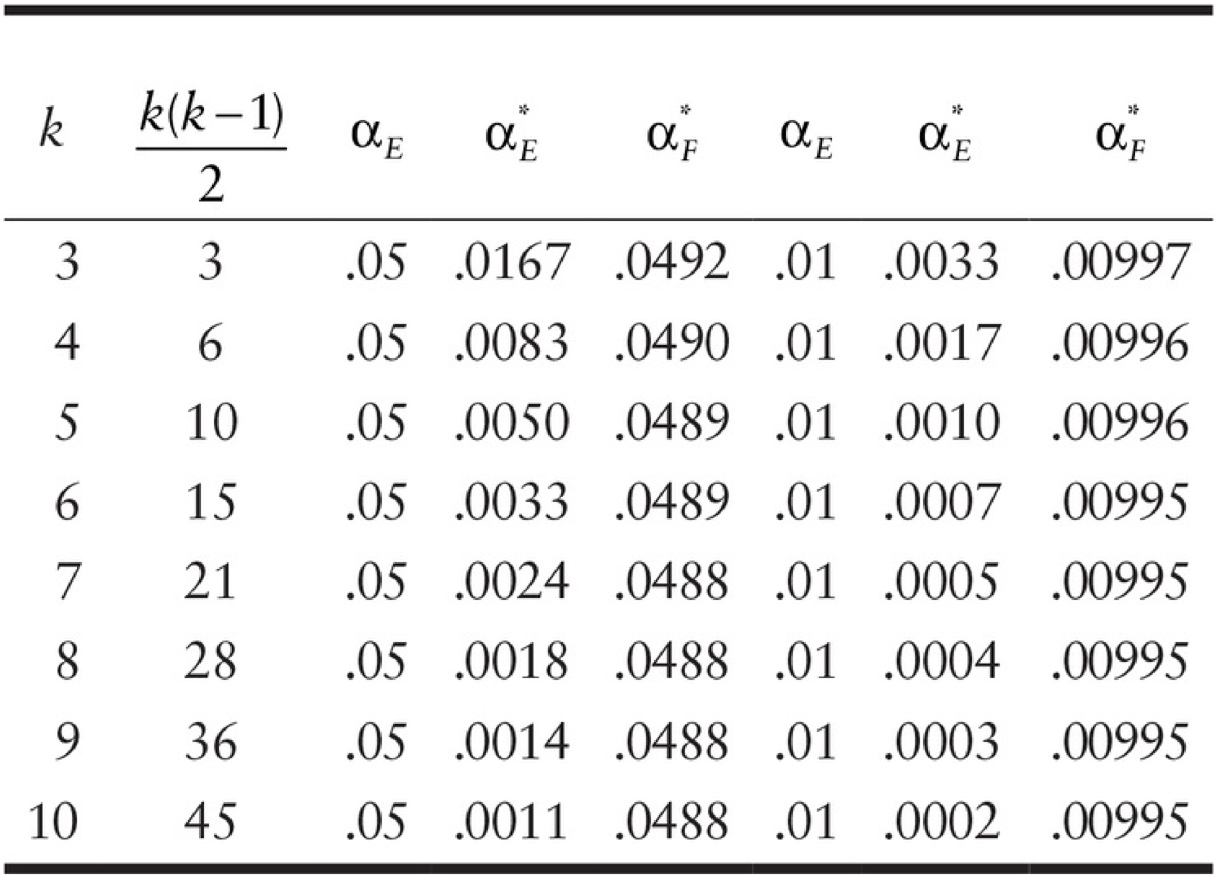
Because the problem of high experiment-wise error rates is motivated by probability, the solution can be found there as well.

Using the Bonferroni inequality, we can obtain a correction for the significance level so that the resulting inferences will satisfy the FWR. Suppose we are testing h hypotheses, H_1, H_2,…, H_h (which could ccorresponding p values p_1, p_2, …, p_h. Under the null hypothesis that all hypotheses are true, then the FWR is the probability we reject at least one of the hypotheses. This gives the following:

$$\alpha_F = \text{FWR} = P(p_1 \leq \alpha_E \text{or } p_2 \leq \alpha_E \text{ or} \ldots p_h \leq \alpha_E).$$

This establishes that FWR ≤ $h\alpha_E$. Hence, by simply setting a *Bonferroni corrected* significance level to , one can guarantee that $\alpha_F \leq \alpha_E$. For the multiple pairwise comparison case, the corresponding Bonferroni corrected significance level would be . Using , the lower bound on the resulting experiment-wise error rate, can be calculated by .

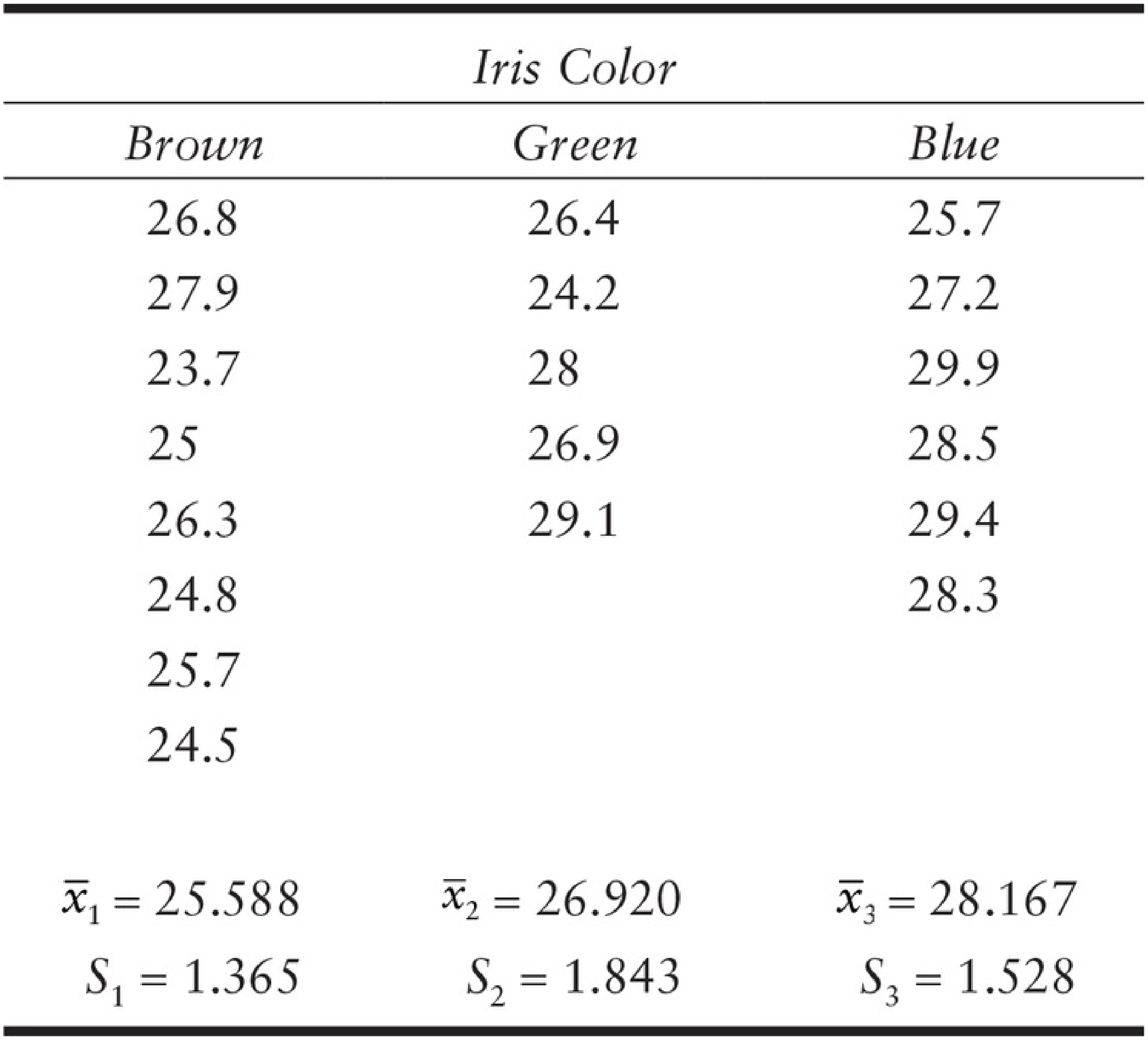
Table 2 shows the number of groups, the number of pairwise comparisons, the associated Bonferroni corrected significance levels, and the resulting lower bound on the experiment-wise error rates. Notice that as the number of groups increase the Bonferroni corrected significance levels, decrease dramatically for both α_E = .05 and .01. Also notice that the resulting lower bound on the experiment-wise error rate, , remains constant and slightly below α_E.



k	$\frac{k(k-1)}{2}$	α_E	α_E^*	α_F^*	α_E	α_E^*	α_F^*
3	3	.05	.0167	.0492	.01	.0033	.00997
4	6	.05	.0083	.0490	.01	.0017	.00996
5	10	.05	.0050	.0489	.01	.0010	.00996
6	15	.05	.0033	.0489	.01	.0007	.00995
7	21	.05	.0024	.0488	.01	.0005	.00995
8	28	.05	.0018	.0488	.01	.0004	.00995
9	36	.05	.0014	.0488	.01	.0003	.00995
10	45	.05	.0011	.0488	.01	.0002	.00995

Bonferroni Procedure for Multiple Comparisons

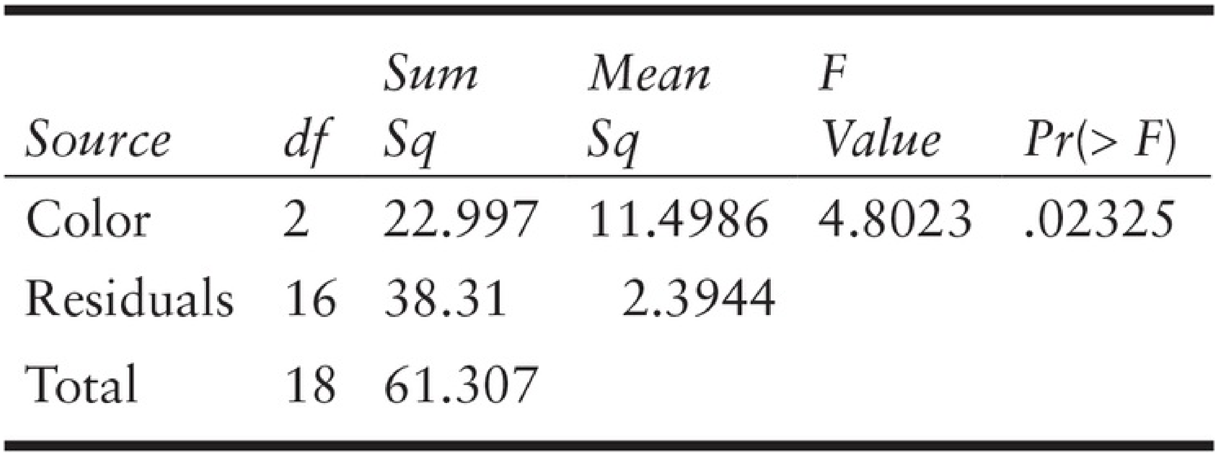
In a 1973 article, James M. Smith and Henryk Misiak attempted to determine whether the individuals’ critical flicker frequency, or the frequency at which the person cannot distinguish a flickering light from a steady, nonflickering light, is related to their iris color. The critical flicker frequency is the highest frequency in which an individual can detect a flicker. This is important as the multiple screens in modern life have various refresh rates and both manufacturers and users want a visually smooth experience. The data from their experiment is given in Table 3 below. Notice that the data are unbalanced in the sense that there are not an equal number of observations in each group (iris color).



Iris Color		
Brown	*Green*	*Blue*
26.8	26.4	25.7
27.9	24.2	27.2
23.7	28	29.9
25	26.9	28.5
26.3	29.1	29.4
24.8		28.3
25.7		
24.5		
$\bar{x}_1 = 25.588$	$\bar{x}_2 = 26.920$	$\bar{x}_3 = 28.167$
$S_1 = 1.365$	$S_2 = 1.843$	$S_3 = 1.528$

Source: Smith, J. M., … Misiak, H. (1973, p. 93).

The first step in the Bonferroni procedure for multiple comparisons is to perform an ANOVA analysis to determine whether differences exist and to obtain the mean square error and degrees of freedom for the error. In this example, a significance level of α = .05 is employed. Table 4 shows the ANOVA table for the flicker frequency data. Notice that the *p* value for the overall test for group differences is .02325 which is below the chosen significance level of α = .05 and indicates that significant difference exists among the mean flicker frequencies across eye color.



Source	df	Sum Sq	Mean Sq	F Value	Pr(> F)
Color	2	22.997	11.4986	4.8023	.02325
Residuals	16	38.31	2.3944		
Total	18	61.307			

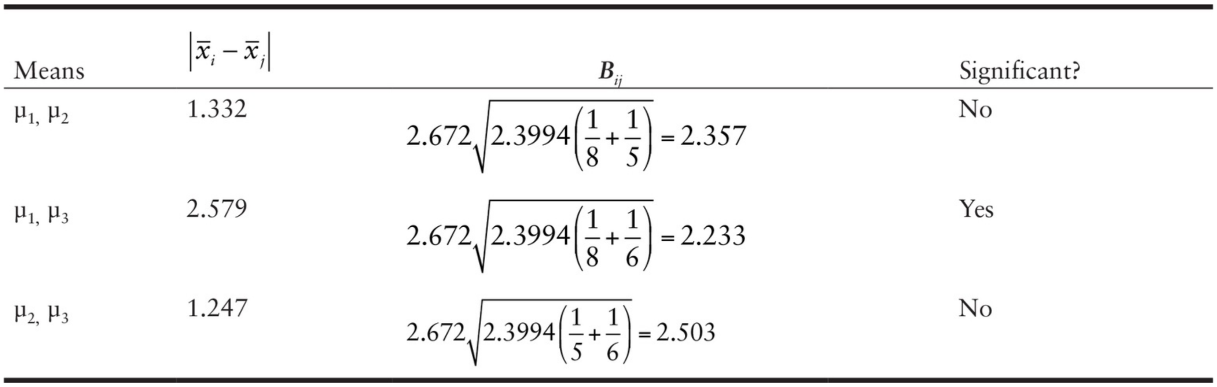
Note: ANOVA = analysis of variance.

Because differences exist in the data, the Bonferroni procedure will be used to determine which means differ. Similar to Fisher's least significant difference procedure, the following formula is used to determine the minimum difference for two means to be significantly different:

$$B_{ij} = t^{*}_{\frac{a_E^{*}}{2},n-k}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

where is the critical value for the *t* distribution with *n* −*k* degrees of freedom and n_i and n_j are the sample sizes of groups *i* and *j*, respectively. Notice that is divided by 2 in the formula to reflect that a two-sided test is being performed. If the absolute difference between and is larger than $\boldsymbol{B}_{ij}$, then the means are considered statistically significantly different. In the case of balanced data, $\boldsymbol{B}_{ij}$ will be a constant number across all pairwise comparisons. This is not the case in the flicker frequency data and $\boldsymbol{B}_{ij}$ will need to be calculated for each pairwise comparison.

In order to perform the multiple comparisons on the flicker frequency data, the critical value from the *t* distribution is needed and in this case is . From Table 4, the MSE is 2.3994 and n_1 = 8, n_2 = 5, and n_1 = 6. Table 5 shows the calculations for the absolute difference in means , $\boldsymbol{B}_{ij}$, and the determination of whether the difference is statistically significant or not. Notice that $\boldsymbol{B}_{ij}$ differs for each comparison due to the unbalanced nature of the data and that in the second case one would declare μ_1 and μ_3 as statistically different. The other two comparisons would not be considered statistically different.



Means	$\lvert \bar{x}_i - \bar{x}_j \rvert$	B_{ij}	Significant?
μ_1, μ_2	1.332	$2.672\sqrt{2.3994\left(\frac{1}{8}+\frac{1}{5}\right)} = 2.357$	No
μ_1, μ_3	2.579	$2.672\sqrt{2.3994\left(\frac{1}{8}+\frac{1}{6}\right)} = 2.233$	Yes
μ_2, μ_3	1.247	$2.672\sqrt{2.3994\left(\frac{1}{5}+\frac{1}{6}\right)} = 2.503$	No

An alternative approach (with less power) for determining whether the groups differ would be to perform all three pairwise *t* tests and evaluate the resulting *p* values with the significance level. In the case of the flicker frequency data, this approach would yield the same inferences as the approach presented earlier.

Edward L. Boone

See also Analysis of Variance; Experimental Designs; Post Hoc Analysis; Type I Error

Further Readings

Montgomery, D. (2012). Design and analysis of experiments (7th ed.). New York, NY: Wiley.

Ott, R. L., & Longnecker, M. (2015). An introduction to statistical methods and data analysis (7th ed.). Pacific Grove, CA: Brooks Cole.

Smith, J. M., & Misiak, H. (1973). The effect of iris color on critical flicker frequency (CFF). The Journal of General Psychology, 89, 91–95.

Turner, J. R., & Thayer, J. (2001). Introduction to analysis of variance. Los Angeles, CA: Sage.

Daniel Lewis Daniel Lewis Lewis, Daniel

Bookmark Method

Bookmark method

216

217

Bookmark Method

The bookmark method is a standard setting method used to establish one or more cut scores associated with interpretable levels of performance on an assessment. In 1995, Daniel Lewis and Howard Mitzel developed the bookmark method that became widely used in the 2000s, with a majority of states employing it to meet the requirements of the No Child Left Behind Act (in particular, requirements associated with reporting test results in terms of achievement levels). This entry describes the technical foundations of the bookmark method and how it uniquely executes activities common to many standard setting procedures.

Standard setting is necessary to systematically set one or more cut scores that separate a test scale into two or more categorical levels of achievement such as “failing” and “passing.” The bookmark method is differentiated from other standard setting methods by its use of item response theory and the ordered item booklet (OIB) as a foundation for key standard setting activities. The bookmark method continues to be widely used both internationally and for U.S. state summative assessment programs. Variations of the bookmark method, such as the Mapmark method, have also emerged in practice.

Most standard setting methods assemble a qualified panel of subject-matter experts to participate in a standardized process that includes the following three key activities:

1. the orientation of panelists to the testing program and test of interest,
2. the training of panelists to make ratings that support cut score estimation, and

3. discussion and consensus building among panelists over multiple rounds of ratings.

The bookmark method’s approach to these three standard setting activities uniquely defines the method. The primary bookmark method tool, the OIB, is assembled in print or digitally as an ordered set of scaled test items that is representative of the construct measured by the assessment of interest.

OIB items are displayed in ascending order, by difficulty, which is defined as scale location. The scale location of dichotomous item *i* is the score, S_i, such that an examinee with ability S_i has a specified probability of success, referred to as the response probability. The most common response probability employed in practice, and for discussion in this entry, is 2/3. Thus, a selected response item is located at the scale location where an examinee has a 2/3 probability of success.

Polytomous items are located at multiple scale locations, one for each positive score point. For example, a constructed response item with obtainable scores of 0, 1, and 2 is mapped on two locations—the scale scores where an examinee has a 2/3 probability of achieving at least a 1 and at least a 2, respectively. Thus, the OIB is a set of test items that represent the construct of interest, presented in order of difficulty, with dichotomous items intermingled with polytomous item score points. The OIB supports each of the three key standard setting activities listed earlier. For simplicity, let us assume panelists are setting a single “passing” cut score.

The first activity—the orientation of panelists to the testing program and test of interest—begins with training that is relatively undifferentiated among standard setting procedures including discussion of the nature and consequences of the testing program, the content standards, and scoring rubrics. However, training on the construct measured by the test is uniquely implemented under the bookmark method by a structured study of the items in the OIB. Panelists typically study the OIB in small groups by reviewing the test items in order, from easiest to hardest, and answering and discussing the following two questions for each item:

1. What does this item (score point) measure? That is, what do you know about a student who knows the correct response to this item (obtains at least the given score point)?
2. Why is this item (score point) more difficult than the previous items in the OIB?

This activity is intended to impart to panelists an integrated conceptualization of what the test measures.

The second activity—the training of panelists to make ratings that support cut score estimation—also utilizes the OIB. Panelists make their ratings, for say, the passing cut score by placing a bookmark at the first point in the OIB such that a student who has mastered the content reflected by the items prior to the bookmark has demonstrated a level of achievement sufficient to pass. By defining "mastered" as "having at least a 2/3 likelihood of success," a passing cut score can be associated with each bookmark using item response theory.

The third activity—discussion and consensus building—fosters communication among panelists to support their common understanding of the diverse perspectives reflected by panelists' varied bookmark placements. Panelists make their first ratings independently and without discussion. A second round of ratings occurs after discussion of the rationales that support panelists' different ratings. This substantive discussion is facilitated using a concrete, content-based representation of panelists' ratings by placing a bookmark in the OIB for each panelist's rating—all panelists' ratings are represented in each panelist's OIB. Discussion of panelist differences is based on a review of the items between the first and last of the panelists' bookmarks (which represent the lowest and highest panelist expectations, respectively). This discussion requires panelists whose bookmarks differ to provide rationales for their inclusion or exclusion of items that students are expected to master to pass the test of interest.

During this process, consensus is *fostered* through discussion of differences but is not *required*. A "consensus" cut score is estimated, typically by taking the median rating, as the cut score recommended to the sponsoring agency.

Daniel Lewis

See also Angoff Method; Body of Work Method; Cut Scores; Ebel Method; Item Response Theory; Proficiency Levels in Language; Standard Setting

Further Readings

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. Educational Measurement: Issues and Practice, 25(1), 4–12.

Lewis, D., & Lord-Bessen, J. (2016). Standard setting. In W. J. van der Linden (Ed.), Handbook of item response theory: Vol. 3. Applications. Boca Raton, FL: Chapman … Hall/CRC.

Lewis, D. M., Mitzel, H. C., Mercado, R., & Schulz, M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (2nd ed.). Mahwah, NJ: Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2000). The bookmark procedure: Cognitive perspectives on standard setting. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Erlbaum.

W. Holmes Finch W. Holmes Finch Finch, W. Holmes

Bootstrapping

Bootstrapping

217

221

Bootstrapping

Bootstrapping, or the bootstrap, is a statistical methodology that is frequently used in situations where standard distributional assumptions, such as normality, do not hold. In addition, the bootstrap can be used to estimate standard errors and confidence intervals for parameter estimates. It is particularly useful where there is not a known sampling distribution for the statistic of interest, thereby making calculation of standard errors difficult or impossible. There are a number of variations in the bootstrap that make it useful in a wide variety of situations. Regardless of context or application, the bootstrap is based upon a basic framework of resampling with replacement from the original sample. This entry discusses the basic nonparametric bootstrap, bootstrap confidence intervals, variations in the bootstrap, and when to use the bootstrap.

Basic Nonparametric Bootstrap

As an example, we will consider the problem of estimating the standard error for the mean, *x*. This statistic can be calculated in a straightforward manner using the equation:

$$SE_x = S/\text{Square root}\left(N\right),$$

where *S* is the standard deviation of the sample and *N* is the sample size.

Equation 1 is based upon an assumption that the population distribution underlying the variable *x* is normal. However, if this is not the case, then Equation 1 no longer yields the appropriate standard error estimate of *x*. The bootstrap offers an alternative approach for calculating the standard error. The basic nonparametric bootstrap operates using the following steps:

1. Calculate sample statistic of interest (e.g., *x*) for the original sample.
2. Randomly sample *n* individuals from the original sample of size *n*, with replacement; individuals can appear multiple times in the bootstrap sample, while others may not appear at all.
3. Calculate the mean, *xB**, for the bootstrap sample.
4. Repeat Steps 2 and 3 many times (e.g., *B* = 10,000) to create a sampling distribution for the statistic of interest.
5. Calculate the bootstrap standard error: $SB = 1\ BxB^{*} - x^{*}\ 2B-1$.

To illustrate the bootstrap, consider the following simple example involving a sample of five individuals with scores 8, 3, 6, 1, and 5. The mean of these values is 4.6, and the standard deviation is 2.7. Based on Equation 1, the standard error is S_x = 2.75 = 1.2. Now, let’s draw five bootstrap samples (which would be far too few in practice but helps to illustrate how the bootstrap works) and calculate the mean for each. The samples appear below.

The standard deviation of the means for the five bootstrap samples is 0.68. Thus, based on these five samples, we would report that the bootstrap standard error of the mean is 0.68. In actual practice, we would use many more than five bootstrap samples, perhaps as many as 10,000. To finish this illustration of the basic bootstrap, a total of 1,000 bootstrap samples of the five data points were drawn using the software package SPSS, yielding a bootstrap standard error estimate of 1.06.

Bootstrap Confidence Intervals

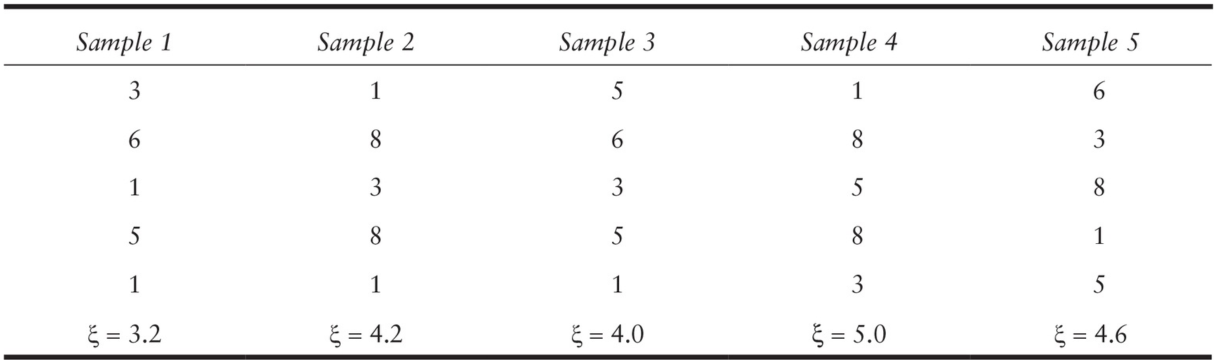
Standard errors are frequently used to construct a confidence interval for the statistic of interest, in this case the mean. Confidence intervals reflect the neighborhood of values within which the population parameter is likely to reside. Using normal theory methods, the confidence interval for the mean is calculated as:

$$x \pm tcvsxn,$$

where the terms in the equation are as defined earlier, with the addition of *tcv*, which is the critical value of the *t* distribution corresponding to the level of confidence that we would like for our interval (e.g., 95%). There are three common methods for constructing confidence intervals using the bootstrap. In the first of these, the bootstrap standard error approach, *sx,* in Equation 2 is replaced by *SB* leading to:

$$x \pm tcvsBn.$$

The second method for constructing a confidence interval using the bootstrap is known as the percentile bootstrap method. It works by taking the bootstrap distribution of the mean and ordering the values from smallest to largest. The lower bound of a 95% confidence interval would correspond to the 2.5th percentile of this distribution, and the upper bound of the confidence interval would correspond to the 97.5th percentile.



Sample 1	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*
3	1	5	1	6
6	8	6	8	3
1	3	3	5	8
5	8	5	8	1
1	1	1	3	5
$\xi = 3.2$	$\xi = 4.2$	$\xi = 4.0$	$\xi = 5.0$	$\xi = 4.6$

The third approach for constructing the confidence interval of the mean using the bootstrap is called the bootstrap-*t* approach. With this method, the following value is calculated for each of the bootstrap samples:

$$t^{*} = xB^{*} - x^{*}SBn,$$

where the terms are as defined previously. The confidence interval is then calculated as:

$$x \pm t95\text{th percetile}^{*}Sxn.$$

The *t*95th percentile* is simply the bootstrap *t* value corresponding to the 95th percentile across the entire bootstrap *t* distribution.

Variations in the Bootstrap

The methodology described earlier represents the basic nonparametric bootstrap.

There are, however, a wide variety of bootstrap algorithms available for use in specific situations. The parametric bootstrap is similar in spirit to the nonparametric approach, except that rather than draw samples from the data itself, bootstrap samples of size *N* are drawn from a known distribution, such as the normal. For example, a psychologist working with IQ scores may elect to draw bootstrap samples from the normal distribution with a mean of 100 and a standard deviation of 15, as these correspond to the population distribution of IQ scores. In so doing, the psychologist is making the tacit assumption that the current sample comes from the population where the distribution of IQ scores matches that described earlier, and thus drawing from that distribution may yield somewhat more representative values than merely drawing from the current sample.

Another alternative bootstrap approach is known as the smooth bootstrap. This methodology is based upon the nonparametric bootstrap but adds a small random value to each of the data points drawn in the process of bootstrap resampling. Thus, in the illustration described in the previous section, each of the five values drawn for each bootstrap sample would have added to it a small random number drawn from, perhaps, the standard normal distribution. The reason for doing this is similar to that underlying the use of the parametric bootstrap, namely, to recognize that the current sample, while hopefully representative of the population, does not contain all possible values of the variable in the population. Thus, by adding a small random number to each sampled value, we increase the breadth of the bootstrap sample.

When the original data are sampled in a clustered fashion, such as students within schools, it is more beneficial to conduct the bootstrap resampling at the higher level of data (e.g., resampling schools rather than children). This approach is known as the block bootstrap. As an example, assume that an educational researcher has randomly sampled 100 schools within a state and that each child in each school is then included in the sample. The researcher would like to estimate a regression model relating academic motivation to academic achievement. Given the clustered nature of the data, the school that a child attends must be considered in any statistical modeling that is done. Thus, in the context of bootstrapping, the resampling will be done at the school level, so that schools are resampled with replacement, and all of the children within the selected schools for a given bootstrap sample are included in the subsequent analysis. In other respects, the block bootstrap works in much the same fashion as the nonparametric bootstrap described earlier, such that a large number of resamples are drawn, and the statistics of interest are calculated. The block bootstrap is the preferred method for bootstrapping clustered data and can also be applied to time series data, in which individuals are measured across time.

In the context of regression and other linear models, there are two additional bootstrapping algorithms that have proven to be useful. The first of these involves the resampling of model residuals rather than the actual sampled data. This approach works as follows:

1. Fit a regression model to the data, such as $y = b_0 + b_1x$.
2. Calculate the residuals from this model, $e = y - y$ for each individual in the sample.
3. Add the residual to the observed dependent variable value for each individual: $y^* = y + e$.
4. Fit the regression model using y^* as the dependent variable rather than y.
5. Repeat Steps 3 and 4 a large number (e.g., 1,000) of times.

This approach to bootstrapping incorporates information about the regression relationship and thus has the advantage of leading to more accurate and representative models than does simply resampling the individual observations as in the standard bootstrap.

Another alternative to the bootstrap is the wild bootstrap, which is similar in spirit to the resampling of residuals approach. It is particularly useful when the regression model exhibits unequal variance. The wild bootstrap involves resampling residuals, and creating a y^* value, just as with the residual bootstrap. However, the residuals are adjusted by a multiplier, v: $y^* = y + ve$. The value v can come from the standard normal distribution or could come from another distribution as is described in the literature. The wild bootstrap has been shown to be particularly useful for the small samples with unequal error variance.

When to Use the Bootstrap

The bootstrap is applicable in a wide range of situations and for many different statistics. Earlier, the entry discussed the simple example of estimating the standard error of the mean and then constructing a confidence interval for the mean. However, the bootstrap can be used for hypothesis testing, estimating standard errors and confidence intervals for complex statistical models such as factor analysis and item response theory, and has application in multilevel modeling.

The bootstrap is an alternative both to standard data analytic techniques that are based on distributional assumptions, such as normal-based methods, and to traditional nonparametric approaches to data analysis that rely on ranks, or permutations of the data. In addition, when the mathematical function underlying a particular statistic or model is not well known or proves intractable to estimate, the bootstrap can serve to be a valuable tool for estimating standard errors and confidence intervals. The bootstrap is useful for work with small samples, particularly if the model is complex and difficult to estimate. Researchers should carefully consider the bootstrap as an alternative to more traditional methods of estimating standard errors and conducting hypothesis tests, particularly when sample sizes are small and data do not meet standard distributional assumptions.

The bootstrap can yield biased estimates, and the researcher should consider the bias corrected and accelerated approach to estimating standard errors and confidence intervals based on the bootstrap. In addition, if the sample itself is not representative of the population, then results from the bootstrap are no more generalizable than those from any other statistical procedure. In summary, the bootstrap is a useful tool for researchers to consider when the model to be fit is complex and the standard error does follow a standard form. It is not, however, a panacea and should be used thoughtfully and with care, just as is the case with any statistical model.

W. Holmes Finch

See also Confidence Interval; Random Assignment; Random Selection; Simple Random Sampling; Standard Error of Measurement

Further Readings

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. Boca Raton, FL: Chapman … Hall/CRC.

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their applications. Cambridge, UK: Cambridge University Press.

Lunnebor, C. E. (2000). Data analysis by resampling: Concepts and applications.

Pacific Grove, CA: Duxbury Press.

Wilcox, R. (2012). Introduction to robust estimation and hypothesis testing. Amsterdam, the Netherlands: Elsevier.

Jennifer A. Brussow Jennifer A. Brussow Brussow, Jennifer A.

Box Plot

Box plot

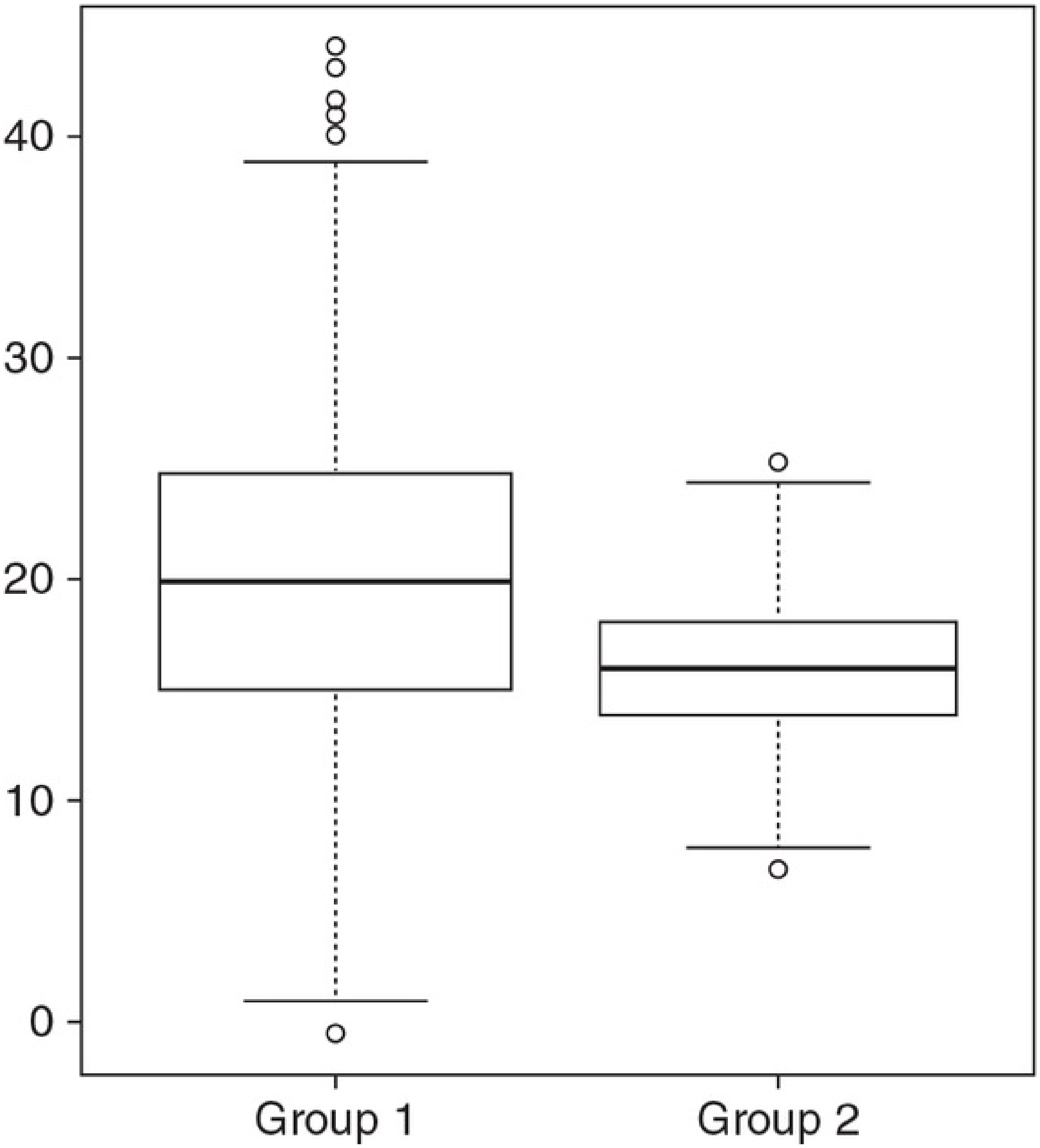
221

222

Box Plot

Box plots (also called box-and-whisker diagrams) are a concise way of displaying the distributions of a group (or groups) of data in terms of its median and quartiles. This way of describing a data set is commonly called a *five-number summary*, where the five numbers are the minimum, first quartile, median, third quartile, and the maximum. While less informative than histograms, box plots are helpful for identifying outliers and comparing distributions between groups. Figure 1 provides an example of a box plot showing the distributions of data for two different groups. The whiskers are represented according to their most common method of calculation: the most extreme values falling within 1.5 times the interquartile range (sometimes abbreviated as IQR).

Figure 1 Box plot with whiskers and outliers

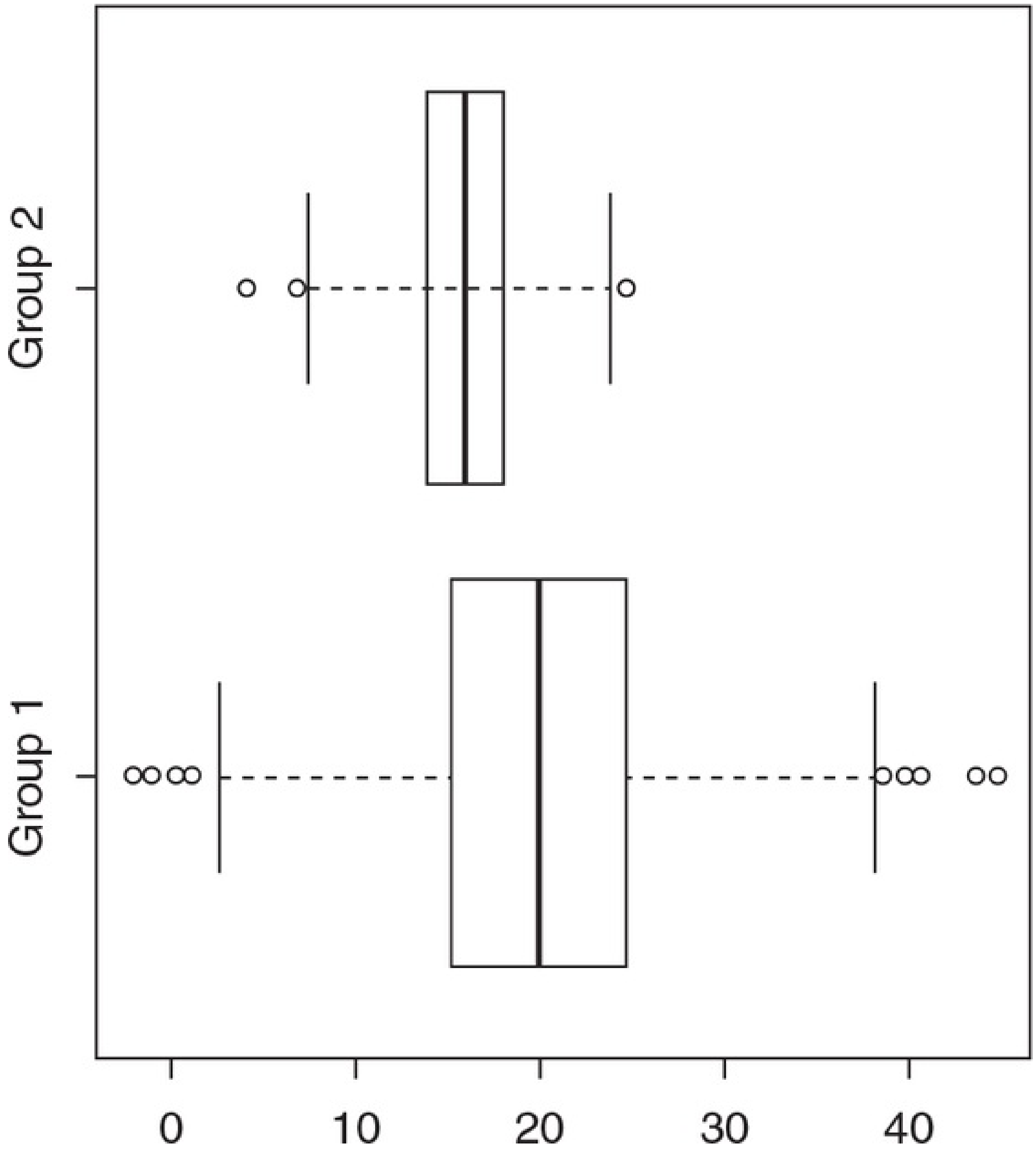


A box plot consists of a box whose upper bound represents the 75th percentile, or third quartile, and lower bound represents the 25th percentile, or first quartile. The boundaries of the box are sometimes called the *upper and lower hinges*, and the distance between the hinges is sometimes referred to as the H-spread. The median, or 50th percentile, is represented by a line bisecting the plot. The median is also referred to as the second quartile. The mean may also be displayed in a box plot by adding a cross or an “X” to the plot. The range between the upper and lower bounds is called the interquartile range (sometimes abbreviated as IQR).

In some instances, “whiskers” are added to this box; the whiskers typically extend to the farthest points in the data that are still within 1.5 times the IQR from the lower and upper quartiles. The unit of 1.5 times the IQR was set by John Tukey when he created the box-and-whisker plot and is sometimes called a *step*. Values falling outside of the whiskers, yet within 3 times the IQR, are considered *outliers* and are represented on the plot with dots or asterisks. Values falling beyond 3 times the IQR are considered *extreme values* and may be denoted with a different type of marking than the outliers.

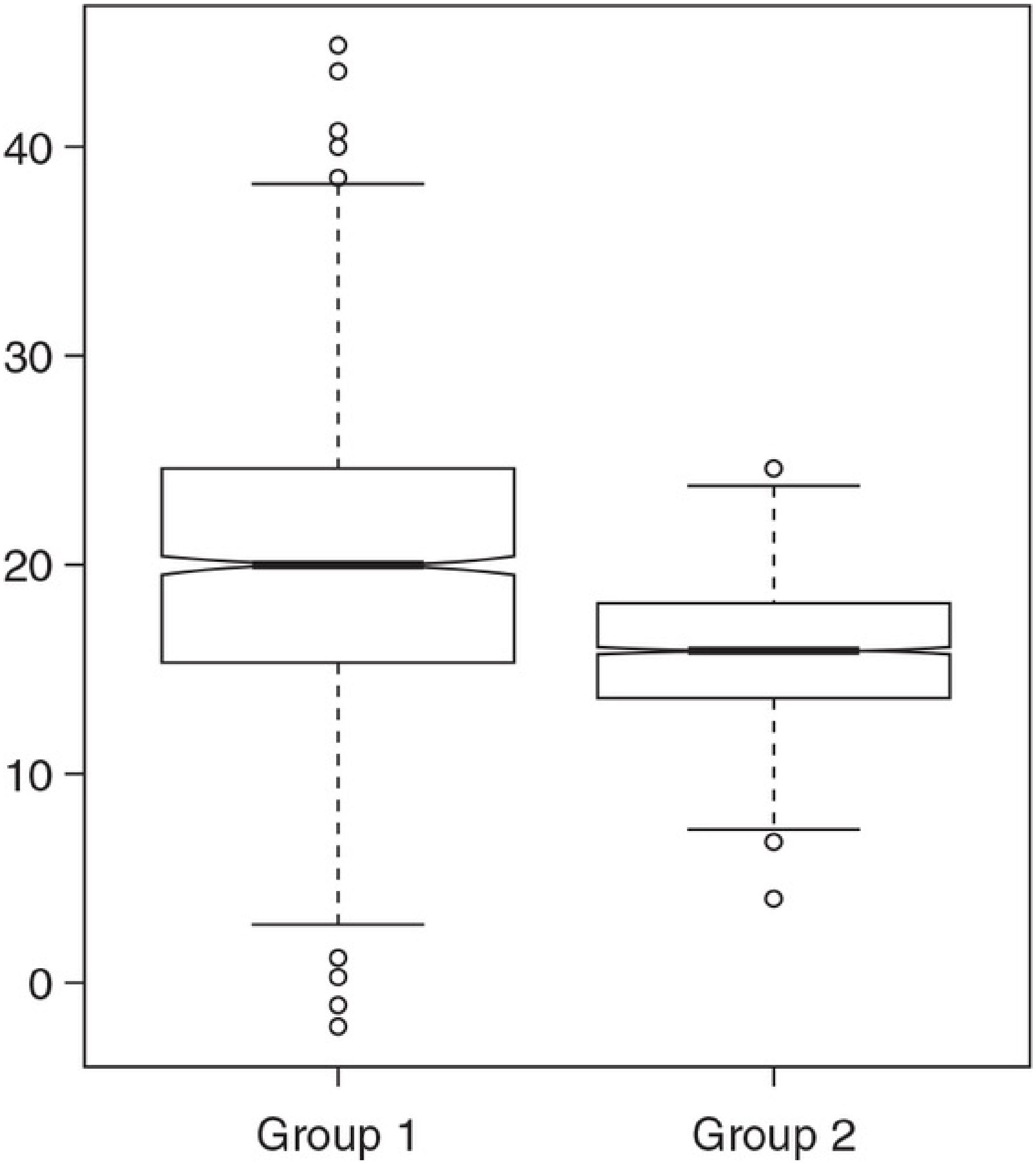
Box plots may be drawn horizontally or vertically, though the vertical format is most commonly encountered. Figure 2 provides an example of a horizontal box plot. Some simple box plots are constructed, so that the whiskers extend all the way to the minimum and maximum values of the data set; in this layout, there would be no outliers or extreme values displayed. Another variation sets the boundaries of the whiskers at 1 standard deviation above and below the mean of the data; still other variations set the boundaries at the ninth and 91st or second and 98th percentiles. When constructing a box plot, it is appropriate to include information regarding the conventions used in its construction in the caption.

Figure 2 Horizontal box plot with whiskers and outliers



Another variation of the box plot includes notches around the median; the width of the notch is proportional to the IQR. Figure 3 shows an example of this variation. Notches are commonly used to offer information on the significance of the difference between medians. If the notches do not overlap, the difference between the medians is likely to be statistically significant.

Figure 3 Box plot with whiskers, outliers, and notches



Although box plots are less informative than histograms, they take up less space within a document and allow the viewer to quickly contrast several groups of data. They also provide more information about the data set than simply reporting the mean and standard deviation for each group because these values can be misleading when dealing with nonnormal data and do not provide information regarding outliers. Box plots are an efficient way to visualize multiple groups of data while also providing information about skewness and outlying values.

Jennifer A. Brussow

See also Bar Graphs; Data Visualization Methods; Histograms; Interquartile Range; Quartile

Further Readings

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. The American Statistician, 32(1), 12–16.

Tukey, J. W. (1977). Box-and-whisker plots. Exploratory Data Analysis, 39–43.

Patricia D. López Patricia D. López López, Patricia D.

Brown v. Board of Education Brown v. board of education

222

226

Brown v. Board of Education

This entry provides a brief historical overview of the U.S. Supreme Court's landmark decision in *Brown v. Board of Education of Topeka* and elaborates on how the process of achieving equal access to educational opportunities for all public school students has taken shape. It also discusses the role the *Brown* decision has played in educational assessment, research, and practice, with a focus on the sociopolitical aspects of these practices.

The 1954 decision in *Brown v. Board of Education* set forth an historical shift in the U.S. public educational system's responsibility to provide educational opportunities for all students, regardless of race. The case was filed on behalf of Oliver Brown, a parent of a Black child who was denied access to a segregated White school in Topeka, Kansas. Segregation can be understood as a practice that divides and orders individuals along racial lines.

Although the case falls under one single name, there were actually two decisions—namely, *Brown I* (1954) and *Brown II* (1955)—both of which were unanimous. *Brown I* put an end to the "separate but equal" doctrine, arguing that public schools that separate students into different facilities based on race are not equal nor can they be made equal, thereby marking them unconstitutional. The court unanimously held that the racial segregation of children and youth in public schools violates the Equal Protection Clause of the Fourteenth Amendment. This clause states that "no state shall make or enforce any law which shall … deny to any person within its jurisdiction the equal protection of the laws." Furthermore, the court asserted that the Fourteenth Amendment guarantees all children and youth access to an equal education, regardless of race. Prior to this decision, a state was allowed to divide people along racial lines, designating some schools Whites-only as long as it provided facilities and teachers that were of equal quality.

In the following year, the same court put forth a second decision referred to as *Brown II* that delegated responsibility to states and districts to implement the new Constitutional principles outlined in *Brown I* “with all deliberate speed.” As part of this full compliance, localities were required to identify methods for addressing and assessing equal access to education among students in newly integrated school settings.

The *Brown* decision is relevant to the topics of education research, assessment, measurement, and evaluation in two meaningful ways: First, the decision helps to define parameters for understanding equal educational opportunity through a lens that focuses on the comparative differences among student subgroups (e.g., students of color vs. White students); and second, the decision advances the notion that school systems fail when they do not provide all students with equal opportunities to perform at the same levels. These two points of relevancy will inform the subsequent sections of this entry.

Assessment, Measurement, and Accountability to *Brown*

The *Brown* decision played an instrumental role in establishing a precedent on the importance of equalized access to educational opportunities among students of color that were not previously existent in the context of lawful segregation. Through a lens that views education as a great equalizer—where schools are viewed as a foundational institution for accessing opportunity—the *Brown v. Board of Education of Topeka* victory paved the way for students of color to have equal access to the same educational opportunities as their White peers. Equality and opportunity were to be defined in terms of buildings, curricula, educator qualifications, and teacher salaries; all of which are objective and measurable factors that place the onus of educational access in the hands of school systems.

As part of this framework’s implementation, it became necessary to develop evaluation methods that demonstrated the extent to which state and districts were in compliance. Accordingly, cognitive tests were increasingly used to assess students.

Standardized tests are hailed as being a cost efficient form of evaluation that is

Standardized tests are hailed as being a cost efficient form of evaluation that is capable of being administrated in mass, and universally, to students in varied contexts. Such approaches are particularly appealing to fiscal decision makers who are primarily concerned with costs. Standardized assessments also allow for comparative analyses that reveal differences in performance levels among student subgroups. These analyses of comparative differences inspired the concept of the achievement gap, which remains a primary focal point in education today.

The achievement gap, or the observed, persistent disparity of educational performance between groups of students, is measured foremost by student performance on standardized testing and at times by other outcome measures such as graduation rates, dropout rates, or grade retention. The concept of the achievement gap brings into consciousness the idea that a student’s ability, as demonstrated by performance on standardized metrics, is the central focal point for understanding access to an equal education. This framework places the onus on the recipient (i.e., students) rather than the institution (i.e., schools), thereby presenting a conceptual deviation from *Brown*’s focus on access to an equal education as a function of institutional availability to opportunities. The latter framing is understood in education as the “opportunity to learn,” where disparities in a student’s educational trajectory are understood through the gaps and differences embodied within and across schools.

With the rise of standards-based reform, which calls for clear, measurable academic standards that all students are required to master, test-based assessment became a common form of evaluation used to understand the access to equal education that schools are extending to all students. In some instances, standardized testing is used as a summative assessment method in tandem with formative and interim (e.g., daily interactions and observations) assessments that identify learning problems or inform instructional adjustments. These practices are referred to as low-stakes testing. At other times, standardized summative assessments are used to determine whether individual students have access to opportunities such as graduation and grade promotion—a practice that the field of education terms high-stakes testing. High-stakes testing is defined as using any test to make important decisions about students, educators, or schools.

High-stakes testing assessment practices have increasingly become the linchpin of larger state and federal accountability policies, such as the No Child Left Behind Act, that are premised on remedying the achievement gap and serve as a compliance mechanism for the larger mandates put forth in *Brown*. These systems are premised on the notion that equal access to educational opportunities can only be realized through a system that rewards progress toward that end goal and applies sanctions for failure to make progress or meet the goal. Accountability policies can be used in ways that punish or deny opportunities, such as retaining students in a grade, firing educators, or closing schools, or in ways that give rewards, such as providing access to higher level classes and providing salary increases or bonuses to educators.

High-stakes testing accountability policies that serve as methods for ensuring equal access to educational opportunities for all students operate on three notable assumptions: First, that standardized testing is an accurate measure of a student's knowledge of subject matter, quality of instruction, and the quality of the education a student has been afforded; second, that schools will be driven to use student-centered practices for improving access to educational quality for all students—regardless of characteristics such as race, socioeconomic status, and gender—as a consequence of the rewards and punishments (sanctions) linked to performance measures; and finally, that students who share common characteristics are homogeneous, whereby one-size-fits-all remedies for extending quality education suffice. Consistent in these assumptions is a framing of students and their educability based upon standardized, test-based performance; this is simultaneously reinforced by a reward structure that labels schools and allocates resources accordingly.

Research and Practice

Today, many argue, schools are just as segregated as they were prior to the *Brown* decision. Notwithstanding persistent issues of race and integration that permeate public school institutions and society alike, assessment and evaluation policy approaches post-*Brown* have presented a disconnect between their premised theory of ensuring equal access to quality educational opportunities and the practice of using standardized performance measures that label and punish students and schools. The social construction of students and schools as "failing" based upon test-based performance measures emulates historical narratives of inferiority, a positioning that many scholars say can deflect from conversations related to resource allocations and relative student needs. Observing how states shift their policies and practices in response to the Every Student Succeeds Act, which was enacted in 2015 to replace the No Child Left Behind Act and changed how schools report performance and are identified for improvement, will be important to the broader conversations on assessment and equity.

Research and researchers play an important role in how the field of education understands access to equal educational opportunity among students of color in the context of post-*Brown* mandates. These different ways of knowing lead to divergent inferences—in some instances wedded to strict, disciplinary rules and parameters (e.g., theoretical frameworks, scope of research questions, or limits of research methodology). In other cases, analyses are interconnected to historical and sociopolitical understandings of education issues (or lack thereof).

In some cases, researchers contribute to, and participate in, discourses that operate within the failure and success dichotomy to discuss the extent to which students are achieving and schools are affording access to educational opportunities. These analyses often evaluate student test score data with an eye toward providing public disclosure of students' educational standings and equated to test-based measures. Proponents of this approach often argue that public transparency will motivate communities to take active roles in responding to undesirable results and subsequently hold their schools accountable.

One point of consideration related to this approach is that it can undermine the ability to obtain a global portrait of academic trajectories and comparisons. By extension, narrow evaluations and inferences contribute to actions such as shaming schools that can produce negative effects thereby ultimately blaming the victim (e.g., students and struggling schools). The focus on test-based measures also isolates symptoms associated with educational inequities devoid of an in-depth understanding of structural, root causes. Finally, a sole reliance on student test scores elevates the visibility and legitimacy of the achievement gap as a framework for understanding inequality and students' access to educational opportunities.

Understanding access to equal education through lenses that are inclusive of sociocultural aspects of education informs yet another approach taken by researchers. In particular, these analyses attend to issues of curriculum, instruction, and resource allocation as being interrelated to equal education. Examples and points of consideration related to this approach include the limits of understanding *Brown* and equal access through a Black–White paradigm that places a sole focus on racial comparisons. This focus alone can leave silent issues of socioeconomic status, gender, and language—all indicators that are similarly used to separate and provide differentiated access to quality instruction among students; and finally, the idea that sameness (i.e., equality) does not equate to fairness (i.e., equity). In particular, the argument that students have relative needs that must be considered when assessing and evaluating their access to equal education.

Finally, emerging research approaches an examination of *Brown* with a focus on sociopolitical factors. These approaches consist of structural and institutional analyses that evaluate decision-making processes, the ethics of measurement and the legitimacy of testing for addressing educational equality, and the political tensions involved in (re)distributing access to educational opportunities. Examples and points of consideration related to this approach include how assessment reconstitutes inequalities and hierarchies in a manner that makes them appear natural while still maintaining a larger system of merit that partly inspired the *Brown* case, and, finally, the role of special interests in using student assessment and the evaluation of schools as a mechanism to privatize education and offset the constitutional mandate to fund public schools.

Patricia D. López

See also Accountability; High-Stakes Tests; Minority Issues in Testing; Paradigm Shift; Policy Evaluation

Further Readings

Balkin, J. M. (2001). What Brown v. Board of Education should have said: The national top legal experts rewrite America's landmark civil rights decision. New York: New York University Press.

Bell, D. (2004). Silent covenants: Brown v. board of education and unfilled hopes for racial reform. New York, NY: Oxford Press.

Frankenberg, E., & DeBary, E. (2011). Integrating schools in a changing society: New policies and legal options for a multiracial generation. Chapel Hill: The University of North Carolina Press.

Haney, W. (1984). Testing reasoning and reasoning about testing. Review of

Educational Research, 54(4), 597–654.

López, P. D. (2012). The process of becoming: The political construction of Texas' lone STAAR system of accountability and college readiness (Doctoral dissertation), University of Texas at Austin.

Nieto, S. (2004). Black, White, and US: The meaning of Brown v. Board of Education for Latinos. Multicultural Perspectives, 6(4), 22–25.

Scheurich, J. J., & Skrla, L. (2004). Educational equity and accountability: Paradigms, policies and politics. New York, NY: Routledge Press.

Legal Citations

Brown v. Board of Education of Topeka I, 347 U.S.483 (1954).

Brown v. Board of Education of Topeka II, 349 U.S. 294 (1955).

Clive R. Boddy Clive R. Boddy Boddy, Clive R.

Bubble Drawing Bubble drawing

226

227

Bubble Drawing

The bubble drawing is a type of projective or enabling technique for research use. It facilitates research participants in describing their thoughts and feelings in relation to a research question of interest. Its advantage is that it is a relatively quick and easy way of accessing and understanding the more emotional considerations in educational choices. This technique can be used in any type of research including educational research. This entry further describes the technique and how it is used.

The drawing is given to research participants (respondents) as a catalyst or stimuli to further discussion. Typically in the drawing, two people are talking to each other with speech bubbles coming out of their mouths and thought bubbles emerging from their heads (minds). The research question is encapsulated in the speech bubble of one of the people (or objects) in the drawings. The speech and thought bubbles of the other person are empty, and it is the job of the research participant to complete these in answer or response to what has been asked or said in the other speech bubble.

For example, if a university or college was researching its attractiveness to students, it may present research participants with a drawing of a building to represent the institution and a drawing of a person to represent the potential student. The university might be saying something like “Why not come and study here, we have a great reputation?” The research participant would be asked to complete, by writing within the bubbles, what the other person—in this case, a potential student—was saying and thinking.

In reply, the speech bubble would tend to contain answers that reflect the rational aspects of the choice of educational establishment. These answers would be more or less the same as would be gained from a similar question in a be more or less the same as would be gained from a similar question in a questionnaire. However, the thought bubble typically contains the underlying and/or emotional and otherwise rarely articulated concerns of the student regarding the particular university or college. This often provides valuable and informative insights.

Just as with other projective techniques, the advantage of the bubble drawing is that it depersonalizes the answers because the research participants are told to fill in the speech and thoughts of the other person in the drawing (not to answer on their own behalf). This depersonalizes the answer and thereby removes some of the potential sensitivity of the answers that are given. This in turn enables the researchers to get a deeper understanding of the “real” concerns of the research participants. This is because the research participants unself-consciously project their own thoughts, feelings, and concerns onto the person in the drawing.

The use of a bubble drawing in research is also assumed to stimulate the nonverbal, less rational, and more emotional parts of the brain, thus facilitating answers that may otherwise be difficult to obtain because they are less consciously considered by the research participant. For example, the bubble drawing technique could be useful in researching the emotional reasons for undergraduate withdrawal (C. R. Boddy, 2010)because it may otherwise be too sensitive for failing students to openly discuss their emotions concerning loneliness, isolation, and academic bewilderment. Their answers to more rational and direct investigations may be biased by social desirability bias—the tendency to give answers that are more socially acceptable or that portray the respondent in a better light.

Using the bubble drawing technique should give a fuller and more comprehensive understanding than that gained from direct questioning alone, because the indirectness of the approach helps to get around, and deactivate, the conscious defenses of research participants.

The bubble technique is also unusual from the research participants’ point of view and can be seen as a nonthreatening and more interesting type of question, one that research participants find stimulating to engage with and even enjoyable to complete. This again puts the research participants more at ease and facilitates unguarded and open replies.

Similarly, instructors have used the bubble drawing technique to gain an understanding of their teaching style and effectiveness from the students’ point of view. This research aimed to provide the lecturer with information from students in order to help develop student-centered learning and teaching opportunities. One issue uncovered related to the (quick) speed of delivery of lectures (to a largely international group of students). Peer observations of the same teaching (by native English speakers) had found that delivery was well timed rather than being too fast. This insight highlights the usefulness of a student-centric approach to research.

The bubble technique is often used in qualitative research, but it can also be incorporated into a quantitative questionnaire. The responses can then be coded and the themes and codes be analyzed statistically. In qualitative research, the research participants can be asked to discuss their answers to stimulate even further discussion and deeper understanding.

Clive R. Boddy

See also Collage Technique; Projective Tests; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Boddy, C. R. (2004). From brand image research to teaching assessment: Using a projective technique borrowed from marketing research to aid an understanding of teaching effectiveness. Journal of Quality Assurance in Education, 12(2), 94–105.

Boddy, C. R. (2010). A paper proposing a projective technique to help understand the non-rational aspects of withdrawal and undergraduate attrition. Ergo: The Journal of the Education Research Group of Adelaide, 1(3), 11–20.

Stephanie Schmitz Stephanie Schmitz Schmitz, Stephanie

Buros *Mental Measurements Yearbook* Buros *mental measurements yearbook*

227

228

Buros *Mental Measurements Yearbook*

The Buros *Mental Measurements Yearbook* (MMY) is a comprehensive compilation of test reviews oriented to test consumers. It is published by the Buros Center for Testing, University of Nebraska–Lincoln within the Educational Psychology Department in the College of Education and Human Sciences. The first MMY was published in 1938; Volume 19 was published in 2014. The MMY is well respected across fields and is used to find tests appropriate for making decisions about employment, as a reference, for discussion within college courses related to psychological testing, and in the courts.

Reviewed tests within the MMY represent a variety of fields, including psychology (e.g., personality, intelligence, behavior), vocational, education, and the business community. To be reviewed in the MMY, each test must meet several criteria; specifically, they must be available commercially, published in English (although some Spanish tests have been reviewed in recent volumes), and new or revised, as well as widely used since the publication of the previous volume. Additionally, supporting documentation of the technical properties of the test has been required following publication of the 14th volume.

Two reviewers holding a doctorate and having psychometric training conducted most of the reviews in the MMY. A typical entry for each test consists of both a description and an evaluation of the instrument. The description section typically includes such information as the name, author, and publisher of the test; the publication date; the purpose and an overview of the test; a description of test materials; and scoring information.

The evaluation section comprises several different types of information. First, information is provided on the development of the test, including its underlying information is provided on the development of the test, including its underlying theories and/or assumptions as well as the process used to develop individual test items. Second, the psychometric properties of the test are reviewed, including information on the reliability and validity of the instrument as well as on the standardization of the test, in which information on the norm sample is discussed. Third, both an overall summary of the test, in which the strengths and weaknesses of the test are highlighted, and a conclusion on and recommendations about the test's quality are provided.

Each volume of the MMY is available in print, and after the Ninth Edition, online. If accessing the MMY online, a brief description of a selected test and publisher contact information is available for free, while a full review of the test may be purchased through the Buros Center for Testing website.

Stephanie Schmitz

See also Intelligence Tests; Personality Assessment; Psychometrics

Further Readings

Buros Center for Testing. (n.d.). Information for reviewers. Retrieved from http://buros.org/reviewers

Buros Center for Testing. (n.d.). Mental measurements yearbook. Retrieved from http://buros.org/mental-measurements-yearbook

Carter, N. F. (2011). Mental measurements yearbook. Reference … User Services Quarterly, 41, 181.

Plake, B. S., & Conoley, J. C. (1995). Using Buros Institute of Mental Measurements materials in counseling and therapy (ERIC Digest ED391987). Retrieved from http://files.eric.ed.gov/fulltext/ED391987.pdf

Szostek, J., & Hobson, C. J. (2011). Employment test evaluation made easy: Effective use of mental measurements yearbooks. Employee Relations Law Journal, 37, 67–74.

C

Scott Bishop Scott Bishop Bishop, Scott

c Parameter *c* parameter

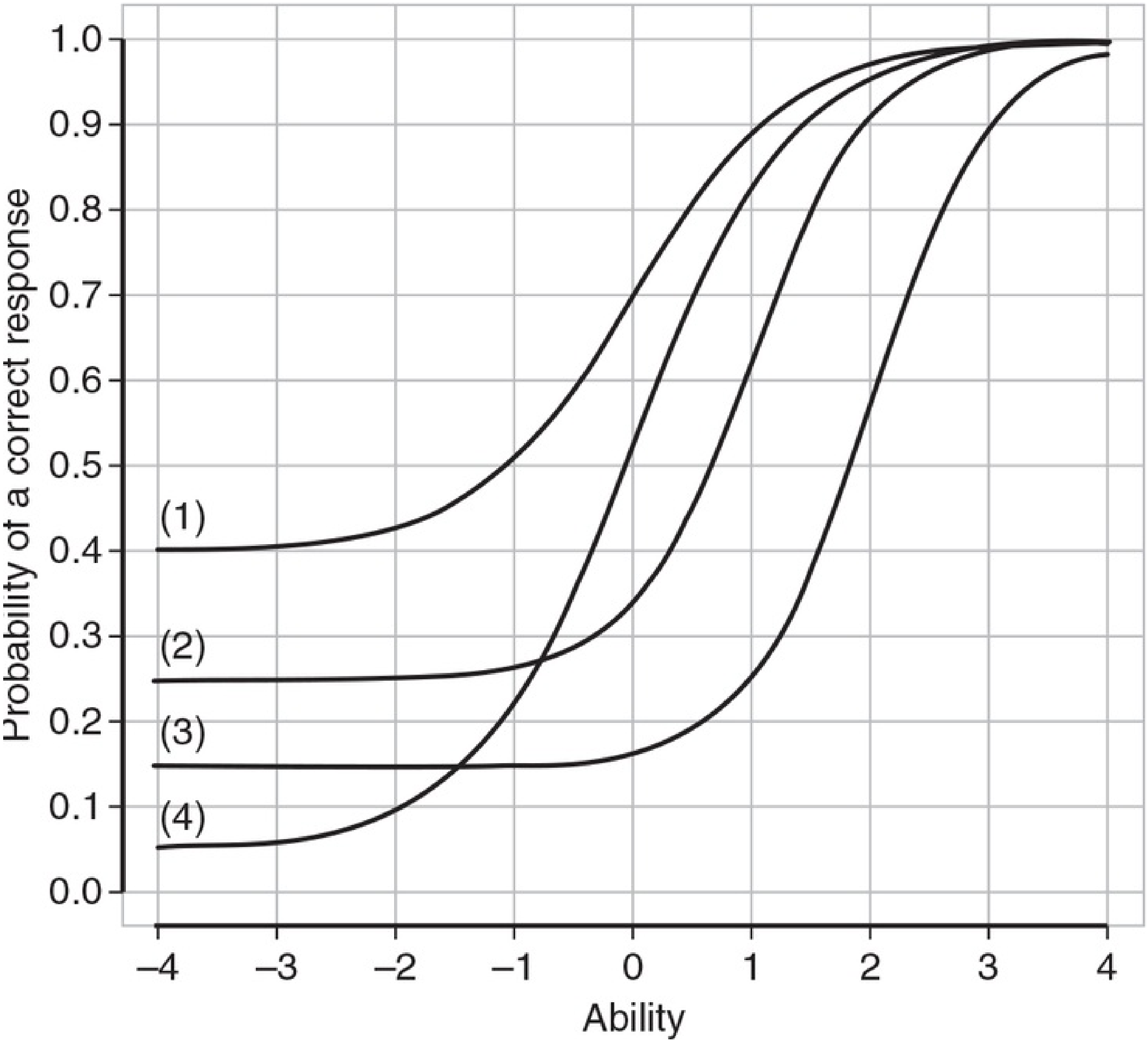
229

231

c Parameter

In item response theory (IRT), the *c* parameter is the lower asymptote of an item characteristic curve (ICC). It is used in the three-parameter logistic (3PL) IRT model, and it is often referred to as the *pseudo-guessing parameter*. In IRT, the probability that an examinee will make a particular response to a test item is modeled on the examinee's standing on the trait that the test measures. As an example, for a mathematics achievement test, IRT can model the probability that an examinee will earn a particular score on a mathematics test item based on the examinee's achievement in math. The ICC visually maps the relationship between an examinee's ability, usually denoted by θ, and the examinee's probability of making a particular item response (Figure 1).

Figure 1 Item characteristic curves for four hypothetical items



Why Is the *c* Parameter Needed?

Selected-response item formats, such as multiple-choice and true–false items, are frequently used on assessments because of their efficiency; selected-response items can sample more of the content domain that a test covers, per unit of testing time, than other item formats. However, some examinees can answer these items correctly as a result of some degree of chance, perhaps depending on how many response options the examinee can accurately eliminate. Guessing is generally not a concern with constructed-response item types. The *c* parameter helps IRT account for an examinee with extremely low ability correctly answering selected-response items.

Unidimensional IRT Models for Dichotomous Items

The most commonly used IRT models assume that (a) a single ability underlies an examinee's response process and (b) items are dichotomously scored (e.g., right *vs*. wrong). The relationship between the probability *p* of a correct response to a dichotomously scored item *i* and examinee ability θ has a monotonically increasing ICC that is roughly *s*-shaped, although it can be compressed and stretched at various points along the ability scale (see Figure 1). Several parameters are used within IRT to control the position and shape of the ICCs. The 3PL model has the following:

1. a discrimination parameter, denoted *a*, that controls the slope of the ICC;
2. the difficulty parameter, denoted *b*, that indicates the point on the θ scale where there is an inflection (i.e., where the concavity of the curve changes) in the ICC; and
3. a pseudo-guessing parameter, denoted *c*, that represents the lower asymptote of the ICC.

The probability of a correct response can then be calculated as:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}.$$

Figure 1 shows 3PL ICCs for four hypothetical items. All 4 items have *b* parameters that are fairly high in value to allow the lower asymptote to be visually prominent at the lowest θ values shown in this graph. Depending on any given item's *b* parameter value and the range of θs shown in the graph, the lower asymptote may not always be as distinct. As shown in Figure 1, the higher the value of *c*, the more likely it is that examinees with extremely low ability will answer the item correctly.

The theoretical value of the *c* parameter should be 1 divided by the number of response options for the item. For an item with two response options, like a true–false item, the theoretical value would be 1/2, for an item with three options, the suspected value would be 1/3, and so on. Although these theoretical expectations are reasonable, in reality, they do not hold in most cases. Very frequently, *c* is lower than the inverse of the number of response options. It is well-known that experienced item writers use common student misconceptions in the incorrect response options and that will make the empirical *c* value lower than the theoretical value. Of course, the empirical *c* value can also be higher than the theoretical value, perhaps when some misconceptions are not very common among students.

It can be informative to compare the 3PL to other univariate IRT models that do not include a *c* parameter. The two-parameter logistic model does not have the pseudo-guessing parameter (or equivalently, one can consider $c = 0$). The one-parameter logistic model also does not have the pseudo-guessing parameter, *c*, and all items are assigned a common slope, meaning items are modeled with only the *b* parameter. Inclusion of the *c* parameter generally results in improved data–model fit for selected-response items. However, IRT software can have difficulty estimating *c* in some cases, such as when there are a small number of test takers. In these cases, one can use a model without *c*, set a prior on *c*, or fix *c* to a specific value (perhaps to a value near to, or just under, the theoretical value). The *c* parameter also has an effect on where an item's maximum information occurs. In one-parameter and two-parameter logistic models, the maximum information occurs at the item's *b* parameter. However, the addition of the *c* parameter changes the location of maximum information, so that it is somewhat higher than the value of the *b* parameter.

Other IRT Models

There are unidimensional IRT models for examinee response processes regarding items that are polytomously scored, and their use is increasing. The most frequently used polytomous IRT models—the generalized partial credit model and the graded response model—do not include a *c* parameter. However, there is multidimensional IRT extension of the 3PL model. Although there are as many *a* parameters as there are dimensions of the multidimensional IRT model, there is only one *b* parameter and one *c* parameter.

Comparison to Classical Psychometrics

In traditional psychometrics, there are analogs to the IRT *a* and *b* parameters (e.g., biserial correlations and item means). However, there is no real-item statistic that compares to the IRT *c* parameter. At the test level, a correction for guessing can be applied to an examinee's total test scores, in which a fractional value for every incorrect answer is subtracted from the examinee's total raw score. For items that have five answer options, the fractional correction is one fourth of a point; for items that have four answer options, the fractional correction is one third of a point; and so on.

Scott Bishop

See also *a* Parameter; *b* Parameter; Item Response Theory

Further Readings

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159–176. doi:10.1177/014662169201600206

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401–412. doi:10.1177/014662168500900409

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Allison Jennifer Ames Allison Jennifer Ames Ames, Allison Jennifer

Jonathan D. Rollins Jonathan D. Rollins Rollins, Jonathan D. III

C Programming Languages C Programming languages

231

234

C Programming Languages

The family of C programming languages, consisting of C, C++, C#, and Objective-C, is a set of similar languages from different paradigms. The first of these languages, C, serves as the foundation for the set because its syntax, structure, and logic strongly influenced the development of the latter languages. Perhaps the most notable difference is that C# and Objective-C are object-oriented languages using classes that gain additional programming properties to build on the C language. The C++ language is inherently capable of object-oriented programming, but its flexibility also allows it to be programmed in the same paradigm as C. Due to the large number of similarities between them, a deeper understanding of C facilitates learning the other languages, as its programming principles can be generalized to the others. For that reason, this entry focuses on C, providing programming principles, sample code, compilation, and applications to educational research, measurement, and evaluation.

Programming Principles in C

Programming languages follow *paradigms* or ways of writing code at different levels of abstraction (i.e., how closely the language resembles the binary machine code of 0s and 1s). C follows an imperative (and more specifically, procedural) paradigm, which means that the code describes *how* to perform a task.

C code accomplishes tasks using *functions,* a term that bears similarity to the mathematical definition. Functions use input to produce output, although this is not always the case. Some functions perform tasks with no explicit input values required, and others use input to perform a task with no specific output values. The *main* function is a requirement in C code when it is compiled to produce an executable program.

A collection of functions written in a single file to be used across multiple programs is called a *library* (which may have a static library file extension, *.a or* .lib, or dynamic library file extension, *.so or* .dll). A library may contain header files (with the file extension *.h), which are preprocessing directives that are evaluated before source code (i.e., the program that a user writes). A primary example of this is the C Standard Library, and it contains a standardized set of header files that allows for basic commonly used functions to be used by programs. More specifically, a header file (e.g., <stdio.h>) contains the standard input/output functions that are not reserved words in the C language. Under the current C11 standard, there are 44 reserved words in C that cannot be used for naming variables or functions; for instance, a variable cannot be named *int* because the compiler understands that term to denote the integer data type. If the end goal is a user-written library, an executable program is not required because libraries typically do not have an execution thread of their own.

There are four basic data types that can be specified in C. First, arithmetic data types are used to store alphanumeric symbols in memory (i.e., *char, int, float,* and *double*). Second, type modifiers can be used to denote the possibility of negative values (i.e., *signed* vs. *unsigned*) or the amount of memory needed (e.g., the number of decimal places) for that particular value (i.e., *short* vs. *long*). Third, enumerated data types are useful for situations involving loops and discrete calculations, such as dummy codes and categorical indicators. Finally, void data types are null values that serve as a proxy for functions returning no values or for declaring pointers to variable locations in memory without specifying a particular data type. However, the classification of these data types could be conceptualized in multiple fashions. It is even possible to declare constants, as opposed to variables, for those values in memory that should not be changed during code execution. Declaring variables is technically separate from initialization. The former creates and names the variable, while the latter gives the variable a starting value.

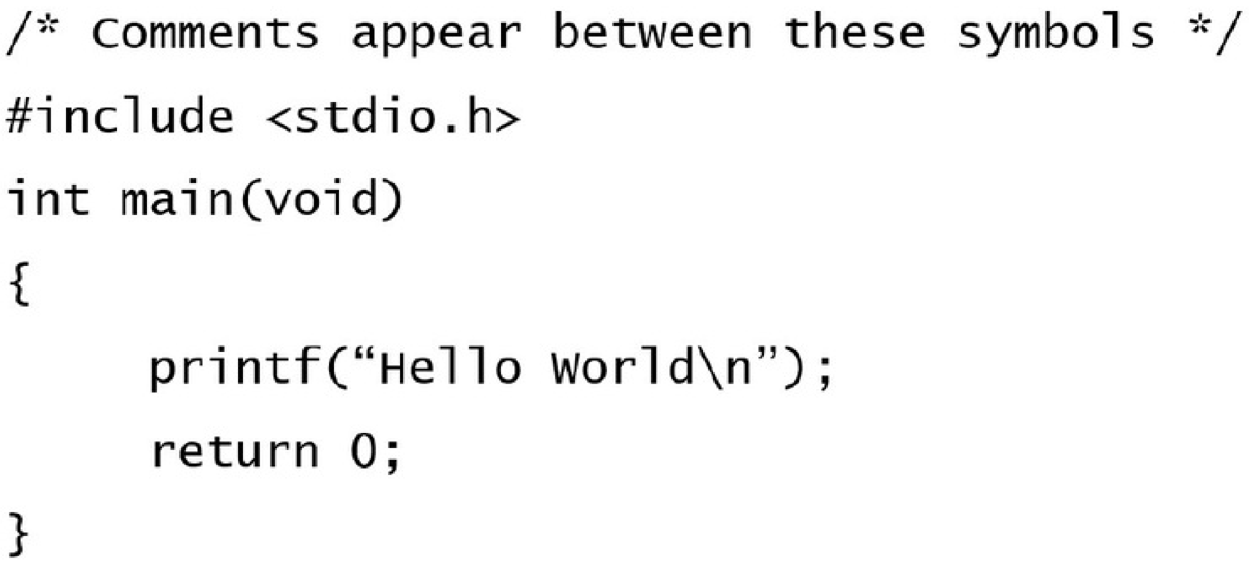
Derived data types are possible as well. That is, there are more complex data types defined using the basic data types described earlier. As an illustration, an array is a collection of elements of the same type within a predetermined-sized range. Conceptually, this can be thought of as a table of values indexed (with the first value starting at an index of 0 as opposed to 1) by dimension using subsequent brackets (e.g., *data*[0][9] would access the value in the first row and 10th column of the two-dimensional array named *data*). A common example of an array is a string because it is a collection of individual characters. Another example of a derived data type is a pointer, which is a variable that holds the memory address of another variable. Furthermore, functions are derived data types because the return value must be a particular basic data type. All in all, there is static type-checking with all of these various data types. This means that the consistency of data types is checked whenever the program is compiled (i.e., created) as opposed to during run time of the program.

The functionality of the code tasks is further manipulated using control structures. These can route logic conditionally (e.g., *if-then-else* and *switch* statements) or iteratively (e.g., *for*, *while*, and *do-while* loops). Additionally, operators can be used inside or outside of control structures to perform actions with the variables in memory. Types of operators include arithmetic (e.g., +, –, ×, /, %), comparison (e.g., ==, !=, <, >), assignment (e.g., =, +=, –=), and logical (e.g., *true, false, not, and*). Furthermore, when multiple operators are used, the grouping of expressions and the order of their evaluation is affected by the precedence and associativity of the operators. For example, the use of parentheses around an expression could change the order in which a mathematical expression is computed: 2 × 2 + 2 versus 2 × (2 + 2) results in 6 and 8, respectively. Reference tables have been created that describe the order of operator precedence.

Example Code

One of the most traditional code examples is writing a basic program that outputs “Hello World” to the terminal window. A variant of this common code is presented in Figure 1. The first line begins with code comments placed between /* and */ to denote the beginning and end of a comment, respectively. Text between these two placeholders can span multiple lines. The importance of commenting code cannot be emphasized enough, especially if code is shared with others or is to be revisited at a later date.

Figure 1 Example of C programming code



```
/* Comments appear between these symbols */
#include <stdio.h>
int main(void)
{
      printf(“Hello World\n”);
      return 0;
}
```

The code example continues with an #*include* preprocessor directive to tell the C compiler that the functions in the standard input/output header file (<stdio.h>) should be made available for the subsequent code to use. The third line begins the declaration of the *main* function, which returns an integer data type as the output. The *void* given to the function is not required technically but is shown here to illustrate an assumption that the compiler would otherwise make. The code within the function is indented for readability purposes. This is a very useful practice in programming because C compilers ignore such spacing.

Within the curly braces on Line 5 of Figure 1, the *printf* function is used to send formatted output to the terminal window. The content inside the double quotes gets printed on the screen. Inside of the quotes, any character that immediately follows a backslash (the combination of which is known as an *escape sequence*) is ignored, so that the compiler can receive additional instructions. The *n* tells the compiler to create a line break. The end of the statement ends with a semicolon, as all expression or declaration statements should. However, preprocessor directives (i.e., *#include*) and control structures (i.e., loops and conditional logic) do not require a semicolon, as evidenced in the example code. Finally, Line 6 returns the integer zero to the terminal to indicate that the program executed correctly. Typically, nonzero values are used by programmers to denote warning or error messages in more complex coding.

Code Compilation

In order to develop an executable program, one must use a compiler. The purpose of the compiler is to take source code (e.g., the C code written by the purpose of the compiler is to take source code (e.g., the C code written by the user) and translate it to the lowest level of software as a series of 0s and 1s (corresponding to the absence or presence of electrical activity in the central processing unit) that the machine hardware can understand directly. This process occurs through several intermediate steps.

Making the assumption that the code is free of compilation errors, both syntactical and logical in nature, the source code will first be modified by the preprocessor. The preprocessor is used to define constants used globally in the program, substitute macro functions in the code with those from library files, and provide directives that stipulate which libraries to include and which code should be compiled according to conditional logic. Following this, the compiler translates the preprocessed code into assembly code. It is similar to machine code but has mnemonics (simpler words, symbols, and numbers) to represent the binary code. Next, the assembler translates the assembly code directly into machine code. After this point, the linker links any libraries directly to the machine code. The final result is an executable binary file.

Using an integrated development environment is perhaps the simplest method for all of the preceding steps to be performed, as it contains a text editor, compiler, and linker. Although many compilers exist, a more commonly used one is the GNU (a recursive acronym for “GNU’s not Unix”) C Compiler. Windows users would use part of the GNU C Compiler known as Minimalist GNU for Windows (MinGW), Linux users typically have access to a GNU compiler that is included in the operating system distribution, and Mac users may use the Apple Developer Tools for compilation.

Applications in Educational Research, Measurement, and Evaluation

Although many software applications used by researchers, measurement practitioners, and evaluators are written in higher level languages, often with extensive graphical user interfaces, C code has an advantage of being very quick for statistical and psychometric calculations. As such, it is not unusual for other programming languages to call dynamic link libraries (*.dll) written in C to quickly and efficiently perform routine tasks. Fortunately, there are repositories of C code, available under open-source licensing conditions, related specifically to psychometric analyses.

One such repository is provided by the Center for Advanced Studies in Measurement and Assessment from the College of Education at the University of Iowa. On their website, there are multiple projects that contain C executable programs. In particular, the *Equating Recipes* project contains over 25,000 lines of C source code for users to use, or appropriately adapt, as needed. This resource provides excellent didactic examples and can feasibly be integrated into research projects and operational practices. Another library of functions, freely available for those interested in item response theory, is the *libirt* library. It contains functions for parameter estimation of parametric and nonparametric models for dichotomous and polytomous data. Multiple estimation algorithms are available for users to implement as well.

C undergirds major software systems and environments that are often used in application. Over half of the core functions in the R programming language and environment are underwritten using C. Because of the open-source nature of R, many of the C functions used for statistical/psychometric calculations can be viewed through the Comprehensive R Archive Network. While the previous examples have highlighted resources that are freely available, C is even used for commercial software. For example, SAS is written in the C language. More specifically, SAS *data* and *proc* steps are interpreted and correspond with C code that is executed incognito. Overall, both R and SAS can be used to call compiled C code, although this functionality extends to other statistical languages and environments as well.

Allison Jennifer Ames and Jonathan D. Rollins III

See also Computer Programming in Quantitative Analysis; Quantitative Research Methods; R; SAS

Further Readings

Banahan, M., Brady, D., & Doran, M. (2007). The C book (2nd ed.). Retrieved from http://publications.gbdirect.co.uk/c_book (Originally published by Addison-Wesley, 1991).

Gookin, D. (2014). Beginning programming with C for dummies. Hoboken, NJ: For Dummies, a Wiley Brand.

International Organization for Standardization. (2011–2012). ISO/IEC 9899:2011 (3rd ed.) [C11 Standard]. Geneva, Switzerland: Author.

Kernighan, B., & Ritchie, D. (1988). The C programming language. Englewood Cliffs, NJ: Prentice Hall.

McGrath, M. (2012). C programming in easy steps (4th ed.). Warwickshire, UK: In Easy Steps.

Zhang, T. (2000). Sams teach yourself C in 24 hours (2nd ed.). Indianapolis, IN: Sams.

Websites

Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa. Computer programs: http://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Comprehensive R Archive Network: http://cran.r-project.org

libirt, Item Response Theory Library: http://psychometricon.net/libirt/

Raman Arora Raman Arora Arora, Raman

Canonical Correlation

Canonical correlation

234

236

Canonical Correlation

Canonical correlation is a statistical measure for expressing the relationship between two sets of variables. Formally, given two random vectors **x** ∈ R^{dx} and **y** ∈ R^{dy} with some joint (unknown) distribution *D*, the canonical correlation analysis (CCA) seeks vectors **u** ∈ R^{dx} and **v** ∈ R^{dy}, such that the random vectors when projected along these directions, that is, variables *u* > *x* and *v* > *y*, are maximally correlated. Equivalently, we can write CCA as the following optimization problem: find **u** ∈ R^{dx}, **v** ∈ R^{dy} that:

$$\text{Maximize}_{dx\ dy}\ \rho\left(u > x, v > y\right),$$

$$\mathrm{u} \in \mathrm{R},\ \mathrm{v} \in \mathrm{R}$$

where the correlation, ρ(*u* > *x*, *v* > *y*), between two random variables, is defined as . Assuming that vectors **x** and **y** are 0 mean, we can write CCA as the problem var(*u* > *x*) var(*u* > *x*) of finding *u* ∈ R^{dx}, v ∈ R^{dy} that: (2)

Affine Invariance

A simple observation suggests that if we scale either *u* or *v* or both, the objective does not change. That is, if *u* 7→ α*u* and/or *v* 7→ β*v*, the objective does not change. This has a profound implication—canonical correlation is statistical measure that is invariant to affine transformations, unlike other measures, for instance, those optimized by principal component analysis and partial least squares.

CCA as a constrained optimization problem: Because the CCA objective is affine invariant, we might as well choose the scaling coefficients, α, β ∈R such that:

$$\mathbb{E}\left[u^\top xx^\top u\right]=1, \mathbb{E}\left[v^\top yy^\top v\right]=1.$$

This yields an equivalent constrained optimization problem:

$$\text{maximize } \mathbb{E}(x,y)\left[u^\top xy^\top v\right],$$

$$u \in \mathrm{R}^{dx}, v \in \mathrm{R}^{dy}$$

$$\text{subject to } \quad \mathbb{E}_x\left[u^\top xx^\top u\right]=1, \mathbb{E}_y\left[v^\top yy^\top v\right]=1.$$

CCA Solution

We can show, using the Lagrange multiplier’s method, that the solution to the constrained optimization problem above is given by choosing *u* to be the top eigenvector of C_{-xx1} C_{xy} C_{-yy1} C_{yx} and choosing *v* = C−*yy*√1C$_{\lambda}$*yxu*.

Simultaneous Solution to CCA

Subsequent CCA directions can be found by deflation by constraining them to be uncorrelated to previous ones. Alternatively, we can solve for to p*k* CCA dimensions simultaneously by solving for the following optimization problem: Find that

$$\text{maximize: } \mathbb{E}_{(x,y)\sim D}\left[\text{trace}\left(U^\top xy^\top V\right)\right]$$

$$\text{subject to: } \mathbb{E}\left[U^\top xx^\top U\right]=I_k, \mathbb{E}\left[V^\top yy^\top V\right]=I_k.$$

In other words, CCA can be posed as the problem of finding the most correlated *k*-dimensional subspace of *D* and columns of *U* are given as the top-*k* eigenvectors of C_{-xx1} C_{xy} C_{-yy1} C_{yx} 29th Conference on Neural Information Processing Systems (2016), Barcelona, Spain.

CCA as Minimizing Reconstruction Error

Given *n* data (x_{i},y_{i}) in $R^{\mathrm{dx}} \times R^{\mathrm{dy}}$, from an unknown *D*, find $U \in \mathrm{R}^{\mathrm{dx} \times \mathrm{k}}$, $V \in \mathrm{R}^{\mathrm{dy} \times \mathrm{k}}$

$$\text{minimize: } \mathbb{E}_{x,y\sim D} \| U^\top x - V^\top y \|_2^2,$$

$$\text{subject to: } U > C_{xx} U = I_k, V > C_{yy} V = I_k.$$

Expand the objective:

$$\mathbb{E} \| U^T x - V^T y \|_2^2$$

$$= \mathbb{E}\left[x^\top U U^\top x\right] - 2\mathbb{E}\left[x^\top U V^\top y\right] + \mathbb{E}\left[y^\top V V^\top y\right]$$

$$= E^h \mathrm{trace} U^\top x x^\top U)]$$

$$- 2\mathbb{E}\left[\mathrm{trace}\ U^\top x y^\top V\right] + E^h \mathrm{trace}\ V^\top y y^\top V)].$$

History

CCA is a classical technique in multivariate statistics, proposed by Harold Hotelling in 1935, for measuring correlations between two random vectors. Hotelling first studied this problem to predict success in college. In his 1935 article, titled *The Most Predictable Criterion*,” he argued that no single regression equation provides fully adequate solution and that regressing on the dependent variate with the largest multiple correlation with predictors yields best accuracy.

Applications

CCA is widely used in multivariate analysis, finance, management sciences, chemometrics, bioinformatics, and neuroscience. CCA has been successfully applied to various tasks in speech, natural language processing, and computer vision. CCA admits a nonstandard stochastic optimization problem where not only the objective but the constraints are stochastic or equivalently the objective is a ratio of two expectations rather than an expectation of a loss function. Consequently, the CCA objective does not decompose over the sample and designing stochastic approximation algorithms for CCA remains a challenging open problem.

Raman Arora

See also Correlation; Multivariate Analysis of Variance; Multiple Linear Regression; Part Correlations; Partial Correlations

Further Readings

Adrian, B., Raman, A., & Mark, D. (2016). Learning multiview embeddings of Twitter users. In ACL.

Aria, H., Percy, L., Taylor, B.-K., & Dan, K. (2008). Learning bilingual lexicons from monolingual corpora. In ACL.

Hotelling, H. (1935). The most predictable criterion. Journal of Education Psychology, 26, 139–142.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28(3/4), 321–377.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 16(12), 2639–2664.

Paramveer, D., Dean, P. F., & Lyle, H. U. (2011). Multiview learning of word embeddings via CCA. In Advances in Neural Information Processing Systems (pp. 199–207).

Raman, A., Andrew, C., Karen, L., & Nathan, S. (2012). Stochastic optimization for PCA and PLS. In Allerton Conference (pp. 861–868).

Raman, A., & Karen, L. (2012). Kernel CCA for multiview learning of acoustic features using articulatory measurements. In Proceedings of the MLSLP.

Raman, A., & Karen, L. (2013). Multiview CCA-based acoustic features for phonetic recognition across speakers and domains. In Proceedings of the ICASSP.

Sujeeth, B., Raman, A., Karen, L., & Mark, H.-J. (2008). Multiview acoustic feature learning using articulatory measurements. In M. B. Blaschko & C. H. Lampert (Eds.), International Workshop on Statistical Machine Learning for Speech Recognition, 2012. Correlational spectral clustering. In *CVPR*.

Weiran, W., Raman, A., Karen, L., & Jeff, B. (2015a). On deep multiview representation learning. In ICML.

Weiran, W., Raman, A., Karen, L., & Jeff, A. B. (2015b). Unsupervised learning of acoustic features via deep canonical correlation analysis. In ICASSP.

Weiran, W., Raman, A., Nati, S., & Karen, L. (2015). Stochastic optimization for deep CCA via nonlinear orthogonal iterations. In 53nd Annual Allerton Conference on Communication, Control and Computing.

Weiran, W., Raman, A., Karen, L., & Jeff, B. (2016). On deep multiview representation learning: Objectives and optimization. Retrieved from arXiv preprint arXiv:1602.01024.

Linda Dale Bloomberg Linda Dale Bloomberg Bloomberg, Linda Dale

Case Study Method

Case study method

236

239

Case Study Method

A case study is an in-depth exploration from multiple perspectives of the richness and complexity of a particular social unit, system, or phenomenon. Its primary purpose is to generate understanding and insights in order to gain knowledge and inform professional practice, policy development, and community or social action. Case study research is typically extensive; it draws on multiple methods of data collection and involves multiple data sources. This method culminates in the production of a detailed description of a setting and its participants, accompanied by an analysis of the data for themes, patterns, and issues. A case study is therefore both a process of inquiry about the case at hand and the product of that inquiry. The case study method is employed across disciplines, including education, health care, social work, history, sociology, management studies, and organizational studies. This entry outlines the defining characteristics of the case study method, provides types of case studies, and describes the role of the researcher.

Defining Characteristics of Case Studies

A review of case studies reported in the literature yields several defining characteristics:

- Clear boundaries—The researcher begins by identifying a specific case or set of cases to be studied. Each case is an entity that is described within certain parameters, such as a specific time frame, place, event, and process. Hence, the case becomes a *bounded system*. Typically, case study researchers analyze the real-life cases that are currently in progress so that they can gather accurate information that is not lost by time.

- Purposeful sampling—Selecting the case requires that the researcher define the unit of analysis and establish a rationale for why the particular case was selected in terms of purpose and intended use, why the specific boundaries were chosen to surround the case, and why specific categories of information were sought.
- Design flexibility—Reliance on a single source of data is typically not sufficient to develop the necessary in-depth understanding and insights. Many forms of qualitative data are therefore collected, including interview, direct observations, participant observation, and physical artifacts (audiovisual materials, documents, and archival records). In addition, some quantitative data, including survey and/or census information, may be collected to augment the qualitative data.
- Thick narrative description—Key to understanding the data is that the report provides thick narrative description of the case, including the current context, history, chronology of events, and a day-to-day rendering of the activities of the case. This description enables deeper understanding on the part of the reader.
- Thematic analysis—In addition to description, the researcher seeks to identify topics or issues that emanate from the findings and that shed light on understanding the complexity of the case. When multiple cases are selected, a typical format includes a detailed description of each case as well as reports of themes within each case (*within-case analysis*) followed by thematic analysis across cases (*cross-case analysis*). Themes aggregate information into larger clusters of ideas and illustrate similarities and differences. Themes can also be presented as a theoretical or conceptual model.
- Transferability—One myth about case study research is that findings cannot be applied beyond the cases studied. This viewpoint is based on statistical generalization, which relies on the use of representative random samples in order to extrapolate findings to a larger general population. Gaining a complex and rich understanding of the data through intense in-depth exploration means that the findings from just one case may hold a wealth of transferable information and knowledge that can be applied in other similar contexts, settings, and conditions. As such, transferability, rather than generalizability, becomes the goal of the case study method.

Types of Case Study Methodologies

Case Study Design

A single case can be selected for in-depth study, or several cases can be selected so that they can be compared. The intent or objective of conducting a case study plays an important role regarding the choice of research design, and there are three design variations: *intrinsic case study, instrumental case study*, and *collective* or *multiple case study*.

A single intrinsic case study can be conducted to illustrate a case that needs to be documented and described. The research focuses on the case itself to the extent that it represents a unique situation or holds intrinsic or unusual interest. Alternatively, the intent of the study may be to understand a specific issue, problem, or concern, and a case or cases would be selected as a vehicle to illustrate and better understand the underlying concern. This would be a single instrumental case study. Finally, if more than one case is involved, this would be a collective case study or multiple case study, with the intent to compare and contrast perspectives regarding the same issue. The focus is on the analysis of diverse cases to determine how it confirms the findings within or between cases or if it calls the findings into question.

Case Study Research Approach

Depending on the researcher's methodological perspective and the overall research questions, there are several types of case study approaches:

- Exploratory—This type of case study is selected for its data gathering possibilities, and also for what it may reflect or represent regarding other similar cases. Pursuing an exploratory design allows the researcher to gain new insights based on the data, with the goal of generating specific ideas or theories that might be used to test ideas regarding similar cases.
- Descriptive—A descriptive case study is selected when the researcher seeks to portray the specifics of a social phenomenon or issue that is not well conceptualized or understood. The goal is to seek rich detail regarding the inner processes of the given case and to provide multiple ways of understanding the layers of meaning inherent in the case through various data-gathering techniques.
- Explanatory/causal—This type of case study is usually associated with a quantitatively driven case study design, in which the researcher begins with

> a specific agenda or set of hypotheses to test. A qualitative approach entails comparing and contrasting the data as well as seeking evidence of negative cases (data that do not fit with the hypotheses) as a way to build validity for the findings.

Analytic strategies are chosen according to the unique opportunities and challenges the case presents. The approach to data analysis also differs depending on the research design and the intent of the study. Some case studies report on the case in its entirety (holistic analysis), whereas others involve the analysis of specific aspects within the case (embedded analysis). When multiple cases are examined, the typical analytic strategy is to provide detailed descriptions of themes within each case (within-case analysis), followed by thematic analysis across cases (cross-case analysis). In the final interpretive phase of analysis, the researcher derives conclusions from the findings and analysis and discusses the underlying meaning behind the findings.

Role of the Researcher

At the outset, a researcher must determine whether the case study approach is appropriate for analyzing the chosen research problem. A case study is a suitable approach when the researcher has a clearly identifiable and bounded case or cases and seeks to achieve an in-depth understanding of the case context. Selecting a case to study requires that the researcher establish a rationale for a purposeful sampling strategy as well as clear indications regarding the boundaries of the case. In many instances, case studies may not have clear beginning and end points, and deciding on boundaries that adequately surround the case can be challenging.

In conducting case study research, identifying and describing contexts (which are typically complex, overlapping, and multidimensional) are vital to generate deep meaning and convey understanding. The case is investigated from different angles by gathering data on multiple dimensions, and methods are selected by the researcher based on their effectiveness in gathering data about key aspects of the case. Data collection methods can include interviews, oral history, critical incidents, ethnographic observation, and document analysis. Qualitative case study designs do not preclude the use of quantitative methods such as surveys, which can be used to gather information in a more standardized manner to achieve a more precise measure of particular factors that are part of the case. In selecting the set of data collection methods, the researcher should take into account the alignment between research questions and the type of data needed to address those questions.

The research is typically presented as a report that contains thick narrative description, and in the final interpretive phase of analysis, the researcher derives conclusions and explains the underlying meaning behind the findings. This phase constitutes the lessons learned from the case. Meaning comes from learning about the issue or concern (i.e., an instrumental case study) or learning about a unique or unusual situation (i.e., an intrinsic case study). The researcher’s conclusions, recommendations, and personal reflection on conducting the study contribute to the reader’s overall understanding of the case analyzed.

Given the interpretive nature of qualitative inquiry, rather than merely identifying and isolating a case, the researcher can reconstruct it. As a result, the academic discussion has departed from arguing the ability of the case study method to establish generalizations, instead becoming redirected toward *phronesis* (practical, contextualized knowledge that is responsive to its environment) and transferability; that is, how (if at all) the understanding and knowledge gained can be applied to similar contexts, settings, and conditions. Toward this end, the researcher attempts to address the issue of transferability by way of rich description that will provide the basis for a qualitative account’s claim to relevance in some broader context.

Indeed, much of the discussion around case study research has concerned its value because its findings may be unable to be generalized beyond the case itself. In practical terms, this leads to the view that, rather than seeking guidance for practice from bodies of theory or generalized knowledge, the case study approach can offer ways of providing insights into social life based on *exemplary knowledge*—that is, by viewing and studying something in its completeness and richness and attempting to understand this. Through such exemplary knowledge, one can develop analytical insights and make connections with the experiences of others. The researcher therefore undertakes a case study to make the case understandable. This understanding may be what the reader learns from the case or its application to other similar cases, thereby constructing practical knowledge that is reflective of and responsive to its environment.

Linda Dale Bloomberg

See also Educational Research, History of; Generalizability; Qualitative Data Analysis; Qualitative Research Methods; Reliability; Representativeness; Sample Size

Further Readings

Creswell, J. W., & Poth, C. N. (2017). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). Thousand Oaks, CA: Sage.

Flyvbjerg, B. (2011). Case study. In N. K. Denzin & Y. S. Lincoln (Eds.), The SAGE handbook of qualitative research (4th ed., pp. 301–316). Thousand Oaks, CA: Sage.

Hamilton, L., & Corbett-Whittier, C. (2013). Using case study in education research. Thousand Oaks, CA: Sage.

Stake, R. E. (1995). The art of case study research. Thousand Oaks, CA: Sage.

Thomas, G., & Myers, K. (2015). The anatomy of the case study. Thousand Oaks, CA: Sage.

Yin, R. (2017). Case study research: Design and methods: Vol. 5. Applied social research methods (6th ed.). Thousand Oaks, CA: Sage.

Sara Tomek Sara Tomek Tomek, Sara

Categorical Data Analysis Categorical data analysis

239

243

Categorical Data Analysis

Categorical data analysis is a field of statistical analysis devoted to the analysis of dependent variables that are categorical in nature. Development of analytic techniques for inference utilizing categorical random variables began around 1900 when Karl Pearson introduced the chi-square statistic (χ^2). From this first introduction of tests of two-way contingency tables, the field has developed to include not only analyses of contingency tables but also more sophisticated analytic techniques such as the generalized linear mixed model. This entry defines categorical variables, outlines the most frequently utilized probability distributions for categorical variables, describes the most commonly used statistical analyses in the field of categorical data analysis, and discusses estimation methods for parameter estimates.

Categorical Variables

Categorical variables are a class of random variables whose outcomes fall into discrete categories as opposed to a continuous range of numbers. Discrete categorical variables can be categorized based on their level of measurement, either nominal or ordinal. Nominal categorical variables contain categories of responses that have an arbitrary ordering. That is, variables measured on this scale cannot be ranked or ordered based on their observed outcomes. The categories are simply placeholders for the outcomes. As an example, gender is measured on a nominal scale, as the following two outcomes, male and female, cannot be ordered in a meaningful way. Ordinal categorical variables, in contrast, contain categories of responses that have a natural ordering to them. The observed outcomes can be ranked or ordered based on this natural ordering, which provides meaning to the categories. As an example, age categories are measured using an ordinal scale, as the following two age categories, 20–29 and measured using an ordinal scale, as the following two age categories, 20–29 and 30–39, have a meaningful order to them. The second category, 30–39, represents subjects who are older than those in the first category, 20–29.

Probability Distributions

The use of inferential statistics requires an assumption of the distributional properties of the variables of interest. The distributional assumption of the categorical dependent variable provides the theoretical distribution of responses in the population, which is the basis for the statistical analysis being performed. For categorical data, the four most common distributions utilized in inferential statistics are the binomial distribution, the multinomial distribution, the hypergeometric distribution, and the Poisson distribution.

Binomial Distribution

The binomial distribution for random variable X calculates the probability of observing the count, Y, of the number of successes in a fixed number of trials of a Bernoulli experiment. A Bernoulli experiment is a random event in which there are two outcomes that have a fixed probability of occurring. In a binomial distribution, one of those outcomes is deemed a “success.” In a total of n trails, these successes are counted and the outcome X is the frequency of occurrence of a successful outcome. For example, if the Bernoulli experiment was flipping a coin, the outcome X could be the number of heads that occur in $n = 5$ trials (note that X ranges from 0 to n).

Multinomial Distribution

The multinomial distribution calculates the probability of observing the counts of each category when multiple outcomes are possible. The multinomial distribution is different from the binomial distribution in that there are three or more outcomes possible in each random experiment. Rather than utilizing the probability of a success, the probability of each outcome is calculated and creates a distribution of the counts of each outcome category. This is a multivariate distribution of all of the outcomes, where each individual outcome falls into a binomial distribution (that category vs. not in that category). The probabilities are calculated based on the occurrence of each category. For example, when asked to select a new drink, the options are “Drink A,” “Drink

B," and "Drink C." The multinomial distribution will look at the probability of the frequencies of the three outcomes simultaneously in a sample. One example would be the probability of observing the counts of 4, 3, and 3, respectively, in a sample of 10 subjects.

Hypergeometric Distribution

The hypergeometric distribution is similar to the binomial distribution in that it is looking at the count of events. The difference between the hypergeometric distribution and the binomial distribution is that the trials are not independent in the hypergeometric distribution as the subsequent trials occur without replacement. The resulting probability distribution of *X* will count the number of times a specific outcome occurs within a particular number of trials, where the number of outcomes available is fixed and the sampling of outcomes occurs without replacement. For example, in a box of 20 light bulbs, it is known that four of them do not work. One could calculate the probability that if one selects three light bulbs, *X* is the count of the number of working light bulbs. When sampling without replacement, the probabilities are not constant, as the current outcome affects the probability of any subsequent outcome. However, as a note, when the sample size is extremely large and the conditional probabilities do not change substantially, the hypergeometric distribution converges to the binomial distribution.

Poisson Distribution

The Poisson distribution for a random variable *X* calculates the probability of the number of times a particular event is observed. The Poisson is similar to the binomial distribution in that it is counting up the number of times an event is observed. However, the Poisson is different from the binomial in one important way: The exact probabilities are not known. In the Poisson distribution, probabilities are based on the observed frequencies, or the average outcomes observed within the data. The count of the number of outcomes within a specific unit of measurement (or number of trials) is calculated based on the average probability of occurrence. For example, in a Poisson distribution, one may count the number of phone calls, *X*, that is received in an hour if it is known that on average eight phone calls are received every 15 minutes (note that a trial can be thought of as a single minute).

Statistical Analyses

Contingency Tables

Analyses of frequencies of observed outcomes of categorical variables, or contingency tables, are the foundation of categorical data analysis. Analysis of a simple one-way contingency table, with one categorical variable, can test deviance from a theoretical frequency distribution. Analysis of a two-way contingency table, with two categorical variables, can test for dependence between two categorical variables (i.e., the extent to which the distribution of outcomes in a second variable deviate from those expected based on the distribution of outcomes in the first variable). Analysis of a three-way contingency table, with three categorical variables, allows for testing of interaction, or moderation effects, within the relationship of the variables.

The main statistical test for most contingency tables is that of the chi-square test. This test will look at the observed frequencies in the table as compared to the theoretical frequencies that are assumed under the null hypothesis. In a 2 × 2 contingency table, odds ratios and relative risk measures can be calculated and compared along with comparisons of proportions. A proportion measures the probability of a single outcome, the relative risk looks at the ratio of probabilities of two outcomes, while the odds ratio looks at the ratio of odds, or likelihood of occurrence, for two outcomes. What is consistent between these three is that they all can be used to evaluate probabilities of events in relation to a second outcome.

In very small samples, probabilities can be computed using Fisher’s exact test. For larger two-way tables, tests of independence are typically calculated based on the analysis of conditional probabilities. If variables are ordinal rather than nominal, linear trends can be analyzed using a Spearman correlation and tested using the Cochran–Mantel–Haenszel statistic.

Three-way contingency tables allow the analysis of independence between each pair of variables as well as conditional associations among all three variables (i.e., whether the level of dependence between the first two variables differs based on the outcome of the third variable). Contingency tables themselves are the basis of almost all of the subsequent analyses discussed in this entry.

Generalized Linear Models

Generalized Linear Models

Although the general linear model assumes a normal distribution for the response variable, the class of models called the generalized linear model allows the response variable to follow a distribution other than the normal distribution. Generalized linear models must have the following components: (a) The response variable must be a random variable, (2) the relationship between the independent variables and the response variable must be linear in form, and (3) the model must contain a link function that brings Components 1 and 2 together. This link function reflects functional form of the probability distribution underlying the response variable, *Y*. Common link functions for dichotomous responses are (a) the linear probability model, which models the probability of a dichotomous response as a linear function, (b) the logit link, which models the probability of a dichotomous response as an exponential function, and (c) the probit link, which models the probability of a dichotomous response in relation to the standard normal distribution. The logit link is the most common link function that is applied to variables with multinomial outcomes. When looking at count data, the log link is the most common link function, which models probabilities along a logistic distribution. The fit of different generalized linear models is typically evaluated using the Wald statistic or by comparing likelihood ratio statistics.

Logistic Regression

Logistic regression is a special case of a generalized linear model that utilizes the logit link to model the probabilities of dichotomous outcomes. The logit link is defined as the natural logarithm of the odds ratio. Individual parameter estimates must be interpreted after conversion back to the scale of the dependent variable (i.e., reversing the logit link). In logistic regression, though the response variable must be dichotomous, the independent variables can represent any level of measurement. That is, they can be both continuous and categorical. Estimated probabilities are compared in relation to the independent variables. Interpretation of these effects will depend upon the level of measurement of the independent variable. For categorical independent variables, probability estimates can be compared for each of the dependent variable categories based on each level of the independent variable. For continuous independent variables, probability estimates are compared in relation to a change in the level of the independent variable. In multiple logistic regression, model fitting is important to obtain the most parsimonious model most parsimonious model.

Multinomial Logistic Regression

Multinomial logistic regression is a special case of the generalized linear model as well. This model will analyze the probabilities of multiple response categories simultaneously. The multinomial logistic regression model also utilizes the logit link function. However, as the multinomial distribution is multivariate, the multinomial logistic regression model analyzes each adjacent category in the multinomial distribution simultaneously with individual logistic regression functions. That is, the log odds of each adjacent category can be analyzed simultaneously in a multinomial logistic regression. The probabilities of each category can then be evaluated in relation to the previous category. Interpretation of these model parameters is important with relation to the scale of measurement, as the previous category is arbitrary for nominal categories. The placement of nominal categories in the model will alter the parameter estimates. As this model is an extension of logistic regression, similar recommendations are given for the multiple multinomial logistic regression models with regard to the independent variables that can be utilized and model parsimony.

Log-Linear Models

Log-linear models are an extension of both generalized linear models and of the analysis described previously for contingency tables. Unlike the other generalized linear models, all variables in a log-linear model must be categorical in nature. The main distinction between log-linear models and analysis of a contingency table using a chi-square statistic is the distinction between an independent and dependent variable, as opposed to analyzing the association between two variables. Log-linear models aid in the interpretation of multiway contingency tables, as model estimates utilize the conditional distributions of the variables.

As in the assumption used for contingency tables, the main assumption under the null hypothesis in log-linear models is that of independence between the variables. Log-linear models utilize the logit function, which analyzes the log odds of moving between categories. If this assumption of independence does not hold and dependence is found, the parameters in the log-linear model allow analysts to more easily separate out the row effects from the column effects in the dependence. That is, how does the independent variable affect the probability of moving between two adjacent categories of the dependent variable. This is especially useful when three-way contingency tables are analyzed, as this model can distinguish between each marginal main effect, each pairwise conditional interaction effect, as well as analyzing a three-way interaction effect. Unlike analysis of variance models, the parameters in log-linear models do not reflect a hierarchy. In that sense, evaluating fit of the model to determine the most parsimonious and predictive model is important.

Generalized Linear Mixed Models

Generalized linear mixed models are yet another extension of generalized linear models. Generalized linear mixed models are appropriate when the categorical dependent is independently distributed (i.e., the responses are not clustered temporally, spatially, or in any other way, and errors are therefore uncorrelated). However, when this independence assumption is violated, standard errors are underestimated, the generalized linear mixed models must be used instead. Put differently, generalized linear mixed models accommodate one or more random effect, whereas generalized linear models only accommodate fixed effects. The difference between a random effect and fixed effect is the interpretation of the individual levels of the measure. When the individual levels of the variable are important, or they encompass all of the possible outcomes, then the independent variable is said to be a fixed effect. Gender is an example of a fixed effect. A random effect occurs when the individual values encompass a random selection of all of the possible outcomes of that variable. In education, teachers are typically assumed to be a random effect, as the teachers selected are a random sample of all teachers. That is, one do not need to measure differences between individual teachers (i.e., how much Mrs. A and Mrs. B differ), but the model will account for the differences that exist among all teachers as a collective.

The first application of these generalized linear mixed models was within item-response theory. For example, the Rasch model is a generalized linear mixed model using a logit link function with a dichotomous outcome of correct response on a test and a random independent variable which is the ability of the subject. Although differing link functions can also be used, the most common link is the logit link.

When a generalized linear mixed model contains two independent variables that are random effects, and these two variables are said to be nested within one other (e.g., students and teachers; students within a teacher’s classroom), these can be (e.g., students and teachers, students within a teacher's classroom), these can be analyzed within a multilevel modeling framework.

Other Models

As this is a brief description of categorical data analysis, not every analysis can be described in this entry. There are many additional methods that are included within the umbrella of categorical data analysis. A few of these methods include probit model, complementary log-log model, conditional logistic regression, inter-rater reliability analysis, latent class analysis, and cluster analysis.

Estimation of Models

The most common method for estimating the parameter estimates of models involving categorical data is through the use of maximum likelihood estimation. However, this method may not be the best when subject responses are not independent. In this instance, the generalized estimating equations control for nonindependence in its estimation and may therefore produce more robust parameter estimates. Since the 1960s, the use of Bayesian techniques for producing parameter estimates for categorical data analytic models has been well researched. Bayesian methods provide additional distributional information that is used to aid in the estimation of the final parameter estimates for these models. Most of the aforementioned methods have Bayesian analogs.

Many computer software programs that run statistical analysis have estimation procedures for categorical data analysis built into their systems. SAS, SPSS, R, S-PLUS, STATA, and SYSTAT all have the function to compute most, if not all, of the statistical analyses mentioned in this entry.

Sara Tomek

See also Bayesian Statistics; Bernoulli Distribution; Binomial Test; Chi-Square Test; Cluster Analysis; Inter-Rater Reliability; Item Response Theory; Latent Class Analysis; Levels of Measurement; Logistic Regression; Mann-Whitney Test; Mantel-Haenszel Test; Maximum Likelihood Estimation; McNemar Change Test; Nominal-Level Measurement; Odds Ratio; Ordinal-Level Measurement; Poisson Distribution; Probit Transformation; Rankings; Rasch Model; Spearman Correlation Coefficient; Two-Way Chi-Square

Further Readings

Agresti, A. (2012). Categorical data analysis (3rd ed.). Hoboken, NJ: Wiley.

Agresti, A. (2015). Foundations of linear and generalized linear models. Hoboken, NJ: Wiley.

Agresti, A., & Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. Statistical Methods and Applications, 14, 297–330.

Allison, P. D. (2012). Logistic regression using SAS: Theory and application (2nd ed.). Cary, NC: SAS Institute.

Azen, R., & Walker, C. M. (2010). Categorical data analysis for the behavioral and social sciences. New York, NY: Routledge.

Friendly, M., & Meyer, D. (2015). Discrete data analysis with R: Visualization and modeling techniques for categorical and count data. Boca Raton, FL: CRC Press.

Hosmer, D. W.Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Hoboken, NJ: Wiley.

Nussbaum, E. M. (2015). Categorical and nonparametric data analysis: Choosing the best statistical technique. New York, NY: Routledge.

Tang, W., He, H., & Tu, X. M. (2012). Applied categorical and count data analysis. Boca Raton, FL: CRC Press.

Tutz, G. (2011). Regression for categorical data (1st ed.). Cambridge, UK: Cambridge University Press.

van der Ark, L. A., Croon, M. A., & Sijtsma, K. (Eds.). (2005). New developments in categorical data analysis for the social and behavioral sciences. Mahwah, NJ: Erlbaum.

A. Alexander Beaujean A. Alexander Beaujean Beaujean, A. Alexander

Cattell–Horn–Carroll Theory of Intelligence Cattell–horn–carroll theory of intelligence

243

247

Cattell–Horn–Carroll Theory of Intelligence

The Cattell–Horn–Carroll (CHC) theory of intelligence is a psychometric taxonomy designed to explain how and why individuals differ in cognitive ability. It provides a common frame of reference and nomenclature to organize cognitive ability research. Its name comes from integrating Raymond Cattell and John Horn's subsequent occurrence theory with John Carroll's three-stratum theory, both of which are largely driven by factor analysis of psychometric measures of cognitive ability. This entry first looks at the historical antecedents of CHC theory, the development and purpose of the theory, and criticisms of the theory.

History

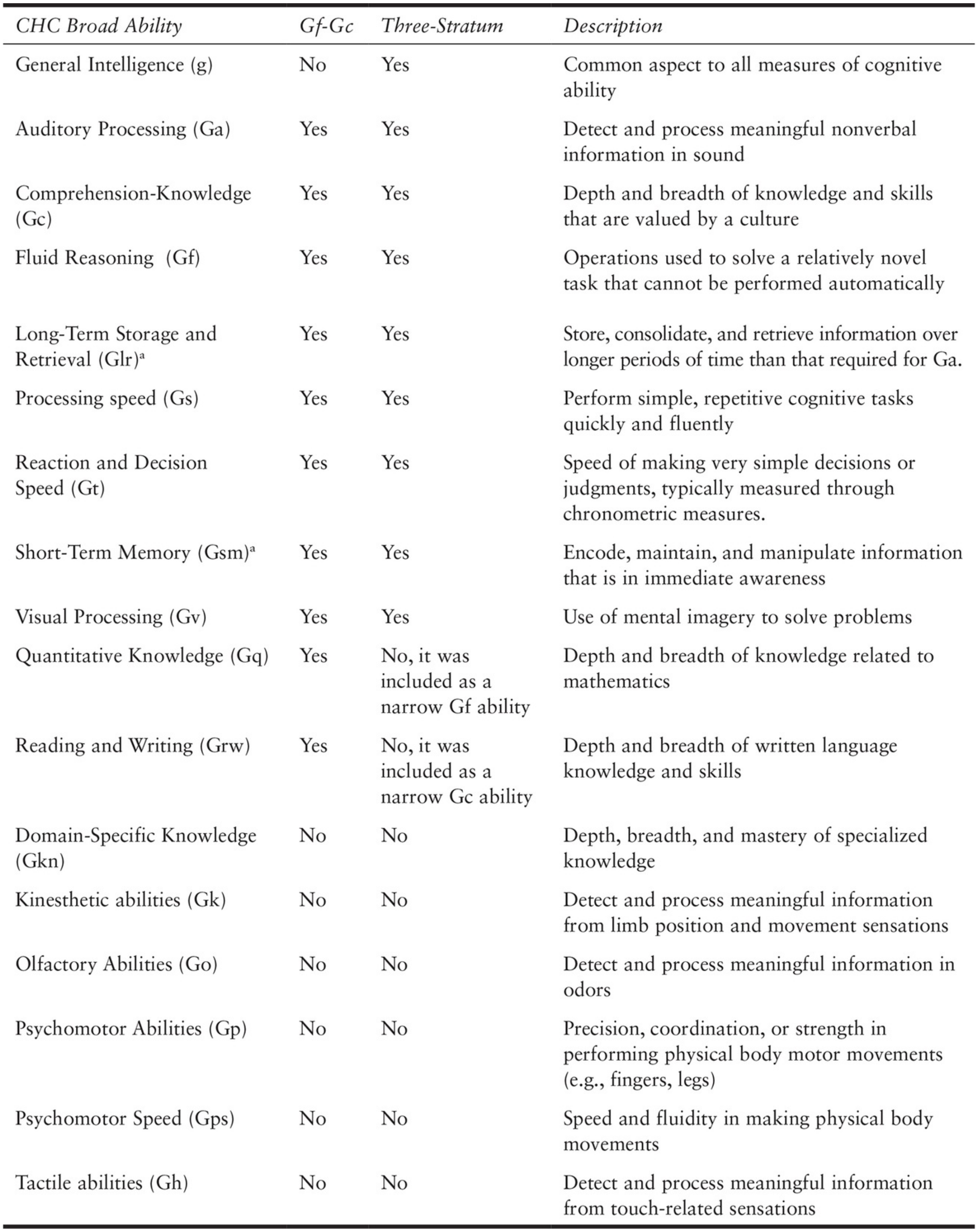
Although CHC theory can trace its heritage to the work of Francis Galton, Charles Spearman, Cyril Burt, Philip Vernon, and L. L. Thurstone, it largely begins with Raymond Cattell. Cattell studied with Spearman in the 1920s. At this time, Spearman was revising his notions about general intelligence (*g*), as it did not appear to be strongly related to scholastic tests or information retention. Cattell later expanded on this idea by noting that there were two abilities common to all measures of cognitive functioning: fluid intelligence (Gf) and crystallized intelligence (Gc) abilities.

Cattell defined Gf similarly to how Spearman defined *g*: a general ability to perceive the relations between the fundamental aspects of any problem. Likewise, he defined Gc as representing Spearman's ideas about tasks that were not good measures of *g*. He thought these tasks required habits established in a particular content area that originally required Gf but no longer need this type of reasoning for the successful completion of problems. In other words, Cattell believed that Gf "invests" in Gc; thus, both Gf and the environment in which a person operates determine the development of Gc.

At the same time, Cattell was developing his Gf-Gc theory, other scholars were finding different common abilities among groups of cognitive ability tests. Scholars came to realize that the number of common abilities that factor analysis could find was arbitrarily large, depending only on the number and similarity of the analyzed tests. Thus, factor analysts developed ways to factor analyze the common factors (i.e., hierarchical models).

The two most common hierarchical approaches were higher order and bifactor. Higher order models assume there are a small number of broad abilities that work through more primary abilities to influence differences in cognitive performance. They are developed by factor analyzing the correlations among primary factors. Bifactor models assume that all abilities directly influence cognitive performance. They are developed by factor analyzing the residual correlations among cognitive performance tasks after extracting a general factor (i.e., the aspect that is in common with all analyzed tasks).

Using higher order models, Cattell and one of his students, John Horn, began significantly expanding Gf-Gc theory. They posited that there was a plethora of primary abilities common to any given set of cognitive tasks, but these primary abilities only influence a small aspect of cognitive functioning. The relations among these primary abilities could be further factor analyzed to find a smaller number of broad abilities. Although these broad abilities included both Gf and Gc, they included other abilities as well; eventually, the number of broad abilities came to a total of 10. All 10 broad abilities and their descriptions are given in Table 1.



CHC Broad Ability	*Gf-Gc*	*Three-Stratum*	*Description*
General Intelligence (g)	No	Yes	Common aspect to all measures of cognitive ability
Auditory Processing (Ga)	Yes	Yes	Detect and process meaningful nonverbal information in sound
Comprehension-Knowledge (Gc)	Yes	Yes	Depth and breadth of knowledge and skills that are valued by a culture
Fluid Reasoning (Gf)	Yes	Yes	Operations used to solve a relatively novel task that cannot be performed automatically
Long-Term Storage and Retrieval (Glr)[a]	Yes	Yes	Store, consolidate, and retrieve information over longer periods of time than that required for Ga.
Processing speed (Gs)	Yes	Yes	Perform simple, repetitive cognitive tasks quickly and fluently
Reaction and Decision Speed (Gt)	Yes	Yes	Speed of making very simple decisions or judgments, typically measured through chronometric measures.
Short-Term Memory (Gsm)[a]	Yes	Yes	Encode, maintain, and manipulate information that is in immediate awareness
Visual Processing (Gv)	Yes	Yes	Use of mental imagery to solve problems
Quantitative Knowledge (Gq)	Yes	No, it was included as a narrow Gf ability	Depth and breadth of knowledge related to mathematics
Reading and Writing (Grw)	Yes	No, it was included as a narrow Gc ability	Depth and breadth of written language knowledge and skills
Domain-Specific Knowledge (Gkn)	No	No	Depth, breadth, and mastery of specialized knowledge
Kinesthetic abilities (Gk)	No	No	Detect and process meaningful information from limb position and movement sensations
Olfactory Abilities (Go)	No	No	Detect and process meaningful information in odors
Psychomotor Abilities (Gp)	No	No	Precision, coordination, or strength in performing physical body motor movements (e.g., fingers, legs)
Psychomotor Speed (Gps)	No	No	Speed and fluidity in making physical body movements
Tactile abilities (Gh)	No	No	Detect and process meaningful information from touch-related sensations

Note: Parenthetical terms are CHC abbreviations for the ability.

[a] There are noticeable differences in how this ability is defined across the three theories.

Although there was a corpus of scholarship supporting Gf-Gc theory, there were many competing theories for how human cognitive ability was structured. To determine what theory had the most empirical support, in the early 1980s, John Carroll began reanalyzing all previously published cognitive ability data sets he could find. Eventually, he found over 460 data sets, all of which he submitted to a common method of exploratory factor extraction and rotation.

From his results, Carroll created a systematic framework for classifying human cognitive ability comprising three different strata. The strata represented his method for differentiating the abilities based on their abstractness. At the least abstract level (Stratum I) are many primary abilities (what he called narrow abilities). At Stratum II are eight broad abilities that he believed represented the basic cognitive characteristics of individuals. They are more abstract than those at Stratum I, and many are similar to the broad factors from Gf-Gc theory (see Table 1). At Stratum III is the most abstract factor: Spearman's *g*.

Although they have many similarities, there are some fundamental differences between the Gf-Gc and three-stratum theories. First, Gf-Gc theory does not include *g*, while *g* is central to the three-stratum theory. Second, each theory specifies a different number of broad and narrow abilities (see Table 1). Third, Gf-Gc theory posits that broad abilities are built upon narrow abilities; thus, broad abilities work through the narrow abilities and are not directly related to performance on cognitive tasks. The three-stratum theory, however, posits that all aspects of cognitive ability are independently operating within an individual and are directly related to any differences on cognitive tasks.

Development and Purpose of CHC Theory

Both Gf-Gc and three-stratum theory remained largely of theoretical interest until Richard Woodcock and Kevin McGrew began work on the revised edition of the Woodcock-Johnson Psycho-Educational Battery (WJ-R). They developed the instrument to map onto Gf-Gc theory by purposefully measuring the broad abilities. Moreover, as part of the WJ-R development of process, they invited Horn and Carroll to a series of meetings to discuss the structure of cognitive ability.

When Woodcock and McGrew started work on the third edition of the Woodcock-Johnson (WJ-III), they again invited Horn and Carroll to consult on the instrument. Between the development of the WJ-R and WJ-III, McGrew started integrating the Gf-Gc and three-stratum theories in order to have a single way to classify various measures of human cognitive ability. It was the publication of the WJ-III, however, that provided the first definition of CHC theory as an amalgamation of the Gf-Gc theory and the three-stratum theory.

Like the three-stratum theory, CHC comprises three strata, each of which represents abilities at different level of abstraction. Like the Gf-Gc theory, initially there were 10 broad abilities at Stratum II, although the number subsequently expanded to 16 abilities (see Table 1). Also like the Gf-Gc theory, most CHC-based factor analysis uses higher order models.

Unlike the Gf-Gc and three-stratum theories, the primary purpose of developing the CHC theory was for clinical purposes: to have a taxonomy to classify individual tests from different cognitive ability instruments as well as develop new instruments. Initially, the process of test classification was done through conducting confirmatory factor analysis of multiple large-scale cognitive ability instruments. This focused largely on classification of tests at the Stratum II level.

Classifications at Stratum I were usually made by finding the consensus classification from a select few scholars about what the tests measure. The major finding from the test classification studies was that no single instrument measured all the Stratum II abilities that CHC scholars thought were important to understand an individual’s cognitive functioning. Moreover, the test an instrument used to measure a given Stratum II ability was not equivalent as some tests did a better job of measuring the constructs than others. This eventually gave birth to the cross-battery approach to cognitive assessment, which is a way to combine test scores from independent instruments for the purposes of a clinical evaluation.

Initially, the WJ-III was the only cognitive instrument whose development was based on CHC. As the theory gained in popularity, however, it began to be used by more test developers. By 2015, most new and revised popular intelligence instruments either are grounded explicitly in CHC theory or pay some form of implied allegiance to it.

Criticisms

Despite the popularity of the CHC theory, it has also been criticized. One major criticism is its lack of focus on *g*. Although most CHC factor models include *g* as a higher order factor, CHC applications typically eschew *g* in favor of the Stratum II abilities. Critics have noted that the prioritization of Stratum II abilities is typically inappropriate as *g* (or its manifestation in a global composite score) explains more variance in test scores and has better psychometric properties; moreover, Stratum II abilities seldom add any additional information in predicting external criteria beyond that provided by *g*.

A related CHC criticism is the number of Stratum II factors. As of 2015, CHC theory had 16 Stratum II abilities, compared to 10 in its original formation and only eight in the three-stratum theory. As is the case with narrow abilities, Stratum II factors can increase almost indefinitely by adding more measures of Stratum I factors in a given analysis. Critics argue that just because these factors can be extracted does not mean that they are clinically useful. Designing instruments to measure many Stratum II factors makes them longer and requires increased administration, scoring, and report writing time. This additional cost is not counterbalanced by the information gained from measuring the additional abilities, however, as there is little evidence that knowledge of levels of Stratum II abilities increases accuracy of diagnosis or intervention planning.

A. Alexander Beaujean

See also *g* Theory of Intelligence; Intelligence Quotient; Intelligence Tests; Multiple Intelligences, Theory of

Further Readings

Beaujean, A. A. (2015). John Carroll’s views on intelligence: Bifactor vs. higher-order models. Journal of Intelligence, 3, 121–136. doi:10.3390/jintelligence3040121

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York, NY: Cambridge University Press.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? Intelligence, 35, 169–182. doi:10.1016/j.intell.2006.07.002

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross battery ability assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment: Vol. 1. Intelligence and achievement (2nd ed., pp. 343–373). New York, NY: Guilford Press.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. Psychology in the Schools, 47, 635–650. doi:10.1002/pits.20496

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), Contemporary intellectual assessment (pp. 151–179). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities: Past, present and future. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests and issues (2nd ed., pp. 136–181). New York, NY: Guilford Press.

McGrew, K. S., & Woodcock, R. W. (2001). Woodcock-Johnson III technical manual. Itasca, IL: Riverside Publishing.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment (3rd ed., pp. 99–144). New York, NY: Guilford Press.

Donald B. Rubin Donald B. Rubin Rubin, Donald B.

Elizabeth R. Zell Elizabeth R. Zell Zell, Elizabeth R.

Causal Inference

Causal inference

247

251

Causal Inference

Causal inference refers to the process of drawing a conclusion that a specific treatment (i.e., intervention) was the “cause” of the effect (or outcome) that was observed. A simple example is concluding that taking an aspirin caused your headache to go away. Inference for causal effects in education might include, for instance, aiming to select programs that improve educational outcomes or identifying events in childhood that explain developments in later life. This entry’s examination of causal inference begins by first exploring the principles of randomized experiments, which are the bedrock for drawing causal inferences. The entry then reviews the design of causal studies, three distinct conceptual modes of causal inference, and complications that can arise that may prevent causal inference.

Basic Principles of Randomized Experiments

Randomized experiments are the gold standard for drawing causal inferences, but drawing such inferences from observational studies is often necessary and requires special care. Here, we use the Rubin causal model (RCM) framework, which begins by defining causal effects using potential outcomes, a formulation originally due to Jerzy Neyman in the context of randomization-based inference in randomized experiments. We use well-accepted statistical principles of design and analysis in experiments to connect to the design and analysis of observational studies.

Randomized controlled trials (RCTs) are commonly used to compare treatments (i.e., interventions). The simplest setting has two groups, with each unit (e.g., person, classroom) having a known probability of assignment into the active treatment or the control treatment, and the units are followed for a predefined period of time to observe outcome variables, generically denoted here by Y. An example would be a posttest score 1 year after randomization.

RCTs ideally have strictly developed protocols specified in advance of implementation. A critical feature of RCTs is that the active versus control treatment is randomly chosen for each unit; thus, in expectation, the treated group and the control group are balanced on measured and unmeasured covariates, where *balance* here means having the same expected distributions of all covariates. Covariates are variables, like age and baseline pretest scores, thought to be correlated with $\boldsymbol{Y}$, but that differ from $\boldsymbol{Y}$ because their values are known to be the same for each unit whether the unit was assigned to the treatment or control group; examples include male–female, age, and educational history of parents. Observed covariates are denoted by $\boldsymbol{X}$.

Assignment of Units in Randomized Experiments

An RCT is a special type of assignment mechanism. Let $W_i = 1$ if the ith unit ($i = 1, \ldots, N$) is assigned to receive the active treatment, and let $W_i = 0$ if the ith unit is assigned to receive the control treatment. In an RCT, the probability that the ith unit assigned active treatment is between 0 and 1; notationally,

$$0 < P\left(W_i = 1 \mid X_i\right) < 1,$$

where the vertical line indicates conditioning, and $\boldsymbol{X}_i$ indicates the values of all observed covariates for unit i; implicitly, the probability in expression (1) does not depend on any values of unobserved covariates or on any values of $\boldsymbol{Y}$ but can depend on $\boldsymbol{X}_i$; this kind of assignment mechanism is called unconfounded.

Although it is common in RCTs with two treatment groups for each unit to have a 50% chance to be assigned to the active or control treatment, this is not required. For example, with an active treatment that is considered likely to be beneficial for older students, to encourage units to enroll in the RCT, investigators might choose to randomly place two thirds of older units into the active treatment group and one third in the control group, whereas younger students would be equally assigned to either treatment.

students would be equally assigned to either treatment.

The causal effect of the active treatment relative to the control treatment for unit i is the comparison of the outcome that would be observed when unit i is assigned active treatment, referred to as $Y_i(1)$, to the outcome that would be observed when unit i is assigned the control treatment, referred to as $Y_i(0)$, with both measured the same length of time after the assignment. In any real-world setting, a unit can only be exposed to either the active treatment or the control treatment. Because we cannot go back in time to give the other treatment, we can only observe $Y_i(1)$ or $Y_i(0)$ for unit i, thus the primary problem facing causal inference is the problem of missing data. Consequently, although these causal effects are defined at the level of the individual unit, they cannot be directly measured.

The collections of observable values of $\boldsymbol{X}$ and $\boldsymbol{Y}(1)$, $\boldsymbol{Y}(0)$ under all possible assignments are called “the science.” For the N units in the study, the science includes (a) the covariates $\boldsymbol{X}$, a matrix with N rows, the ith row being, $\boldsymbol{X}_\mathrm{i}$; (b) the potential outcomes under treatment $\boldsymbol{Y}(1)$, which is a matrix with N rows, the ith row being $\boldsymbol{Y}_i(1)$, which gives values for the outcome variables when unit i is exposed to the active treatment; and (c) $\boldsymbol{Y}(0)$, which is a matrix of the potential outcomes under the control treatment with N rows, the ith row being $\boldsymbol{Y}_i(0)$, which gives the values of the outcome variables for unit i under the control treatment.

The science, the array $(\boldsymbol{X}, \boldsymbol{Y}(1), \boldsymbol{Y}(0))$, represents all observable values of $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the stable unit-treatment value assumption, which asserts that each potential outcome is a function only of the unit label, i, and the assigned treatment W_i. More precisely, for unit i, stable unit-treatment value assumption disallows (a) “hidden” treatments not represented by $W_i = 0$ or $W_i = 1$ as well as (b) interference between units; that is, the potential outcomes $(Y_i(1), Y_i(0))$ for unit i are not affected by the treatments assigned to any other units.

Formal Definition of the Assignment Mechanism

The assignment mechanism gives the probability of the N-component vector of treatment assignments $\boldsymbol{W} = (W_1, W_2, \ldots, W_i, \ldots, W_N)^T$—(the superscript T denotes transpose, so that $\boldsymbol{W}$ is a column vector)—given the science, notationally,

$$P\left(W \mid X, \boldsymbol{Y}(1), Y(0)\right).$$

This notation reveals the possible dependence of the assignment mechanism is not only on the covariates but also on the potential outcomes.

The possible dependence on the potential outcomes in Equation 2 is the bane of observational studies because, for example, teachers may assign the active treatment to students they think are more needy, based on unmeasured assessments, a feature that could violate the unconfounded assumption of Equation 1. RCTs are also probabilistic in the sense that every unit has a positive probability of being assigned either treatment. RCTs possess other advantageous features.

Sometimes the assignment mechanism can be written as proportional to the product of N propensity scores, $e(\boldsymbol{X}_i) = P(W_i = 1|\boldsymbol{X}_i)$, where $e(\boldsymbol{X}_i)$ is the probability that unit i with covariate value $\boldsymbol{X}_i$ is assigned to be actively treated. In an RCT, the N propensity scores are known, whereas in an observational study, they must be estimated—a critical distinction affecting both design and analysis of observational studies.

Causal Estimands

Even though there is no way to calculate unit-level causal effects from observed data because at least one of the potential outcomes is missing, typical causal effects can be estimated. For example, a common estimand compares the average potential outcome under the active treatment with the average potential outcome under the control treatment, , where is the average value across all N units of the $Y_i(1)$, and analogously for $Y_i(0)$. Or the estimand could be the median individual causal effect, $\text{med}_i[Y_i(1) - Y_i(0)]$. Generally, causal estimands are a comparison of $Y_i(1)$ values and $Y_i(0)$ values on a common set of units.

RCTs can provide reliable answers to causal questions because we know the rule used to select the treated and control units, and each unit has a known chance to be in either group. More precisely, consider an RCT with each unit having an equal probability of being in the treatment or control group. The observed $Y_i(1)$ values are simply a random sample from all $Y_i(1)$ and so fairly represent all

$Y_i(1)$; analogously, for the observed $Y_i(0)$ values, fairly representing all $Y_i(0)$. With nonrandomized studies, it is usually difficult to use the observed values of $Y_i(1)$ to estimate fairly the missing values of $Y_i(1)$, and analogously for the values $Y_i(0)$ because of possible baseline differences between observed and missing $Y_i(1)$ values, and between observed and missing $Y_i(0)$ values, as in the example of teachers who assign students they perceive as more needy at baseline (in unmeasured ways) to the active treatment.

Design of Causal Studies

The first task in the design of any causal study is to try to create, using only values of $\boldsymbol{X}$, active and control groups that have nearly the same distributions of $\boldsymbol{X}$. This task is easier when one can use randomization to assign treatments, and there is a vast literature on the design of RCTs. The literature on the proper design of observational studies is far more recent and often utilizes estimated propensity scores and associated diagnostics for assessing the balance in active and control $\boldsymbol{X}$ distributions.

Modes of Causal Inference from Data

There are three distinct conceptual modes of causal inference in RCTs: one due to Ronald Fisher, one due to Jerzy Neyman, and one due to Donald Rubin. These are extended to nonrandomized studies in the RCM.

The Fisherian approach is closely related to the mathematical idea of proof by contradiction and begins with a sharp null hypothesis, which is that the treatments have absolutely no effect on the potential outcomes. This null hypothesis is called “sharp” because under it, all potential outcomes are known from the actual observed values of the potential outcomes; for each unit, either $Y_\mathrm{i}(1)$ or $Y_\mathrm{i}(0)$ is observed, and by assumption they are equal. Under the null hypothesis, it follows that the value of any statistic such as the difference in the observed averages for units exposed to Treatment 1 and units exposed to Treatment 0, , is known, not only for the observed assignment but also for all possible assignments $\boldsymbol{W}$. From this fact, we can calculate the significance level (or *p* value) of the observed . Neymanism randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over their distributions induced by the assignment mechanism to calculate a confidence interval (e.g., 95%) for the typical causal effect. This mode is currently dominant in educational investigations of causal effects.

The third mode of inference (Bayesian) for causal effects requires a probability model for the science, $P(\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1))$. A virtue of the RCM framework used here is that it separates the science and a model for it, from what we do to learn about the science—the assignment mechanism. This approach directly and explicitly confronts the missing potential outcomes by multiply imputing them. That is, the RCM perspective takes the specification for the assignment mechanism and the specification for the science and derives the conditional distribution, called the posterior predictive distribution (i.e., posterior because it conditions on all the observed data and predictive because it is based on predicting the missing potential outcomes), of the missing potential outcomes given all observed values (i.e., $\boldsymbol{X}$, $\boldsymbol{W}$, and the observed potential outcomes). This approach relies on current computational environments that rely on simulation, here the simulation of the missing potential outcomes.

Complications

Many complications may, and often do, occur in real-world studies for causal effects, many of which can be handled much more flexibly with the Bayesian approach than with randomization-based methods. Of course, the models for the science can be difficult to formulate in a practically reliable manner. In addition, Neymanian-style evaluations are still important. Fisherian p values are a special case of Bayesian posterior predictive p values. Thus, the wise investigator should understand all three modes.

Most of the field of classical experimental design is devoted to issues that arise with more than two treatment conditions and covariates that can define blocking structures. The common modes here are randomization based.

Missing data, due perhaps to unit dropout, can complicate analyses more than one would expect. Methods such as multiple imputation, the expectation–maximization algorithm, data augmentation, and the Gibbs sampler are more compatible with the Bayesian approach to causal inference than the other modes.

Another common complication is noncompliance with assigned treatment, which is often unavoidable in education investigations. Further complications include partially defined outcomes, such as final exam scores that are only well defined for students who are still in school at the time of measurement. In the real world, complications typically do not appear simply one at a time. For example, an RCT can suffer from missing data in both covariates and longitudinal outcomes and also from noncompliance and partially defined outcomes. Many of the aforementioned complications can be viewed as special cases of principal stratification.

Donald B. Rubin and Elizabeth R. Zell

See also Bayesian Statistics; Compliance; Experimental Designs; Markov Chain Monte Carlo Methods; Missing Data Analysis; Outcomes; Propensity Scores; Random Assignment

Further Readings

Fairlie, T., Zell, E. R., & Schrag, S. (2013). Effectiveness of intrapartum antibiotic prophylaxis from prevention of early-onset group B streptococcal disease. Obstetrics … Gynecology, 121, 570–577.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. Biometrics, 58, 21–29.

Imbens, G., & Rubin, D. B. (2015). Causal inference in statistics, and in the social and biomedical sciences. New York, NY: Cambridge University Press.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). New York, NY: Wiley.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 6, 34–58.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine, 127, 757–763.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys. New York, NY: Wiley.

Rubin, D. B., & Zell, E. R. (2016). Causality in experiments and observational studies. In S. J. Henly (Ed.), Routledge international handbook of advanced quantitative methods in nursing research (Chapter 15). Oxford, UK: Routledge.

Gavin W. Fulmer Gavin W. Fulmer Fulmer, Gavin W.

Causal-Comparative Research Causal-comparative research

251

254

Causal-Comparative Research

Causal-comparative research is a family of research designs used to examine potential causes for observed differences found among existing groups. Causal-comparative research is useful for the study of causes where experimental assignment or manipulation is infeasible, unethical, or in some way prohibited. It is frequently used with large-scale survey data such as Programme for International Student Assessment or National Assessment of Educational Progress but also common in smaller scale studies. It is similar to correlational research designs, except that the independent variable to be tested is categorical (e.g., school or class membership) and the analysis explicitly attempts to test causality. Although some scholars debate the conceptual distinction between causal-comparative and correlational designs in education research and recommend merging correlational and causal-comparative under the heading “nonexperimental quantitative research,” the distinction is still present in many methods textbooks. This entry presents the basic principles of causal-comparative research and steps to conduct a causal-comparative study.

Basic Principles

Causal-comparative research begins with a known or expected outcome—a dependent variable that is the *effect*—and a group distinction to be compared as a possible *cause* for the effect. The researcher compares two or more *intact groups* to test the cause. When data on both the effect and the potential causes are already known by the researcher—hence the situation under study has already completely transpired—the study is retrospective. Retrospective causal-comparative studies are therefore *ex post facto*, or “after the fact” studies, because all data about group differences and about potential causes are obtained after both cause and effect have occurred. When the potential causes are studied contemporaneously before effects are observed, the study is a prospective one. Such cases are sometimes called *natural experiments*. However, regardless of whether retrospective or prospective, causal-comparative studies do not have experimental manipulation by a researcher and so technically cannot be classified as experimental research.

Alternative to Experimental Research

Causal-comparative research is an alternative to experimental and quasi-experimental designs, but the distinction with experimental research is an important one. In general, experimental designs involve some manipulation by the researcher of a causal intervention or treatment of some kind. For the so-called true experiments, the distinction with causal-comparative research is readily apparent: A true experiment has random assignment of participants to an experimental condition. The distinction with quasi-experimental research may be less obvious in certain cases. In some quasi-experimental studies, the researcher may work with intact groups just as in causal-comparative research, but a quasi-experiment would have a manipulation of some kind. By contrast, in causal-comparative research, the researcher does not control the study conditions.

Consider the following example: A researcher studying the potential causes for observed differences in elementary classes' average mathematics achievement chooses to focus on use of newer textbooks in some of the classes. In this example, the effect is the difference in mathematics achievement, and the potential cause is the use of the new textbook. The two groups to be compared are classes using the newer textbooks and classes using the older textbooks. In a true experiment, the researcher would form groups by randomly assigning students into classes using the new textbooks or an older textbook. In a quasi-experiment, the researcher would not be able to assign the students into classes randomly but may still be able to assign some classes to use the new textbook as a comparison with classes using the older textbook. In a causal-comparative study, the researcher does not assign students to classes and does not influence which textbooks are used. Instead, the researcher finds classes already using the new textbook and compares them with classes using the older textbook.

The lack of experimental manipulation makes causal-comparative research similar to correlational studies; it is often presented in methods textbooks in the same chapter with correlational designs. One important difference is that a causal-comparative study must involve two or more groups being compared with the intention of uncovering cause, whereas correlational designs typically focus on descriptive or trend analysis (whether within a single group or across groups) and need not assert that group differences are caused by the group membership variable in any way.

Causal-comparative research is particularly useful in situations where a researcher cannot influence either the group membership or the experiences of the groups. That is, situations where experimental manipulation is impossible or in contexts where group membership and the potential causes are limited due to feasibility, ethics, or legal reasons. For example, in a study of potential sex differences in reading ability, it is technically infeasible to “assign” students to be boys or girls. Likewise, any study of biological or social variables that cannot be manipulated must be causal–comparative—such as the effects of age, race, and ethnicity.

In other situations, it may be impractical or unethical to control assignment or to manipulate the environment. Consider a study of education interventions among incarcerated youth and the effect on the youths’ recidivism, comparing different subgroups in the detention center or across different detention centers. Research in this context is restricted in group membership to incarcerated youth by definition. Practically, any one youth detention center may have one or few classes for a given age range, prohibiting multiple experimental groups. Policy may also prevent detention centers from implementing various interventions without some evidence of potential benefit of the new intervention. Furthermore, ethically speaking, the youths may be in a vulnerable situation that would affect their ability to freely opt into or out of the study’s experimental conditions. Causal-comparative research could be the only ethical and practical approach to study in such circumstances.

Limitations

There are two serious limitations in causal-comparative designs that researchers must recognize: the fallacy of homogeneity and the post hoc fallacy. The *fallacy of homogeneity* is an error of assuming groups are internally homogeneous. It arises when researchers assume that a demographic group (e.g., women, persons of color) is sufficiently internally similar to allow meaningful comparison with other groups, when in fact all groups are internally varied on some other variable that also influences the effect (e.g., socioeconomic status [SES]). The *post hoc fallacy* is an error in attributing causation where no cause can be established. It arises when researchers presume that an observed correlational relationship implies a causal relationship. Both fallacies cannot be eliminated—they can only be controlled through careful and thorough consideration of alternative explanations for an observed effect as discussed in the next section.

Conducting a Causal-Comparative Study

To conduct a causal-comparative study, a researcher must identify an effect and potential causes within a context and among groups. This order is not obligatory; a researcher is very likely to encounter a context and intact groups and then begin to consider potential causes for an observed effect. Then, the researcher must identify and attempt to eliminate alternative explanations for the findings. Finally, the researcher analyzes group differences to test the proposed causal relationship and alternative explanations.

Effects and Causes

The first step in causal–comparative is to identify the effect and propose a cause. For new studies, this may arise from practical experience with the context. For secondary analyses of large-scale data, this involves reading through the documentation on the survey measures. A researcher must provide a strong argument for the mechanism by which the proposed cause is expected to yield the effect, typically with a combination of the following: theoretical analysis of the context, the effect, and various causes; corroboration from other empirical research on causal relationships; and a clear and logical rationale. Without a strong logical, theoretical, and empirical argument, there would be little to demonstrate that the observed relationship is causal.

Identifying or Forming Groups

The next step is to identify or form groups for comparison of the potential cause. Group identification could follow from existing group information, such as in the example on textbook use—one group of students in classes using a newer textbook and one group of students in classes using an older textbook. But, it is also common that researchers identify groups using other data. Such groups could be formed based on organismal data (e.g., sex and age), other demographic could be formed based on organismal data (e.g., sex and age), other demographic data (e.g., gender, race, ethnicity, or religion), or constructed from other responses (e.g., based on a calculated SES index or performance on previous tests).

Group formation can be sensitive. Race, ethnicity, and gender identification can be complex and have multiple, competing interpretations in varied circumstances. In addition, comparing groups by race, ethnicity, or SES may be contentious in some research contexts or scholarly fields. Furthermore, a focus on one identified grouping variable (e.g., ethnicity) may mask other group differences (e.g., by SES) that could also be pertinent, giving rise to the fallacy of homogeneity. There may also be multivariate combinations, or intersections, of some grouping variables that warrant attention. For example, research on gender difference in undergraduate science majors may miss further differences for the subgroup of women from underrepresented minorities. Given the complexity and potential sensitivity, causal-comparative researchers need to exercise caution in the identification of groups for comparison. This further highlights the importance of a logical, theoretical, and empirical argument for the mechanism connecting cause and effect.

Identifying Alternative Explanations

The researcher next identifies alternative explanations for observed effects. The purpose is to recognize other possible explanations, so that they can be examined and potentially eliminated. Identifying alternative explanations is quite similar to the need to identify the causal mechanism—there is a combination of theory, empirical review, and logical argument. An effective approach is to conjecture what a reasonable and informed reader might suggest as a different cause for the effect or a different mechanism between proposed cause and observed effect. Failure to identify and account for alternative explanations may lead to *spurious causation*: when the proposed cause and observed effect actually result from a different cause that was not considered, a case of the post hoc fallacy.

Reconsider the earlier textbook adoption example and suppose that students in classes using the newer textbooks outperformed students using the old textbooks. Is this caused by the textbook? One alternative cause is that schools that have purchased newer textbooks may have more financial resources for book purchases because they are in wealthier neighborhoods and, thus, such students may perform better on tests. If that were the case, then the notion that textbooks caused differences in student performance would be spurious; both the textbooks caused differences in student performance would be spurious: both the observed textbook adoption and the differences in performance are evidence of a different cause that was not considered. One alternative mechanism is that teachers willingly using the new textbooks may be more knowledgeable or more open to teaching in new ways—so it is not the textbook adoption itself but the teachers' use of the new textbooks. Ultimately, researchers should look to competing theories or conceptual frameworks to identify the various possible causes for an effect. Then, the researcher must be sure to gather data on these alternatives to be considered during the analysis phase.

Analysis Approaches

Analyses for causal-comparative studies are varied, but analysis of covariance (ANCOVA) and multiple regression are most common. Both ANCOVA and multiple regression allow the researcher to consider alternative explanations while also testing the proposed causal variable by (a) including other grouping variables in addition to the proposed cause and (b) accounting for other covariates that may influence the relationship between proposed cause and the effect, such as preexisting differences on other measures. ANCOVA is somewhat more common where the proposed cause involves more than two groups, but this is also available in multiple regression using *dummy-coded variables*. Alternative explanations are eliminated by demonstrating they are statistically nonsignificant or that they have weaker relationship with the outcome than the proposed cause (for ANCOVA, using effect estimates like partial h^2; for multiple regression, using standardized coefficients or changes in R^2).

Advances in statistical techniques are also prompting changes in analyses in causal-comparative designs. Hierarchical linear modeling and structural equation modeling are increasingly used for causal-comparative studies, especially for studies of large-scale survey data, as these methods and appropriate software become more widespread in education research. These techniques can allow better estimates for standard errors in situations with nested data (such as students who are part of an intact class) or for testing competing causal relationships simultaneously.

Gavin W. Fulmer

See also Causal Inference; Correlation; Experimental Designs; Scientific

Method

Further Readings

Bliss, S. L., Skinner, C. H., Hautau, B., & Carroll, E. E. (2008). Articles published in four school psychology journals from 2000 to 2005: An analysis of experimental/intervention research. Psychology in the Schools, 45(6), 483–498. doi:10.1002/pits.20318

Fraenkel, J. R., & Wallen, N. E. (2008). How to design and evaluate research in education (7th ed.). New York, NY: McGraw-Hill Education.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2011). Educational research: Competencies for analysis and applications (10th ed.). Upper Saddle River, NJ: Pearson.

Johnson, R. B. (2001). Toward a new classification of nonexperimental quantitative research. Educational Researcher, 30(2), 3–13. Retrieved from https://doi.org/10.3102/0013189X030002003

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. Review of Educational Research, 78 (3), 608–644. doi:10.3102/0034654308324427

Mertens, D. M. (2014). Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods. Thousand Oaks, CA: Sage.

Ceiling Level

Ceiling level

254

254

Ceiling Level

See Basal Level and Ceiling Level

Arthur Charpentier Arthur Charpentier Charpentier, Arthur

Central Limit Theorem

Central limit theorem

254

258

Central Limit Theorem

The central limit theorem is a fundamental theorem of statistics. It prescribes that the sum of a sufficiently large number of independent and identically distributed random variables approximately follows a normal distribution.

History of the Central Limit Theorem

The term *central limit theorem* most likely traces back to Georg Pólya. As he recapitulated at the beginning of an article published in 1920, it was "*generally known that the appearance of the Gaussian probability density* exp (–x2)" in a great many situations "*can be explained by one and the same limit theorem*" which plays "*a central role in probability theory*." Pierre-Simon Laplace had discovered the essentials of this fundamental theorem in 1810, and with the designation *central limit theorem of probability theory*, which was even emphasized in the article's title, Pólya gave it the name that has been in general use ever since.

In this article of 1820, Laplace starts by proving the central limit theorem for some certain probability distributions. He then continues with arbitrary discrete and continuous distributions. But a more general (and rigorous) proof should be attributed to Siméon Denis Poisson. He also intuited that a weaker version could easily be derived. As for Laplace, for Poisson the main purpose of that central limit theorem was to be a tool in calculations, not so much to be a mathematical theorem in itself. Therefore, neither Laplace nor Poisson explicitly formulate any conditions for the theorem to hold. The mathematical formulation of the theorem is attributed to the St. Petersburg School of probability, from 1870 until 1910, with Pafnuty Chebyshev, Andrey Markov, and Aleksandr Liapounov.

Mathematical Formulation

Let X_1,X_2,…,X_n be independent random variables that are identically distributed, with mean μ and finite variance σ^2. Let

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

denote the empirical average, then from the law of large numbers tends to 0 as *n* tends to infinity. The central limit theorem establishes that the distribution of tends to a centered normal distribution when *n* goes to infinity. More specifically,

$$p\left(\sqrt{n}\,\frac{[\bar{X}_n - \mu]}{\sigma} \le x \right) \to \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2} \right) dz.$$

We can also write

$$\sqrt{n}\left(\frac{[\bar{X}_n - \mu]}{\sigma} \right) \xrightarrow{L} N(0,1)$$

or as *n* → ∞.

A Limiting Result as an Approximation

This central limit theorem is used to approximate distributions derived from summing, or averaging, identical random variables.

Consider for instance a course where 7 students out of 8 pass. What is the probability that (at least) 4 failed in a class of 25 students. Let *X* be the dichotomous variable that describes failure: 1 if the student failed and 0 if the student passed. That random variable has a Bernoulli distribution with parameter *p* = 1/8 (with mean 1/8 and variance 7/64). Consequently, if students’ grades are independent, the sum $S_n = X_1 + \ldots + X_n$ follows a binomial distribution, with mean *np* and variance *P* (1 − *p*), which can be approximated, by the central limit theorem, by a normal distribution with mean *np* and variance *np*(1 − *p*). Here, $\mu = 3.125$ while $\sigma^2 = 2.734$. To compute $P(S_n \le 4)$, we can use the cumulative probabilities of either the binomial distribution or the Gaussian approximation. In the first case, the probability is 80.47%,

$$\left(\frac{7}{8}\right)^{25} + 25\left(\frac{7}{8}\right)^{24}\left(\frac{1}{8}\right) + \frac{25 \cdot 24}{2}\left(\frac{7}{8}\right)^{23}\left(\frac{1}{8}\right)^{2} + \frac{25 \cdot 24 \cdot 23}{2 \cdot 3}\left(\frac{7}{8}\right)^{22}\left(\frac{1}{8}\right)^{3} + \frac{25 \cdot 24 \cdot 23 \cdot 22}{2 \cdot 3 \cdot 4}\left(\frac{7}{8}\right)^{21}\left(\frac{1}{8}\right)^{4}$$

In the second case, use a continuity correction and compute the probability that S_n is less than 4 + 1/2. From the central limit theorem:

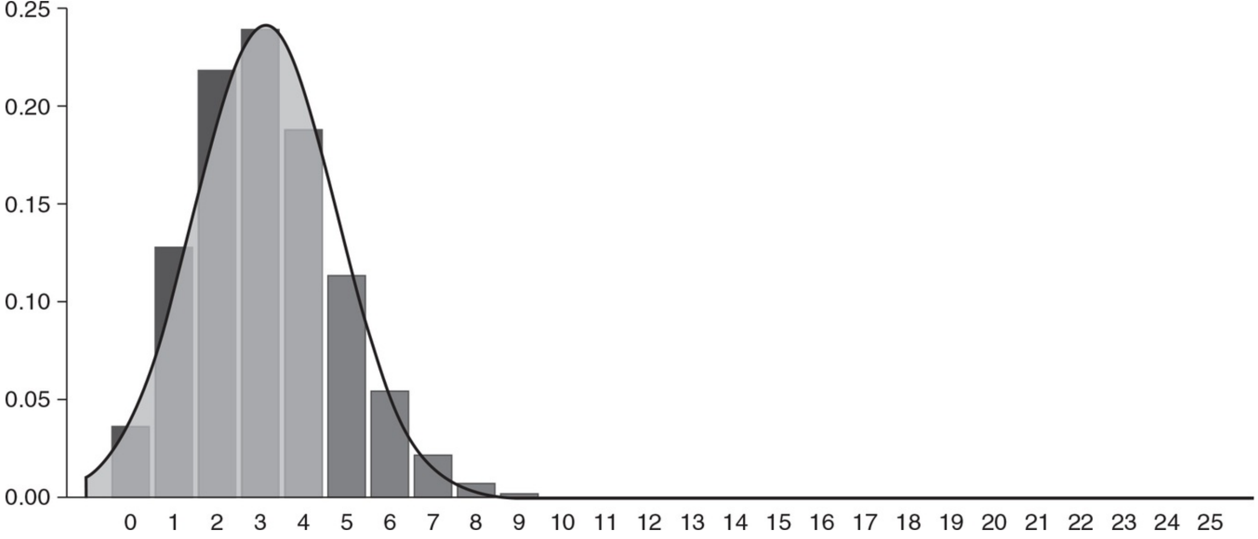
$$\sqrt{n}\,\frac{[\bar{X}_n - \mu]}{\sigma} = \sqrt{25}\,\frac{[4.5/25 - 1/8]}{\sqrt{7/64}} = 0.8315.$$

The probability that a standard Gaussian variable is less than this quantity is:

$$P(Z \le 0.8315) = 79.72\% \text{ where } Z \sim N(0,1),$$

which can be compared with 80.47% obtained without the approximation (see Figure 1). Note that this approximation was obtained by Abraham De Moivre, in 1713, and is usually known as Bernoulli’s law of large numbers.

Figure 1 Gaussian approximation of the binomial distribution



Asymptotic Confidence Intervals

The intuition is that a confidence interval is an interval in which one may be confident that a parameter of interest lies. For instance, that some quantity is measured, but the measurement is subject to a normally distributed error, with known variance σ^2. If X has a $N(\mu,\sigma^2)$ distribution, we know that:

$$P\left(\mu - 1.96 \bullet \sigma < X < \mu + 1.96 \bullet \sigma\right) = 95\%$$

Equivalently, we could write:

$$P\left(X - 1.96 \bullet \sigma < \mu < X + 1.96 \bullet \sigma\right) = 95\%$$

or

$$P(\mu \in [X \pm 1.96 \bullet \sigma]) = 95\%.$$

Thus, if *X* is measured to be *x*, then the 95% confidence interval for μ is [*x* ± 1.96…σ].

In the context of Bernoulli trials (described earlier), the asymptotic 95% confidence interval for *p* is:

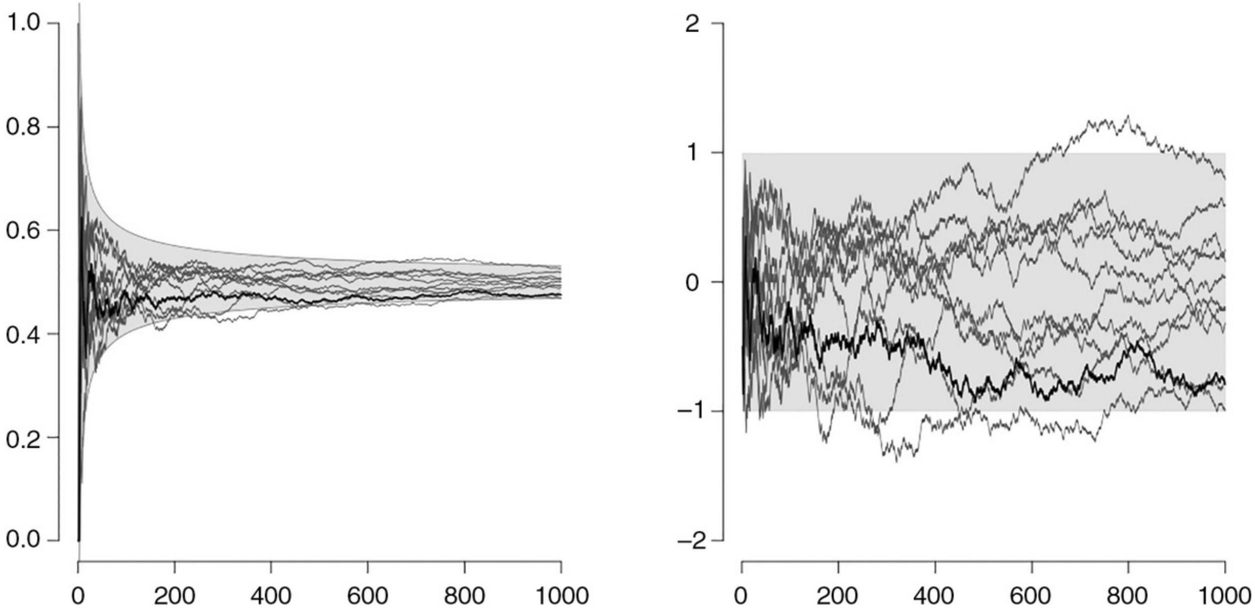
$$\left[\bar{x} \pm \frac{1.96}{\sqrt{\bar{x}(1-\bar{x})}} \cdot \frac{1}{\sqrt{n}}\right].$$

A popular rule of thumb can be derived when *p*~50%. In that context is close to 1.96 (or 2), and a 95% approximated confidence interval is then

$$\left[\bar{x} \pm \frac{1}{\sqrt{n}}\right]$$

(see Figure 2). If that confidence interval provides a good approximation for the 95% confidence interval when *p*~50%, it is an over-estimation when *p* is either much smaller, or much larger.

Figure 2 Law of large numbers on the left, with the convergence of toward *p* as *n* increases, and central limit theorem on the right, with the convergence of towards a Gaussian distribution. The shaded area is the 95% confidence region.



The Delta Method and Method of Moments

This method is used to approximate a general transformation of a parameter that is known to be asymptotically normal. Assume that:

$$\sqrt{n}\,(Z_n - \mu) \xrightarrow{\;L\;} N(0,\sigma^2) \quad \text{as } n \to \infty,$$

then

$$\sqrt{n}\,(h(Z_n) - h(\mu)) \xrightarrow{\;L\;} N\left(0, h'(\mu)^2 \cdot \sigma^2\right) \ \text{as}\, n \to \infty.$$

For some continuous transformation *h* such that *h*'μ ≠ 0.

Consider now a parametric model, in the sense that X_1, X_2, …, X_n are independent random variables, with identical distribution F_θ (which can be a Weibull distribution to model a duration, a Pareto distribution to model the income or the wealth, etc.), with unknown parameter θ. The method of moments is a method of estimating parameters based on equating population and sample values of certain moments of the distribution. For instance, if $E[X] = \mu(\theta)$, then the estimator of the unknown parameter is given by equation or equivalently . From the central limit theorem, we know that:

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{\;L\;} N(0,\sigma^2) \quad \text{as } n\, \infty$$

and applying the delta method with $h = \mu^{-1}$, then:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\;L\;} N\left(0, h'(h^{-1}(\theta))\right)^2 \cdot \sigma^2 \quad \text{as } n \to \infty$$

where a numerical approximation for the variance can be derived. This method has a long history, and has been intensively studied. Furthermore, this asymptotic normality can be used to compute a confidence interval, and also to derive an asymptotic testing procedure.

An Asymptotic Testing Procedure

Based on that asymptotic normality, it is possible to derive a simple testing procedure. Consider a test of the hypothesis H_0: θ = 0 against H_1: θ ≠ 0, usually called a “significant” test for parameter θ (or significance of an explanatory variance in the context of regression model). Under the assumption that H_0 is valid, then for some variance s^2, that can be computed using the delta method. The *p* value associated with that test is:

$$p = P\left(|Z| > \left| \frac{\hat{\theta}_{obs}}{s} \right| \right)$$

where is the observed empirical estimator of the parameter and *Z* is a standard normal variable. Thus, the *p*-value can easily be computed using quantiles of the standard normal distribution. Here, the *p*-value is above 5% if:

$$-1.96 < \frac{\hat{\theta}_{\text{obs}}}{s} < 1.96.$$

Weaker Forms of the Central Limit Theorem

As stated by Laplace, the central limit theorem relies on strong assumption. Hopefully, most of them can be relaxed. In a first variant of the theorem, random variables have to be independent, but not necessarily identically distributed. If random variables X_{i} have averages μ_i and variances , then μ and σ_2 in the central limit theorem should be replaced by averages of μ_{i} and s, with an additional technical assumption related to the existence of some higher moments (the so-called Lyapunov condition).

For a second variant of the theorem, random variables can be dependent, as in ergodic Markov chain, or in autoregressive time series. In that context, if X_1, X_2, …, X_{n} is a stationary time series, with mean μ, then define:

$$\sigma^2 = \lim_{n\to\infty} \frac{E[S_n^2]}{n}$$

and with that limit, the central limit theorem hold:

$$P\left(\sqrt{n}\,\frac{[\bar{X}_n - \mu]}{\sigma} \le x \right) \to \Phi(x),$$

even if the variance term has here a different interpretation.

Finally, a third variant that can be mentioned is the one obtained by Paul Lévy about asymptotic properties of the empirical average, when the variance is not finite (actually, even when the first moment in not finite). In that case, the limiting distribution is no longer Gaussian.

Arthur Charpentier

See also Bernoulli Distribution; *F* Distribution; *t* Tests

Further Readings

de Laplace, P.S. (1810). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. Mémoires de l’Académie Royale des Sciences de Paris, 10. See http://www.cs.xu.edu/math/Sources/Laplace/approximations%20of%20formul

Le Cam, L. (1986). The central limit theorem around 1935. Statistical Science 1(1): 78–96.

Polyà, G. (1920). Ueber den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem. Mathematische Zeitschrift 8, 171–181. See http://gdz.sub.uni-goettingen.de/dms/load/img/?PPN=PPN266833020_0008

Alisa Palmer Branham Alisa Palmer Branham Branham, Alisa Palmer

Certification

Certification

258

259

Certification

Educator certification is the process of documenting an individual's qualifications to practice teaching, administration, or special services in a public school. The standards and regulations of certification are dynamic and reflect the complex political and social issues that affect public education in the United States. The underlying purpose for certification is to ensure high-quality, competent educators.

In the 18th and 19th centuries, teachers were hired based on their ability to pass a locally accepted evaluation and the possession of "good moral character." Over time, states exerted more control over certification, and institutions of higher education developed pedagogy-based programs in efforts to boost professionalism. Typical requirements for new teachers in the 21st century include a minimum of a bachelor's degree, completion of an approved program of teacher preparation, passing scores on standardized certification exams, and a criminal background clearance. Most states also require recent knowledge or experience, referred to as recency, and there may be additional training mandated in topics such as first aid or identifying child abuse.

Traditional approved programs for teacher preparation consist of coursework in content and pedagogy, along with field experiences including a university-supervised student teaching experience. Alternative routes to certification come in many forms and typically allow a person with a prior bachelor's degree to start teaching full time while completing pedagogy courses or professional development in a state-approved program. In alternative programs, paid classroom experience replaces the supervised student teaching. These programs may be offered through a university, a school district, or a state agency.

may be offered through a university, a school district, or a state agency.

An initial certification program provides the training for the main/first teaching certificate. Additional content areas may be added to the initial certificate in several ways as allowed by each state; options may include completing an additional approved program, passing a list of specific courses, or passing the appropriate state subject area exam. For example, in some states, a licensed elementary K–6 teacher might be allowed to add an endorsement in middle-level math 5–8 by passing a state subject exam.

Certification terminology varies state by state and country by country. Some states issue a *certificate* while others issue a *license*, and California issues a *credential*. Educators must file for a new certificate when they move to a new state. Reciprocity allows some states to issue a certificate if the applicant holds a similar certificate in another state based on comparable requirements.

There is no reliable centralized source for certification/licensure information across all 50 states. Information on requirements in each state can be obtained from local experts, including personnel at state agencies, certification officers at institutions of higher education, and human resources staff in local school districts.

Teacher salaries frequently constitute a major portion of the state’s overall budget, so there are fiscal as well as qualitative reasons for certification. In most states, proper certification is required, so that educational personnel may be paid from the correct pool of money. Supply and demand may affect changes in certification requirements too, as states adapt in order to bring new candidates into the teaching ranks.

Alisa Palmer Branham

See also Common Core State Standards; Teachers’ Associations

Further Readings

Angus, D. Professionalism and the public good: A brief history of teacher certification. Retrieved from https://edex.s3-us-west-2.amazonaws.com/publication/pdfs/angus_7.pdf

Ravitch, D. A brief history of teacher professionalism. Retrieved from

http://www2.ed.gov/print/admins/tchrqual/learn/preparingteachersconference/r

James Wollack James Wollack Wollack, James

Rachel Watkins Schoenig Rachel Watkins Schoenig Schoenig, Rachel Watkins

Cheating

Cheating

259

265

Cheating

In general terms, cheating is an action taken by an individual to intentionally bias assessment results. The “individual” involved can be anyone with knowledge of or access to testing materials and/or the testing process: “testing materials” include test items, test booklets, scoring templates, answer sheets, score reports, or databases for item responses or test scores; and “testing processes” include test development, technical aspects of test delivery, test proctoring, test scoring, and test reporting. Cheating may involve one or more examinees, educators, third-party test preparation entities, testing staff, parents, or representatives of the testing company or its various partners and vendors. Although many actions may result in biased test scores, cheating requires that the actions are done with the goal of biasing the results. Similarly, whether the assessment results are actually biased, or biased in the intended direction, is irrelevant under this definition.

Cheating is important because it creates a fundamental fairness issue among examinees, and to the extent it allows individuals to acquire a license or credential to practice in a discipline for which they are unqualified may also present a possible threat to the health, safety, and well-being of the public. This entry begins by providing a general context for cheating and follows with discussions of preventing, deterring, and impeding cheating, detecting cheating, and deciding how to address cheating.

The Cheating Context

Although no compelling trend data exist to suggest that the overall prevalence of cheating has changed dramatically since the 1960s, since the start of the 21st century, the prevalence and magnitude of cheating have garnered more national and international media attention. Furthermore, the methods used for cheating have evolved as technology has evolved.

Cheating on assessments occurs across the globe and can involve individuals at any age and educational level. Individuals may feel more pressure to cheat when they perceive exam results will have a reputational, financial, or employment impact. The greater the impact of assessment results, whether positive or negative, the more likely individuals will be to engage in cheating. Because assessment results can significantly impact a variety of stakeholders, numerous individuals with access to testing materials or the assessment process may have an incentive to cheat.

Cheating can have measurement, societal, and financial consequences. Cheating on assessments can result in inaccurate data and invalid measurement results. When an individual cheats on a norm-referenced test, such as a classroom exam or an admissions/employment test, the cheater can gain an unfair advantage over others. When educators cheat, the conclusions drawn from invalid scores may result in a failure to provide students adequate instruction, inappropriate district-wide adjustments in curriculum and staffing, and skewed teacher evaluations or unwarranted salary adjustments. When examinees cheat to obtain professional credentials, individuals may be allowed to practice in an area that has direct impact on the health and safety of the public. Depending on the extent of cheating, society may determine legal consequences, such as civil or criminal liability, are appropriate. Tangible financial costs are associated with investigating cheating, invalidating scores, terminating testing staff, and engaging in administrative, civil, or criminal actions.

There are several common methods used to cheat on standardized exams, including copying, unauthorized use of exam aids (e.g., notes written on paper or clothing or answers stored or transmitted through digital means), use of a surrogate tester, examination preknowledge, and tampering. Examinees may attempt to gain access to secure test materials, including through theft of paper materials or digital hacking, in an attempt to gain knowledge of exam questions and answers prior to the exam administration. During administration, examinees may capture test content or answers with the intent of compromising score validity for subsequent administrations. Educators or training program employees may provide students with preknowledge by viewing the test in employees may provide students with preknowledge by viewing the test in advance of the administration and teaching the questions and answers to students. Educators or test administrators may tamper with test results by coaching examinees during the test or changing responses after testing has concluded.

Left unchecked, cheating can become widespread. Thus, it is helpful for testing programs and score users, such as school districts, universities, credentialing programs, and employment entities, to have in place a holistic framework for addressing cheating, including steps to deter or prevent cheating, tools to detect potential cheating incidents, and processes to decide how to respond to incidents. Furthermore, because numerous individuals can engage in conduct that threatens valid assessment results, it is necessary to consider deterrence, detection, and decision-making across a broad range of actors.

Preventing, Deterring, and Impeding Cheating

The first line of defense against cheating is to attempt to prevent it entirely or, alternatively, to deter it from happening or impede its effectiveness in the event it does occur. Testing programs use multiple strategies to prevent, deter, and impede cheating on tests.

Test Design

Test developers have a range of tools available to make cheating more difficult and less likely to be successful. Different testing modalities pose different advantages and disadvantages with respect to cheating. Single-form, linear testing, in which all examinees see the same items in the same order, presents the greatest cheating risk because examinees who copy from neighboring examinees have a high probability of improving their performance, and examinees entering the testing room with prior information or preknowledge about operational test questions are guaranteed to see those items during the exam. Scrambling test items so they appear in different orders for different examinees may help prevent or reduce answer copying as well as reduce its negative impact. Using multiple equated test forms has many security advantages, including reducing the likelihood of neighboring examinees seeing the same items and reducing the likelihood examinees with preknowledge will be administered the compromised items. In addition, programs can ensure that retesting examinees are administered a different form of the test with limited content overlap administered a different form of the test with limited content overlap.

Computerized adaptive testing (CAT), in which each examinee receives a tailored exam, optimized for the person’s performance level, offers many security advantages over paper-based or computer-based linear testing. Answer copying is essentially eliminated in CAT; moreover, variable length CATs are inherently self-correcting for unusual responses, as might be the case for examinees with preknowledge of difficult items, provided there are no constraints on the length of the exam because in CAT, spuriously correct answers lead to more difficult questions for which the examinee’s probability of correct response is lower. As a countermeasure against examinees entering with preknowledge, CATs often rely on several large, regularly rotated item pools and use a variety of methods to control the exposure rates of individual items.

Perhaps the most serious security vulnerability of computer-based tests is that to improve access and convenience, tests are often administered over extended periods, referred to as testing windows, ranging from several days to a few months. Although programs can implement security measures between testing windows, such as using entirely new item pools and limiting examinees to a single testing attempt per window, different examinees testing within a common window will often see a high degree of overlap in test content. Utilizing testing windows raises the risk of examinees entering with preknowledge obtained from an examinee who tested earlier in the window. Narrow testing windows can help reduce this type of cheating.

Communication and Contracting

To help deter cheating, testing programs should clearly communicate to examinees and testing staff what is considered appropriate (or prohibited) behavior and what materials are allowed (or not) in the testing room. Programs should also clearly denote copyrights and communicate potential consequences for violations of test security policies or agreements. Sanctions for cheating typically include outcomes ranging from score cancelation to legal action, including civil claims and criminal prosecution.

Information on cheating may be communicated to examinees through pretest instructions or in a contract agreed to at the time of registration or immediately prior to the test administration. It may also be communicated again following the exam to reinforce examinee obligations, including the obligation to maintain the confidentiality of exam content. In addition, testing program employees, test confidentiality of exam content. In addition, testing program employees, test administration staff, and vendors are typically asked to sign confidentiality agreements, which require those individuals to maintain the security and confidentiality of the testing materials to which they have access.

Check-in

One of the most common and potentially successful cheating strategies is to access and utilize prohibited items during the exam, such as notes, the Internet, communication devices, or imaging equipment. Because detecting prohibited items during testing is difficult, it is common to take steps to locate and exclude them from the exam environment. Testing staff often require examinees to turn out their pockets to demonstrate they are empty, place hats, sunglasses, or other disallowed accessories in a storage locker outside the testing room, or submit suspicious outerwear, such as baggy sweatshirts or overcoats, to inspection prior to entering the testing room. Metal detection devices can be used to scan rooms or examinees for prohibited technology.

To address the risk of surrogate testing, examinees are commonly required to provide identification prior to entering the testing room, such as a government-issued photo ID, which can be cross-referenced by testing staff with the examinee's registration data to ensure physical likeness and identical matching of demographic data collected during the registration process. Many programs also use biometric data, such as palm vein scans or voice recognition, to authenticate candidates. Identification procedures are often used every time an examinee enters the testing room, including after scheduled and unscheduled breaks. Examinees without appropriate matching identification credentials are typically not allowed to test.

Test Administration and Proctoring

During group examinations, examinees are ideally randomly assigned seats by testing staff to deter collaboration among examinees. Seating examinees facing front with wide space between them makes answer copying or communication with other examinees more difficult.

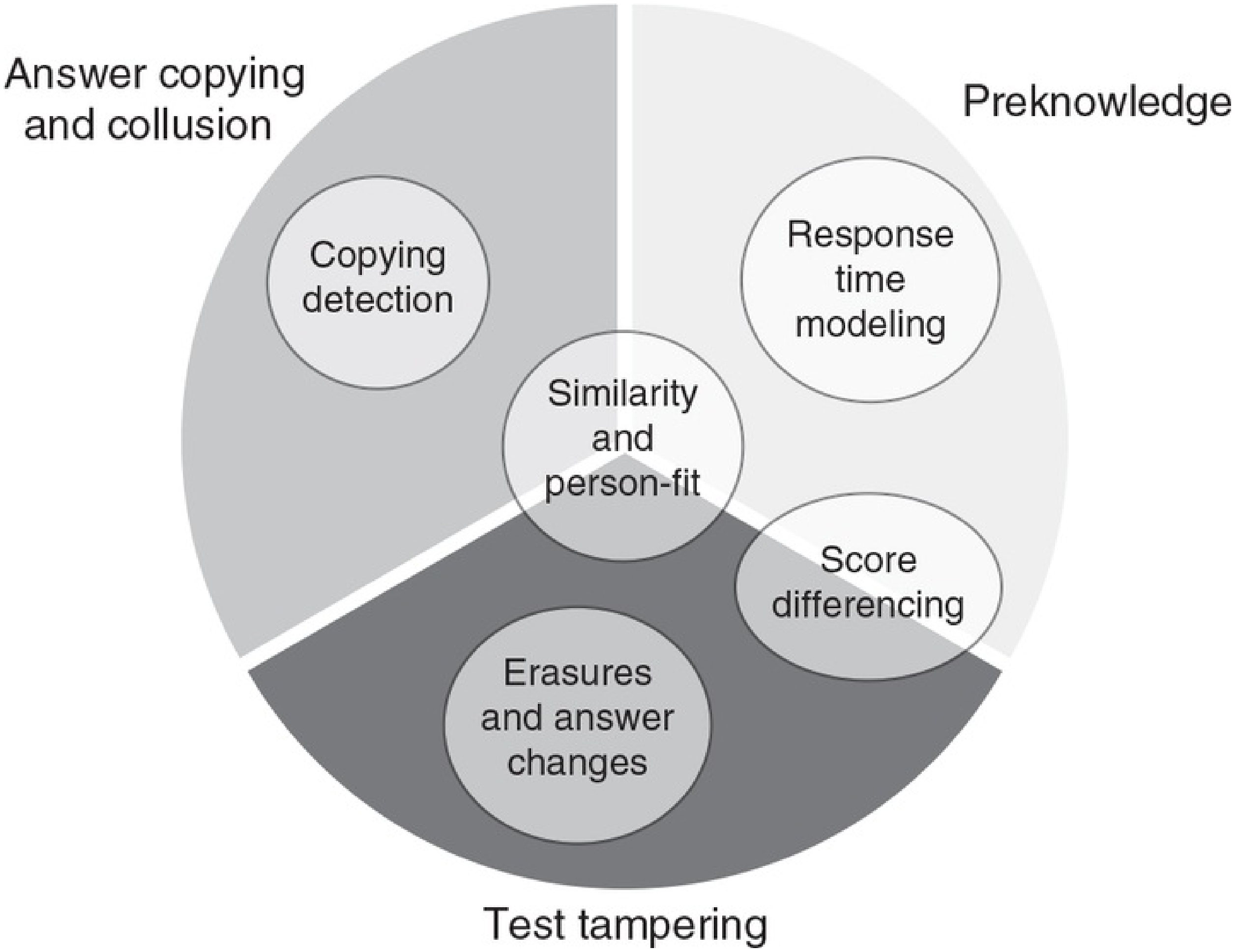
Managing examinees during breaks is critical, especially during scheduled breaks, where all examinees are breaking simultaneously and communication between examinees is expected. Having a defined break area that restricts examinees’ abilities to interact with individuals outside the testing center is common. Testing staff often remind examinees not to discuss test content or access technology or test materials during the break and monitor examinees for compliance.

Proctors should be free of conflicts of interest to help reduce the risk of cheating by the proctor. Enough proctors should be used to adequately monitor testing activities and actively look for signs of cheating. Testing staff should be familiar with and adhere to the standardization protocol for each test administered. Ideally, staff should be able to identify and respond to a wide range of cheating behaviors and unusual situations (such as fire alarms) which could increase opportunities for cheating.

Detecting and Investigating Cheating

Despite testing programs’ best efforts, prevention strategies are not always successful and examinees manage to successfully cheat on exams. Consequently, best practice is to also devote resources to detecting cheating after it happens. The late 20th and early 21st centuries have seen numerous advances in statistical methodologies designed explicitly for detecting different forms of test cheating. Figure 1 depicts the broad categories of statistical approaches and the types of cheating they are commonly used to detect.

Figure 1 Statistical methodologies and common uses for detecting cheating



Person-Fit

Cheating often produces response patterns that are not explained well by the psychometric models used to fit the data. Person-fit involves conducting a statistical test to identify nonmodel-fitting examinees. Many different person-fit measures exist, and in practice, testing for fit is often used as a first step to reduce the pool of potential cheaters to a more reasonable number. However, because there are many legitimate reasons why an examinee may produce a nonfitting response pattern, such as time pressure, creative responding, carelessness, and unconventional instruction, person-fit measures often detect a lower proportion of actual cheaters than do other statistical methods.

Copying Detection and Similarity

Early work on cheating detection focused almost entirely on answer copying.

With answer copying becoming less prevalent due to computerized testing, for many programs, focus has shifted to answer similarity. Copying and similarity indexes flag examinees who produce an observed number of answer matches that significantly exceeds the number of matches predicted under the model. Computationally, the difference between the two rests on the directionality of the index. With answer copying indexes, the expected number of matches is computed conditionally on the estimated trait level for the suspected copier and the answer string for the suspected source, thereby producing different index values depending on which examinee is evaluated as the copier. With similarity indexes, the expected number of matches is computed by finding the probability of matching on any of the item alternatives, conditioned on the trait levels for both examinees. The symmetry of similarity indexes allows them to be used to identify groups of examinees with similar responses; hence, they have proven useful for detecting other forms of cheating, such as preknowledge and test tampering. For both copying and similarity indexes, research has found that using all items, rather than only jointly incorrect items, results in the highest detection rates.

Score Differencing

Score differencing detects preknowledge by identifying examinees whose performance on a set of compromised items is significantly different than is expected based on their performance on secure items. Score differencing works best when it is known which items have been compromised, as might be the case when items are found on the Internet or examinees are found with printouts of live test content. However, when no known compromise exists, programs will often compare examinees’ performances on pilot and operational items or on items with many exposures and items with few exposures. Because preknowledge usually results in inflated scores, it is typically easier to identify examinees with preknowledge when difficult items become compromised. However, in special cases where programs include items that are modified slightly from the compromised versions, blind reliance on preknowledge will lower the examinee’s score and score differencing will work better if the modified items are easy.

Gain scores represent a special type of score differencing, in which an examinee’s current test score is compared to a previous one. Retest candidates typically improve their performance; however, candidates who improve by more than the expected amount may be suspected of having preknowledge or having engaged in some other type of cheating. In educational accountability settings, students, classes, or districts that show unusual growth may be flagged for investigation into potential inappropriate coaching or educator tampering.

Erasures and Answer Changes

Research has found that examinees change their test answers infrequently; most estimates are that changes occur on approximately 2% of the items. Furthermore, when examinees do change their answers, approximately half the changes are expected to involve switching from a right to a wrong answer or from one wrong answer to another. Consequently, examinees with large numbers of answers that have been changed from wrong to right are suspicious and may indicate test tampering. Answer changing methods have most commonly been applied to K–12 accountability exams. The most frequently used answer change methods involve comparing the number or average number of wrong to right changes per student, often aggregated to the class, school, or district level. Although most of these methods have not been well researched and their effectiveness at detecting different amounts and magnitudes of tampering is not well understood, they have been used to help uncover many large-scale instances of educator cheating.

A second, newer class of answer change methodologies adopts a score differencing approach by comparing performance across those items for which the initial answer was changed with the expected score on those items based on the examinee’s performance on all other items. Although the research on these methods is limited, they are theoretically preferable to count-based methods and appear to work quite well at detecting moderate-to-large amounts of tampering.

Response Time Methods

In computer-based testing, it is possible to collect data on the amount of time examinees spend answering each question. Response time has been shown to vary considerably across items, as a function of reading load, cognitive load, and the amount of computation, as well as intraindividual differences. Several response time models exist which predict the amount of time necessary to complete an item, based on both item and person characteristics. Irregular response time patterns—spending considerably more or less time than expected on certain items—can be useful for detecting examinees with preknowledge or attempting to steal exam content. Response time models are sensitive to examinees who answer many items very quickly, particularly if those items were largely answered correctly or were believed to be compromised. One unique feature of response time models is that multiple examinees producing both correct answers and unusually short response times on common items can help identify compromised items that had previously been believed to be secure.

Other Methods for Detecting Cheating

Cheating detection is not limited to statistical approaches. Although proctoring was discussed earlier as a deterrent to cheating, it is also one of the more common mechanisms for detecting cheating. Many programs establish an anonymous reporting service that enables individuals to report potential test security concerns through a secure hotline or webform. Frequent web monitoring can identify examinee attempts to hire surrogates or share live items. Use of nonscore-related data, such as examinee travel patterns or irregularity reports at test sites, can help identify or predict patterns of unusual behavior that can be used to trigger an investigation or increase monitoring.

Deciding How to Address Cheating

Once a potential cheating incident has been detected, several decisions are necessary. The first is whether and to what extent the potential incident will be investigated. There is professional debate concerning whether statistical evidence, standing alone, is sufficient to cancel scores. The traditional and conservative approach is to use statistical evidence to trigger an investigation. However, as methods have advanced, some programs are using results from statistical analyses to void or cancel scores.

If additional investigation is conducted, qualified and objective individuals ideally should conduct the investigation. The investigators may collect a wide range of evidence, such as interviews of examinees and testing staff, collection and evaluation of relevant materials, additional statistical analyses, and biometric data evaluation. Depending on who is implicated in the investigation, it may be necessary to include union or legal representatives in the investigation.

After the investigation concludes, the next decision is the appropriate body to evaluate the evidence and the process to be used for deciding consequences, if any. This may be impacted by state laws, agency regulations, local policies, and contract terms. Depending on the circumstances, the decision body may range contract terms. Depending on the circumstances, the decision body may range from a teacher to an ethics committee, independent panel, or credentialing authority. Processes for deciding consequences may range from an informal meeting to a formal, recorded hearing and written decision by an administrative body. Typically, these bodies are asked to determine whether a score is invalid or whether cheating occurred by a preponderance of the evidence, and decisions rendered will be held to a reasonableness standard.

The next decision is what consequences to impose for individuals found to have engaged in cheating. Consequences will vary with the circumstances and evidentiary findings. For example, where cheating has impacted score validity, scores may be cancelled or voided at either an individual or group level and credentials may be revoked. Notice of the decision may be provided to third parties, such as school districts, colleges, employers, or law enforcement. Individuals found to have engaged in cheating may be required to undergo ethics training, be prohibited from retesting (or administering assessments, in the case of an educator or test administrator) for a period of time, be expelled from school, or have their credentials rescinded.

The final decision is whether, based on the investigative findings, steps can be taken to improve the assessment process. For example, in an effort to deter future misconduct, school districts or colleges and universities may decide to invest in additional staff training or create more messaging to students to ensure all stakeholders understand the consequences of cheating.

James Wollack and Rachel Watkins Schoenig

See also Test Security; Tests

Further Readings

Cizek, G. J. (1999). Cheating on tests: How to do it, detect it, and prevent it. Mahwah, NJ: Erlbaum.

Cizek, G. J., & Wollack, J. A. (Eds.). (2017). Handbook of quantitative methods for detecting cheating on tests. New York, NY: Routledge.

Josephson Institute. (2012). 2012 report card on the ethics of American youth. Los Angeles, CA: Author.

Kingston, N. M., & Clark, A. K. (2014). Test fraud: Statistical detection and methodology. New York, NY: Routledge.

Lang, J. M. (2013). Cheating lessons: Learning from academic dishonesty. Cambridge, MA: President and Fellows of Harvard College.

National Council on Measurement in Education. (2012). Testing and data integrity in the administration of statewide student assessment programs. Madison, WI: Author.

Olson, J. F., & Fremer, J. J. (2013). TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities. Washington, DC: Council of Chief State School Officers.

Wollack, J. A., & Case, S. M. (2016). Maintaining fairness through test administration. In N. J. Dorans & L. L. Cook (Eds.), Fairness in educational assessment and measurement (pp. 33–53). New York, NY: Routledge.

Wollack, J. A., & Cizek, G. J. (2017). Security issues in professional certification/licensure testing. In S. Davis-Becker & C. Buckendahl (Eds.) Testing in the professions (pp. 178–209). New York, NY: Routledge.

Wollack, J. A., & Fremer, J. J. (Eds.). (2013). Handbook of test security. New York, NY: Routledge.

Meagan M. Patterson Meagan M. Patterson Patterson, Meagan M.

Childhood

Childhood

265

268

Childhood

Historically, childhood has been defined as the time period prior to adulthood or maturity. A more precise contemporary definition frames childhood as the period between the end of infancy and the beginning of adolescence or from approximately 2–12 years of age. The juvenile period of physical and cognitive immaturity is substantially longer relative to the overall life span for humans compared to other species. This lengthy period of immaturity is due to the complexity of the human brain.

Humans' lengthy childhood allows for substantial flexibility in development, such that individuals can adapt to a wide range of physical and cultural environments. Understanding of children and their development is critical to educational policy and practice. This entry discusses the history of the construct of childhood, methods for studying children and their development, and major domains of child development.

History

For much of history, children were viewed as the property of their parents (particularly their fathers), and the goal of child-rearing was to have children assume adult responsibilities as quickly as possible. It was common for children to work full days on farms or in factories. This version of childhood persists in many parts of the world. However, in developed countries, the view of children and childhood has changed substantially, such that children are generally protected from adult responsibilities such as paid work, and parents devote substantial time and resources to ensuring that their children are healthy and happy. A number of social, economic, and cultural factors contributed to these changes, including a move away from agrarian economies, increased availability of formal schooling, and decreased infant and child mortality rates.

Along with changes regarding children’s value and roles, there have also been changes throughout history in views of the fundamental nature of children. For much of Western history, perceptions of children were shaped by the notion of “original sin”—a belief that people are born with an inclination toward evil and that children must be trained away from these tendencies by adults, often through the use of harsh discipline. In the 18th century, this view was called into question by the writings of two philosophers, Jean-Jacques Rousseau and John Locke.

Rousseau viewed children as “noble savages” who were born with an innate sense of right and wrong and were largely harmed, rather than helped, by adult intervention. Rousseau argued that the best outcomes would arise from allowing children to develop naturally, with minimal intervention from adults. Locke, in contrast, viewed the child as a blank slate, beginning from nothing and being shaped, either positively or negatively, by the actions of parents and other adults.

Both Locke and Rousseau argued that harsh treatment of children was not beneficial for their development. Their writings influenced the emerging field of developmental psychology in a variety of ways, including the examination of the roles of nature (innate characteristics, traits, and predispositions, such as those emphasized by Rousseau) and nurture (environmental influences, such as the actions of parents and teachers, as emphasized by Locke) in children’s development.

Although philosophers and scholars have reflected on the nature of childhood and goals of child-rearing for centuries, the formal study of children’s development has a relatively short history. The first careful observations of children’s development we published in the mid-19th century included Charles Darwin’s “baby biographies.” Systematic theories and empirical study of child development emerged even later, with the work of early psychologists including Sigmund Freud, G. Stanley Hall, Melanie Klein, Jean Piaget, B. F. Skinner, Lev Vygotsky, and John Watson. These researchers sought to document and explain typical and atypical patterns of children’s development and the factors that influenced development.

Methods for Studying Children

Methods for Studying Children

Researchers use a wide variety of methods for studying children’s development. These methods include observations of behavior (both naturalistic and structured), interviews, surveys, case studies, and psychophysiological methods (measures of the physical manifestations of thoughts or emotions such as heart rate or brain activity).

In order to study the effects of age and development, researchers may use methods that examine children at different ages. Such studies might include longitudinal, cross-sectional, sequential, or microgenetic research designs. Longitudinal studies examine the same person at multiple points in time (e.g., assessing a group of children each year from ages 5 to 9). Cross-sectional studies, in contrast, examine children of different ages (e.g., 5-, 7-, and 9-year-olds) at the same point in time.

Sequential studies combine elements of cross-sectional and longitudinal designs, beginning with participants of different ages and following them over time. Microgenetic designs aim to examine children’s behavior or cognition while it is changing; these designs assess children who are thought to be on the verge of an important developmental change by studying individuals at multiple time points over a fairly short span of time (typically several months to a year).

Stages

Childhood is frequently divided into three stages: early, middle, and late childhood. Early childhood spans the period from 2 to 5 or 6 years of age (the age at which children typically begin formal schooling). Middle childhood begins around age 6 and lasts until approximately age 10, roughly corresponding to the elementary or primary school years. Late childhood (sometimes referred to as preadolescence) spans approximately ages 10–12 years, at which point most children have begun puberty and thus entered adolescence. Preadolescents are sometimes informally referred to as “tweens,” indicating their status as being in between childhood and adolescence.

Developmental Contexts

There are three developmental contexts that are viewed as most influential for children and thus most widely studied by researchers: the family, the school, and children and thus most widely studied by researchers: the family, the school, and the peer group. Relationships and experiences within these contexts can have substantial influences on children’s development. In addition, interactions of these contexts can have meaningful effects on the child. For example, parental involvement with schooling may promote children’s academic achievement, whereas conflict between parents and teachers may have a negative impact on academic performance.

Developmental Tasks

Developmental tasks are the fundamental skills and abilities that must be acquired for optimal development at a given life stage. Accomplishment of these tasks provides a foundation for progress toward subsequent developmental stages. Key developmental tasks of childhood include developing a sense of personal identity, internalizing rules and moral standards for behavior, establishing peer relationships, managing emotions, learning to solve problems independently, and engaging with school.

Socialization

Socialization is the process by which an individual acquires the attitudes, knowledge, and skills needed to succeed within a particular social or cultural context. Socialization of the child is a primary goal of parents, teachers, and other adults who regularly interact with children. The goals of socialization will vary depending on the social and cultural context. For example, in individualistic cultures, there is typically a view that the child is born helpless and dependent, and a primary goal of socialization is to promote independence and self-reliance. In contrast, in collectivist cultures, the child is often viewed as having been born independent or separate from others, and a primary goal of socialization is to foster feelings of connection to family and community members. Cultures also vary widely in other aspects of socialization, such as flexibility of gender roles, tolerance of aggressive behavior, and the importance of deference to elders.

Goodness of Fit

One important aspect of promoting optimal child development is the goodness of fit between a child’s developmental level and the environments the child experiences. For example, parents or teachers who have unrealistic expectations for children’s abilities may become frustrated with a child’s lack of performance and become harsh or rejecting with the child.

The educational construct of developmentally appropriate practice recognizes the importance of matching educational requirements to children’s cognitive and socioemotional development. In some cases, researchers have criticized the educational system for the lack of attention to children’s developmental needs. For example, when students transition from elementary to middle school, the environment often becomes more restrictive just as students are developmentally ready for greater autonomy; this lack of fit between students and their environment may decrease academic motivation and engagement.

Research on goodness of fit also emphasizes the importance of fit between a child’s characteristics (such as temperament) and the child’s environment. For example, a shy child who is born into an outgoing family may struggle with the expectations placed on her by her parents or siblings. Although goodness of fit is important throughout development, issues of fit may be especially important in childhood because children generally have less control and choice over their environments relative to adults (i.e., a child is less able than an adult to choose to leave a situation that is not a good fit).

Cognitive Development

Over the course of childhood, thought becomes increasingly logical and flexible. Executive function, the ability to monitor and regulate one’s thinking and decision making, also increases over the course of childhood. As children move from early to middle childhood, their attention spans and memory capacities increase. Thus, children are increasingly able to understand and use deliberate learning strategies to retain information and accomplish academic tasks. Throughout the course of childhood, an environment that provides appropriate levels of cognitive stimulation (e.g., talking to children, reading to and with children, exploring nature) promotes cognitive growth and development.

Emotional Development

Basic emotions, such as anger, joy, and surprise, are present from infancy. Self-conscious emotions, such as guilt and pride, emerge in the early childhood years, once the child has developed a sense of self. Over the course of childhood, children become more skilled at recognizing emotions in themselves and others and at regulating and managing their own emotions.

By middle childhood, children have a range of emotion regulation strategies at their disposal, including cognitive reframing, emotion-centered coping (i.e., behavioral strategies intended to change one’s emotional state, such as watching a funny movie), and problem-centered coping (i.e., behavioral strategies intended to change an upsetting situation, such as deciding to study harder next time after failing a test).

Identity Development

As children develop, their sense of self becomes more detailed and differentiated (moving from a generalized self-view to awareness of one’s strengths and weaknesses in a variety of areas such as academic performance, physical skills, and relationships with peers). With development, self-views also become more accurate, as children become more aware of their own capabilities in comparison to peers and objective performance standards. Awareness of one’s membership in important social categories (such as gender, race/ethnicity, and religion) increases across childhood; these social group memberships may play an important role in identity development.

Moral Development

Moral development includes multiple elements: moral reasoning (thinking about moral situations and actions), prosocial behavior (engaging in moral behaviors such as helping and sharing), moral emotions (such as feelings of guilt following a moral transgression), and development of conscience (a personal sense of right and wrong). Research indicates that allowing children ample opportunities for prosocial behavior, combined with the use of inductive discipline (which focuses on teaching children about the consequences of their behavior for self and others), is effective in promoting moral development.

Family Relationships

Relationships with parents and (for those children who have them) siblings are a major influence on children’s development. Research indicates that the best outcomes occur when parents are warm and supportive toward their children while also setting clear limits for acceptable behavior. Parents should set clear rules, explain the reasons for these rules, and enforce appropriate consequences when rules are not followed. Optimal parenting also takes into account the child’s developmental level and individual characteristics (e.g., a reserved child may need to be encouraged to take more risks, whereas a more adventurous child may need to be discouraged from doing so).

Peer Relationships

Friendships are important throughout childhood. Interactions with friends provide an opportunity to build social and cognitive skills, and positive relationships with friends can provide a buffer against negative experiences (such as bullying). As children grow older, friends become an increasingly important source of emotional support.

Conflicts among peers are relatively common. These conflicts can be learning experiences, encouraging children to think about events from another person’s perspective. With increasing age, most children become less likely to resort to physical aggression during conflicts with peers, due to increasing ability to take others’ perspectives and to use alternative strategies (such as verbal negotiation) to resolve conflicts. For children whose levels of aggression remain high, there is an increased risk of rejection by peers due to this aggressive behavior.

Risk and Resilience

There are a variety of risks or threats to optimal development that children may experience. These include physical or mental illness, exposure to environmental toxins, experience with trauma (such as physical or sexual abuse), and family conflict. Children whose families lack economic resources are more vulnerable to a variety of threats, including poorer health, unsafe living conditions, and exposure to trauma.

Although exposure to these risk factors is generally associated with poorer developmental outcomes, many children express resilience or an ability to thrive despite negative experiences or difficult life circumstances. Factors that promote resilience include intelligence, high self-esteem, strong self-regulation abilities, and warm, supportive relationships with family members, peers, and teachers.

Meagan M. Patterson

See also Adolescence; Cognitive Development, Theory of; Erikson’s Stages of Psychosocial Development; Kindergarten; Puberty; Resilience

Further Readings

American Academy of Pediatrics. (2014). Caring for your baby and young child: Birth to age 5 (6th ed.). New York, NY: Bantam Books.

Berk, L. E. (2012). Child development (9th ed.). Upper Saddle River, NJ: Pearson.

Bjorklund, D. F. (1997). The role of immaturity in human development. Psychological Bulletin, 122, 153–169.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. Annual Review of Psychology, 53, 371–399.

Hunt, M. (2007). The developmentalists. In The story of psychology (Rev. ed., pp. 401–458). New York, NY: Random House.

Lancy, D. F. (2015). The anthropology of childhood: Cherubs, chattel, changelings (2nd ed.). New York, NY: Cambridge University Press.

Masten, A. S., & Coatsworth, J. D. (1998). The development of competence in favorable and unfavorable environments: Lessons from research on successful children. American Psychologist, 53, 205–220.

Brian S. Gordon Brian S. Gordon Gordon, Brian S.

ChiSquare Test

Chisquare test

268

271

ChiSquare Test

The chisquare test refers to a family of statistical tests that have been utilized to determine whether the observed (sampling) distribution or outcome differs significantly from an a priori or theoretically anticipated outcome or distribution. More simply stated, the test is formulated to determine whether the difference observed was due to a chance occurrence. This entry further describes the chisquare test and looks at its basic principles, applications, and limitations.

Although the most common chisquare test statistic is Pearson’s chisquare test, there are other test statistics that exist with the same theoretical foundation including Yates’s chisquare test, Tukey’s test of additivity, Cochran–Mantel–Haenszel test, and likelihood ratio tests. Although the chisquare test has been applied to a plethora of statistical applications, the fundamental utilization has been as a goodness-of-fit statistic and difference test by comparing a hypothesized distribution to an observed distribution.

In 1900, Karl Pearson developed the chisquare test and published his work entitled “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” where he outlined the limitations of related measures and the functional utility of the test he developed. While the idea of determining whether standard distributions gave acceptable fits to data sets was well established early in Pearson’s career, detailed in his 1900 paper, he was determined to derive a test procedure that further advanced the problem of goodness of fit. As a result, the formulation of the chisquare statistic stands as one of the greatest statistical achievements of the 20th century.

Basic Principles and Applications

Generally speaking, a chisquare test (also commonly referred to as χ^2) refers to a bevy of statistical hypothesis tests where the objective is to compare a sample distribution to a theorized distribution to confirm (or refute) a null hypothesis. Two important conditions that must exist for the chisquare test are independence and sample size or distribution. For independence, each case that contributes to the overall count or data set must be independent of all other cases that make up the overall count. Second, each particular scenario must have a specified number of cases within the data set to perform the analysis. The literature points to a number of arbitrary cutoffs for the overall sample size.

The chisquare test has most often been utilized in two types of comparison situations: a test of goodness of fit or a test of independence. One of the most common uses of the chisquare test is to determine whether a frequency data set can be adequately represented by a specified distribution function. More clearly, a chisquare test is appropriate when you are trying to determine whether sample data are consistent with a hypothesized distribution. The test includes the following procedures: Compute the chisquare statistic, determine the degrees of freedom, select the desired confidence level or *p* value, compare the chisquare value to the critical value in a chisquare distribution table, and decide to accept or reject the null hypothesis on the basis that the observed distribution differs from the theoretical distribution based upon whether the chisquare value exceeds or is less than the critical value.

The chisquare test is also commonly used as a test of independence. By this, it is meant that two variables are compared in a contingency table to determine whether they are related. More generally speaking, the test of independence compares the distribution of categorical variables to see the degree to which they differ from one another. If the two distributions are identical, the chisquare statistic is 0. However, if there is a significant difference between distributions, the result will be a much higher number.

The basic formula for the chisquare statistic used in a chisquare test is as follows:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

In this particular formula, the subscript “*c*” denotes the degrees of freedom, whereas “*O*” is the observed value and “*E*” is the expected value. The summation symbol indicates that this calculation needs to be performed for every data point in your data set. The chisquare statistic is a single number that tells you how much difference exists between the observed and expected frequencies. The chisquare statistic can only be used on numbers and is not applicable to percentages, proportions, or medians, for example.

One of the most common ways to utilize the chisquare statistic is in hypothesis testing and involves the utilization of a chisquare table. After you have calculated the chisquare statistic and have the degrees of freedom, you can consult a chisquare distribution table to determine your *p* value, which will inform your decision to accept or reject the null hypothesis. Generally speaking, the smaller the *p* value, the higher the likelihood of rejecting the null hypothesis, as the likelihood for Type I error is minimal.

To better understand the utility of the chisquare test, a common example utilized is to examine the rolling patterns of two dice in a casino game. For any dice game, the odds that a specific number would come up when thrown should be 1 in 6. However, if we are looking to prove that the player is cheating and may be using “loaded” dice, we can use a chisquare test to determine whether certain numbers coming up more than others represents a meaningful difference or if it is by mere chance. The first step would involve establishing a hypothesis for how you believe the dice should act after a number of throws.

In the casino game example, if there are 60 throws, you would expect each number (1–6) to come up 10 times. Therefore, we could create a table that contains our expected frequency of each number to come up versus what we observe. Let’s say over 60 throws of the dice, the following values come up: 1 = 5, 2 = 16, 3 = 17, 4 = 6, 5 = 7, and 6 = 9. It is important to remember that the expected number for each value is 10. Based on the frequency of certain numbers appearing more than others, it would lend credence to the idea that the die might be loaded to produce more 2s and 3s than the rest of the numbers. However, this could occur by chance as well.

A chisquare test can be used to determine whether the difference between the observed and the expected is meaningful and significant. By using the expected and the observed numbers as well as the number of observations, we can utilize the formula that was discussed earlier to derive our chisquare statistic. Once we have our chisquare statistic, we can consult the chisquare distribution table to determine our *p* value (ensuring also that we know the degrees of freedom). The significance level (remember, the smaller the *p* value, the more likely that we will reject the null hypothesis) provides guidance as to whether we want to reject or accept the null hypothesis (also informing us of the likelihood of committing a Type I or Type II error). The chisquare test represents a family of statistical tools and has been used to develop other measures that can be utilized with other statistical tests. In this section, two such statistical methods are described: analysis of variance and structural equation modeling (SEM). In regression analysis, Tukey's test of additivity is a common statistical test that utilizes the principles of the chisquare test. Specifically, Tukey's test is commonly applied to a two-way analysis of variance. The overarching goal of Tukey's test is to test for interaction when a variable is added to the overall factorial model (typically referred to as an "additive" model). More clearly stated, if a researcher believes interaction (the effect of one variable differs depending on the level of another variable) is an issue, Tukey's test can be used to test the level of interaction effect with the variable added to the factor model.

In the realm of SEM, a chisquare statistic is a goodness-of-fit test and a common statistical tool used to measure the difference between observed and estimated covariance matrices. It is the only goodness-of-fit measure used for SEM that has a direct significance level attached to its testing and forms the basis of many other goodness-of-fit measures. Unlike how the chisquare test is interpreted in other statistical disciplines, for SEM, we actually desire a low chisquare value and in turn, the larger the *p* value, the better. The reason behind this is that the null hypothesis for SEM model testing is that the estimated covariance matrix and the observed matrix are equal. As a result, the smaller the difference between the estimated and observed covariance matrix, the better the hypothesized model fits the data.

Limitations

Although the chisquare test has formed the basis for numerous other statistical tests and represents a leap forward in the realm of statistics, its use comes with a number of limitations. First, it has been shown that a number of studies have incorrectly applied the chisquare test in a variety of research contexts. The most common sources of error were (a) the lack of independence among the measures/variables tested, (b) theoretical frequencies that are too small, (c) use of nonfrequency data, (d) incorrect determination of the number of degrees of freedom, and (e) failure to equalize the sum of observed frequencies and the sum of the theorized frequencies.

In the realm of SEM, the utilization of the chisquare statistic as the sole goodness-of-fit measure is problematic as well. The chisquare statistic is highly susceptible to being influenced by the overall sample size due to the fact that it is a mathematical function of *N* as well as the difference between the observed and estimated covariance matrix. As the sample size increases, so does the chisquare statistic, even if the differences in the matrices remain constant. Second, the number of observed variables also can influence it in that the more observed variables that are present in the SEM model, the higher the chisquare value will be. Although it is the only goodness-of-fit measure with a significance level attached to it, researchers must understand how this statistic is susceptible to the sample size and the number of observed variables in the model.

Brian S. Gordon

See also Analysis of Variance; Goodness-of-Fit Tests; Hypothesis Testing; Interaction; *p* Value; Sample Size; Structural Equation Modeling; Two-Way ChiSquare; Type I Error; Type II Error

Further Readings

Chernoff, H., & Lehmann, E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. The Annals of Mathematical Statistics, 25(3), 579–586.

Greenwood, P. E., & Nikulin, M. S. (1996). A guide to chisquared testing. New York, NY: Wiley.

Lewis, D., & Burke, C. J. (1949). The use and misuse of the chisquare test. Psychological Bulletin, 46(6), 433–489.

Lorga, S., Lubin, L., & Parigi, P. (2003, April). About the chisquare test. Retrieved from http://ccnmtl.columbia.edu/projects/qmss/the_chisquare_test/about_the_chisqu

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(5), 157–175.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). New York, NY: Wadsworth.

Plackett, R. L. (1983). Karl Pearson and the chisquared test. International Statistical Review, 51(1), 59–72.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chisquare test statistic for moment structural analysis. Psychometrika, 66(4), 507–514.

Tukey, J. (1949). One degree of freedom for non-additivity. Biometrics, 5(3), 232–242.

John M. Hintze John M. Hintze Hintze, John M.

CIPP Evaluation Model

CIPP evaluation model

271

275

CIPP Evaluation Model

The CIPP model of evaluation developed by Daniel Stufflebeam is a decision-oriented evaluation approach designed to help those in charge of administering programs to make sound decisions. Designed as a multifaceted approach evaluation, the CIPP model provides a comprehensive framework for conducting both formative and summative evaluations of programs, projects, personnel, products, and organizations by focusing on context, input, process, and product.

Fundamental to the CIPP model is the belief that the most important purpose of evaluation is not to *prove* but rather to *improve*. In this manner, evaluation is viewed as a functional activity where the primary purpose is to strengthen and improve programs and provide a recursive approach to continuous program improvement. That is, evaluation is not seen as a time-limited activity but rather as an ongoing critical component of the programmatic enterprise. This entry describes the evaluation approaches that fall within the CIPP model and discusses how CIPP approaches can be part of the efforts to promote continuous quality improvement and enhancement.

CIPP Categories and Procedures

Consistent with an improvement focus, the CIPP model places a premium on guiding planning and implementation of development efforts. In doing so, the model’s intent is to supply evaluators with timely and useful information for stakeholders, so that they may identify appropriate areas of development, form sound goals and activity plans, strengthen existing programs and services, determine whether and when goals and activity plans need to be altered, and develop plans for the dissemination of effective practices. The utility of the model is judged relative to the relevance, importance, timeliness, clarity, and credibility of findings rather than through common technical adequacy criteria as often found in requirements for internal and external validity.

The model itself comprises four different evaluation approaches that are designed to assist managers and administrators in responding to differing informational and decision-making needs. Although it is not required that each evaluation use each of the four techniques, programmatic enterprises could certainly make use of various components of all four in parallel as an integrated method of continuous program improvement. What follows is a discussion of each of the four separate subevaluation approaches relative to their objectives, methods, and uses. Brief descriptions of certain techniques that evaluators might find useful for conducting each type of evaluation are included.

Context Evaluation

The primary purposes of a context evaluation are to assess needs, problems, assets, and opportunities within a defined environment. *Needs* include those things that are necessary for an organization to fulfill its *defensible purpose*. *Problems* represent determents or impediments that must be overcome in order for the organization to successfully meet targeted needs. *Assets* include resources and expertise that are available and accessible that can be used to help fulfill the targeted purpose of the program. *Opportunities* represent resources and expertise that could possibly be used to support the efforts of the program in meeting its targeted needs and solving associated problems. *Defensible purposes* define what is to be achieved related to the organization’s stated mission while adhering to ethical and legal standards. A context evaluation’s main objectives are to:

set parameters and describe the setting for the intended service,

identify potential recipients, beneficiaries, and stakeholders of the intended service and assess their needs,

identify possible problems and barriers to meeting assessed needs,

identify relevant, accessible assets, and funding opportunities that could possibly be used to address the targeted needs,

provide a rationale and program theory for the program and develop improvement-oriented goals, and

establish a basis for judging outcomes and the worth or merit of targeted improvement and service efforts.

Context evaluations can occur any time during the delivery of a project, program, or intervention. The methodology of a context evaluation usually involves collecting a variety of types of information about members of the target population and their environment, with the goal of thoroughly understanding the context in which a program might be instituted. Information gathering techniques such as semi-and structured interviews, document reviews, demographic and performance data, hearings and community forums, and assessments are all examples of the wide range of activities that evaluators may use to assess the need, worth, and significance of a possible program or intervention.

Input Evaluation

An input evaluation’s main objective is to develop a program or intervention that attends to the needs determined during the context evaluation. Included here is a critical examination of potentially relevant approaches that have been or are currently in use. The preliminary results of the input evaluation—what programmatic approaches were chosen and over what alternatives—are shared with primary stakeholders so that they may collaboratively share in the decision-making process. In doing so, a major purpose of the input evaluation is to identify and rate potential programmatic and intervention approaches and to assist decision makers in the deliberate examination of alternate strategies to address their targeted needs. In reviewing the state of practice in meeting identified needs and objectives, evaluators may use the following strategies:

review of the relevant empirical literature and assess the program’s strategy for responsiveness to assessed needs and feasibility;
identify and investigate existing programs that could serve as a model for the contemplated program;
assess the program’s strategy against pertinent research and extant literature base;
consultation with experts;
querying of pertinent information sources (e.g., those on the World Wide Web);
review of informational reports and available products and services;
assess the program’s plan for sufficiency, feasibility, and political viability;
compile a draft input evaluation report and distribute to client and agreed upon stakeholders; and

upon stakeholders; and

discuss input evaluation findings with client and agreed upon stakeholders in a feedback workshop or strategy session.

The overall intent of the input evaluation is to assist the client and stakeholders in program planning efforts and to use input evaluation findings to devise a program strategy that is scientifically, economically, socially, politically, and technologically defensible. In using the results of input evaluations, clients and stakeholders must assure that the proposed program strategy is feasible for meeting the assessed needs of the targeted beneficiaries, can be supported within the budgeted amount for programmatic change, and is accountable with respect to the rationale for choosing the selected program strategy and has a defensible operational plan.

Process Evaluation

The process evaluation is a formative, ongoing assessment of the program plan or intervention's implementation as prescribed in the input evaluation and documentation of the fidelity and integrity of the program's specified procedures. A thorough process evaluation is facilitated by assigning an evaluation team member to monitor, observe, and provide a record of program implementation. Here, the main objective is to provide program staff and managers with formative feedback regarding the extent to which programmatic efforts are being carried out as planned in a timely and efficient manner. Information gathered during the process evaluation may be used formatively to alter program plans in cases where initial decisions were unsound or not feasible. Moreover, another objective is to ensure that program staff members are delivering the program as intended and that they accept and can carry out their programmatic roles. Process evaluation may include activities such as:

choosing data collection sources/instruments, a schedule for data collection, and rules for collecting and processing information, including procedures for keeping it secure;

reporting specifications and schedule, including interim reports, and rules for communicating and disseminating findings;

in collaboration with program staff, maintaining a record of program events, problems, costs, and allocations;

periodically interviewing beneficiaries, program leaders, and staff to obtain their assessments of the program's progress;

periodically drafting written reports on process evaluation findings and periodically drafting written reports on process evaluation findings and providing the draft reports to the client and agreed upon stakeholders;

presenting and discussing process evaluation findings in feedback workshops to program leaders and staff; and

presenting a final process evaluation report (often incorporated into a larger evaluation report), so that clients and stakeholders can judge the effectiveness of a program relative to process efforts.

In turn, clients and stakeholders use the results of process evaluations to coordinate and strengthen staff activities and program design, to maintain a record of the program’s progress and costs, and to report on the program’s progress to interested parties such as financial sponsors, executive boards and committees, and other program developers.

Product Evaluation

The primary purpose of a product evaluation is to collect and interpret information and judge the worth or merit of a program relative to information gathered during the context, input, and process evaluations. Here, the main objective is to determine whether program outcomes met the preestablished criteria as defined by stakeholders and beneficiaries. Importantly, product evaluation decisions are made in light of both intended and unintended outcomes that influence both positive and negative outcomes. Product evaluations can be conducted using any one of a number or combination of evaluation approaches including objective oriented (e.g., evaluation of program theory, logic models), management oriented (e.g., Provus’ discrepancy model, utilization-focused evaluation), or participant oriented (e.g., naturalistic evaluation, responsive evaluation, goal-free evaluation, empowerment evaluation).

By using a multifaceted approach, product evaluators can decide whether a given program, project, service, or other enterprise is worth continuing, repeating, or extending to other settings. Results should also provide information that evaluators find useful for modifying the program or replacing it so that the institution will more cost-effectively serve the needs of all stakeholders of a target audience and serve as an essential component of accountability in reporting.

As the CIPP model of evaluation has grown and adapted, product evaluations have been conceptualized and subdivided into four different types of product evaluation depending on the informational needs of program developers and evaluation depending on the informational needs of program developers and evaluators.

Impact evaluations

assess the extent to which a program demonstrated intended effects and served individuals and groups in a manner consistent with the program’s intended beneficiaries. Moreover, results of impact evaluations allow stakeholders to make decisions regarding the extent to which the program reached and served the appropriate beneficiaries or community needs and whether the program’s success was consistent with the program’s intended purpose. That is, the resultant findings are consistent with program theory or they may be plausibly explained by other situational features. Consideration of such consistency is important if stakeholders and evaluators desire to consider other forms of product evaluation.

Effectiveness evaluations

attempt to assess the quality and significance of program outcomes and whether they are consistent with designed program theory. Using a variety of methods as noted earlier, effectiveness evaluations attempt to ascertain the significance of the program’s effects—positive and negative, and intended and unintended—and to obtain information on the nature, cost, and success of similar programs to make bottom-line assessments of a program’s significance.

Sustainability evaluations

attempt to assess the extent to which a program’s contributions are institutionalized and successfully continued over time. Here, evaluators strive to understand and assess the program’s provisions for continuation or lack thereof. Results of sustainability evaluations help decision makers understand whether stakeholders and beneficiaries favor program continuation and considerations for adaptation. In addition, findings from sustainability evaluations allow for decisions to be made regarding the continuing need or demand for continued program services. Lastly, sustainability evaluations assist program managers in setting long-term program goals and assigning authority and responsibility for program continuation.

Transportability evaluations

assess the extent to which a program could successfully be adapted and applied elsewhere. By providing a description of the program and a summary of evaluation findings and quality, evaluators can assist other potential program adopters in judging the program's relevance to their situation and the likelihood of similar results and replicability across differing contexts.

CIPP as a Systems Approach to Continuous Quality Improvement

As a recursive model of program evaluation, CIPP approaches can be integrated into a system-wide approach to continuous quality improvement and enhancement. Within this view, program evaluation appropriately promotes the ongoing examination of program goals in a formative manner rather than a series of oblique one-off investigations. CIPP evaluation treats evaluation as a tool by which evaluators in concert with stakeholders can assist programs in promoting growth for their beneficiaries.

Applied correctly, CIPP evaluation represents a sustained, ongoing effort to help organizations organize and use information systematically to meet the needs and goals of a target audience in a respectful and dignified manner. At its core, the model is designed to promote growth and improvement. When used appropriately, the CIPP model of evaluation presents a sustained, ongoing effort to help an organization's leaders and decision makers organize and use information about the organization in a systematic manner to validate goals, meet the needs of stakeholders and target recipients, and provide an accountable approach to program delivery.

One of many approaches to program evaluation research and inquiry, the CIPP model views evaluation as a flexible approach to continuous program improvement and accountability. Notably, the approach is adaptable in the sense that it is responsive to the realities of applied social science endeavors in a manner not typically found with controlled, randomized experiments which are often narrowly focused on particular aspects of a program. While the goal of such latter approaches is often to "*prove*," the goal of CIPP evaluations is to "*improve*." As such, CIPP models of evaluation are designed specifically to allow users to conduct comprehensive and systematic evaluations of the programs that are delivered in real-world, organic settings, not the highly controlled conditions typically seen in experimental psychology where the goal is to minimize the influence of any extraneous variables. The CIPP model embraces such extraneous variables, as they likely represent the real-world conditions under which programs occur.

Lastly, at the larger level, the CIPP model views evaluation as an essential component of societal progress that can be used as an important piece of promoting the well-being of individuals and groups. The core of this belief rests on the contention that societal groups cannot make their programs or services better unless they critically examine both the strengths and weaknesses of what they are delivering. By carefully and continually examining and validating goals and assessing needs, program developers and service providers can plan effectively and invest their time and resources wisely in a manner that affects society at large.

John M. Hintze

See also Evaluation, History of; External Evaluation; Inputs; Logic Models; Program Evaluation; Program Theory of Change; Utilization-Focused Evaluation

Further Readings

Stufflebeam, D. L. (2001). Evaluation models. New Directions for Evaluation (89).

Stufflebeam, D. L. (2007). CIPP evaluation model checklist (2nd ed.). Retrieved from https://www.wmich.edu/sites/default/files/attachments/u350/2014/cippchecklis

Stufflebeam, D. L., & Shinkfield, A. J. (2014). Evaluation theory, models, and applications (2nd ed.). San Francisco, CA: Jossey-Bass.

Michael E. Dawson Michael E. Dawson Dawson, Michael E.

Anne M. Schell Anne M. Schell Schell, Anne M.

Classical Conditioning Classical conditioning

275

277

Classical Conditioning

Classical conditioning is a simple associative learning process first systematically investigated by Ivan Pavlov (1849–1936). Pavlov was a Russian physiologist who studied the digestive processes in dogs and who incidentally noticed that dogs salivated not only to the presentation of food but also upon hearing the footsteps of the research assistant bringing the food. In follow-up laboratory studies of conditioning, Pavlov and his associates presented a neutral sound of a beating metronome (the conditioned stimulus [CS]) followed by the presentation of food (the unconditioned stimulus [UCS]), which elicited salivation (the unconditioned response [UCR]). After several CS-UCS pairings, the sound of the metronome began to elicit salivation (the conditioned responses [CR]), which it had never done before. The dog was classically conditioned to salivate to the metronome.

Since the time of Pavlov, classical conditioning has been extensively studied in a variety of lower animals as well as in humans. Of particular interest has been conditioning of emotional responses. This entry first looks at how classical conditioning has been studied in humans before discussing the studies of classical conditioning in animals, including research on the brain systems involved in conditioning.

Classical Conditioning in Humans

Emotions such as fear can be readily classically conditioned in both humans and lower animals. A common classical conditioning procedure with humans involves pairing a neutral CS such as a mild tone with an aversive UCS (e.g., moderately intense electric shock). The CR is usually measured by changes in autonomic responses such as heart rate and skin conductance following the CS. A conditioned emotional response can be established sometimes with only one pairing of the CS with the UCS. If the CS is subsequently presented a number of times without the UCS, the CR will gradually decline and completely vanish, a phenomenon known as extinction. However, in the absence of any extinction procedure, a CR may remain over time, even if the person realizes that the response no longer seems rational.

In order to ensure that the autonomic changes are truly conditioned to the CS, rather than general sensitization to all stimuli, it is common to randomly intermix presentation of two different CSs. Of the two CSs, only one, the CS+, is paired with the UCS, whereas the control CS, the CS−, is explicitly not paired with the UCS. Alternatively, the control CS can be randomly associated with the UCS in a separate control group. Evidence of successful conditioning is indicated by greater responding to the CS+ than the control CS.

Many individual differences in aversive conditioning have been reported, some of which are associated with the forms of psychopathology (mental illness). For example, a number of studies have found that psychopaths show impaired classical conditioning with aversive UCSs. This is particularly true for psychopaths with callousness and emotional detachment rather than simply antisocial behavior. Yu Gao and colleagues found that poor autonomic fear conditioning at age 3 was associated with criminal behavior at age 23. All in all, these results are consistent with the hypothesis that low fear of socializing punishments is associated with psychopathy. Classical conditioning is also thought to be the basis of many phobias, in which some originally neutral object (a dog, for instance, the CS) is paired with intense fear and pain (being bitten, the UCS), leading to fear of dogs (the CR). Ease of acquiring conditioned fear responses also may be related to anxiety disorders.

Another type of classical conditioning involves presenting a puff of air to the eye as the UCS and eyeblink as the response measure. Conditioning of the eyeblink response has been reliably demonstrated in both humans and lower animals. This appears to be a different type of conditioning than autonomic emotional conditioning because it occurs more slowly and requires a shorter interval between the onsets of the CS and UCS.

Other studies of human classical conditioning have employed brain imaging measures to determine the central nervous system mechanisms that mediate such measures to determine the central nervous system mechanisms that mediate such conditioning. Activation of both the amygdala and frontal cortex has often been found to occur during fear conditioning. Studies of amygdala-damaged patients also indicate impaired autonomic aversive classical conditioning and this is consistent with studies of lower animals, as reviewed in the following section.

Classical Conditioning in Lower Animals

Fear conditioning in lower animals has been partly responsible for the renaissance of interest in emotion within neuroscience. Like the human conditioning paradigm, the prototypical paradigm is to present a tone to a rat followed by an electric shock. Following tone–shock pairings, a number of defensive behaviors (e.g., freezing), autonomic responses (heart rate), and endocrine responses (hormone release) will be elicited by the tone. Fear conditioning occurs in a wide variety of species.

A major advantage of studying conditioning in lower animals is that the effects of invasive brain lesions can be used to determine the areas essential for fear conditioning. Joseph LeDoux conducted a series of studies in rats during the 1980s and 1990s to trace the pathways in the brain involved in fear conditioning. LeDoux used lesions and chemical tracers beginning at the point at which the tone CS activates the auditory neocortex, then tracing the pathway backward to find the areas of the brain that the stimulus must reach to produce conditioning. On the basis of these studies, LeDoux and his colleagues found that the higher auditory neocortex was not necessary for fear conditioning. Instead, the critical point involved specific areas in the amygdala. Information can reach the critical amygdala areas via direct pathways from the thalamus (the low road) as well as by pathways from the thalamus through the cortex to the amygdala (the high road), but the critical point for forming the CS-UCS connection was in the amygdala.

Although the neocortex is not needed for simple conditioning with one CS, it may be necessary for more complex types of conditioning. Suppose that two auditory CSs are presented, a CS+ that is associated with the shock UCS and a CS− that is not. In this case, the animal must learn to discriminate the predictive meaning of two stimuli, and there is evidence that the neocortex is necessary for that discrimination learning. Studies of brain activity during discrimination fear conditioning in humans have also shown that in most circumstances, activity in the cortical areas, particularly the frontal lobes, is involved.

Michael E. Dawson and Anne M. Schell

See also Anxiety; Applied Behavior Analysis; Behaviorism; Educational Psychology; Punishment; Reinforcement; School-Wide Positive Behavioral Support

Further Readings

Gao, Y., Raine, A., Venables, P. H., Dawson, M. E., & Mednick, S. A. (2010). Association of poor childhood fear conditioning and adult crime. American Journal of Psychiatry, 167, 56–60.

LeDoux, J. (1996). The emotional brain: The mysterious underpinnings of emotional life. New York, NY: Simon … Schuster.

Pavlov, I. (1927). Conditioned reflexes. London, UK: Oxford.

Rescorla, R. A. (1988). Pavlovian conditioning: It’s not what you think it is. American Psychologist, 43, 151–160.

Clifford E. Hauenstein Clifford E. Hauenstein Hauenstein, Clifford E.

Susan E. Embretson Susan E. Embretson Embretson, Susan E.

Classical Test Theory

Classical test theory

277

283

Classical Test Theory

Classical test theory (CTT) is an approach to measurement that considers the relationship between the expected score (or “true” score) and observed score on any given assessment. The word *classical* is used in the sense that the theory is considered to be the first practical application of mathematics to describe this relationship. CTT offers a relatively parsimonious, elegant, and intuitive way to scale individuals according to some theorized latent construct. This entry further describes CTT and its basic principles and estimation procedures, then discusses its framework for determining a measure’s proportion of true score variance, standard error of measurement, item analysis, and validity. Finally, it looks at the limitations to the theory.

Although more contemporary, model-based approaches to measurement, such as item response theory (IRT), have garnered more focus, CTT retains its relevance and importance for several reasons. First, CTT offers a relatively simple and intuitive analysis of response characteristics for an assessment. Even if the goal is to utilize more contemporary methods of measurement, CTT provides an initial framework of analyses to explore data; its relatively simple approach augments data diagnostic efforts. Second, CTT follows a less rigorous set of assumptions than the more complex IRT approach to measurement. It can be easily be applied to a wide variety of testing situations. Third, CTT requires fewer data demands for scaling procedures. Fourth, CTT extends from a framework of computations that are simpler in nature; variance, covariance, and correlation statistics lay the groundwork for CTT. Thus, almost any statistical software or data management program can be employed for most CTT analyses.

software or data management program can be employed for most CTT analyses.

Basic Principles and Estimation Procedures

CTT was born out of the culmination of two particular advances in the field of measurement: first, the growing recognition of symmetrically distributed random errors in measurement (a concept that dates back to Galileo's masterpiece, *Dialogue on Two Main Systems of the Universe: Ptolemaic and Copernicus*). By the latter half of the 19th century, it was well accepted that experimental observations were jointly impacted by a stable, true score and an error in measurement defined as a random variable.

The advent of a metric to describe the degree of relationship between two variables provided the second groundwork for the CTT approach. Francis Galton derived the correlation statistic in 1886 to indicate the extent to which mean deviations in one variable reflect corresponding mean deviations in another variable. This metric laid the foundation for estimating the impact of random errors on the stability of a test score (reliability analysis).

Each of these motivations (randomly distributed error terms and correlation) was considered together in a landmark paper by Charles Spearman in 1904, in which he recognized that observed correlations between tests would be attenuated as a function of the amount of error measured with each test. By many accounts, this paper set the stage for the development of CTT as a proper measurement paradigm. Frederic Lord and Melvin Novick are credited with organizing the psychometric developments of the time into a cohesive framework in their 1968 book, *Statistical Theories of Mental Test Scores*.

Lord and Novick invoke the notion of randomly distributed error terms to develop the following formula, which forms the crux of CTT:

$$X = T + e.$$

Here, any observed score (*X*) is a result of the joint influence of a stable true score (*T*) and a random error term (*e*). Because the observed score is a function of a random variable (*e*), it itself can be considered a random variable.

To understand the error component in practical terms, it is helpful to distinguish random error from systematic error. In general, error represents the impact of all variables extrinsic to the trait of interest. Systematic error represents influences that bias the observed score in a consistent manner. For example, in a math ability assessment consisting of word problems, more linguistically demanding items may result in lower scores for nonnative speakers. Thus, linguistic ability, a variable extrinsic to the trait of interest, would influence scores in a consistent manner from one test administration to another.

Conversely, random error represents those influences extrinsic to the trait of interest that are not stable from one testing occasion to another. For example, distractions in the test environment, fatigue, and guessing may have differential effects on each test administration.

In the CTT model, true score is defined in purely statistical terms as the expected value of observed scores. Intuitively, the expected value can be thought of as a long run average of a series of observations. Computationally, it is defined as:

$$T_j = E = n = 1nXnpn.$$

Where T_j is the true score for subject *j, X* is the observed score for subject *j, n* corresponds to the particular testing occasion, and *p* corresponds to the probability of observing any particular score. If we assume that the frequency of observed scores is proportional to the probability mass function of random variable *X*, then the calculation simplifies to the arithmetic mean of observed scores. Thus:

$$T_j = EX = n = 1nXnpn = n = 1nXnN.$$

It extends from this definition that the expected value of the error is necessarily zero:

$$e = Xj - Tj$$

$$Ee = EXj - Tj = EXj - ETj.$$

And since *EXj* = *Tj*; then *Ee* = *Tj* − *Tj* = 0.

These definitions correspond to multiple testing occasions for a single subject. Invoking identical assumptions, it can be shown that the average true score for a population of subjects can be estimated from the average of all observed scores in a sample. Similarly, the average error term for a population of subjects can be estimated as the average of error terms for a sample; thus, population error = 0.

We might draw the practical conclusion that we can expect the average error

We might draw the practical conclusion that we can expect the average error over many testing occasions to approach zero. Practically speaking, we can expect random influences to positively impact the observed score just as often as we can expect them to depreciate it. Half the time, random error improves the observed score relative to the true score, and half the time, it decreases the observed score relative to the true score.

From this groundwork, a few additional corollaries can be derived:

1. The correlation between true and error scores in a population is equal to zero (both within and across measures).
2. The correlation between error scores on two separate measures is zero, assuming the observed scores are randomly drawn from independent distributions.
3. The variance of the error term for a group of examinees is taken as the expected value of the within person score variance, over all *n* persons.
4. Because we assume *CovT, e* = 0 (as referred to in no. 1), then necessarily the variance of the observed score in a population is the sum of the variance of true scores and error scores:

$$\begin{aligned}\mathrm{Var}X &= \mathrm{Var}T + \mathrm{Var}e + 2CovT,e \\ &= \mathrm{Var}T + \mathrm{Var}e + 20 \\ &= \mathrm{Var}T + \mathrm{Var}\left(e\right).\end{aligned}$$

At this point, it is important to emphasize a few notions regarding true scores in CTT. First, the true score in CTT is not defined by a particular physical, biological, or substantive indicator. Rather, the true score is defined according to the moment of an observed distribution of scores. Thus, the true score is dependent upon the measurement process itself. Importantly, this means that any consistent bias in measurement (systematic error) cannot be disentangled from the true score. That is, score variation due to systematic influences is absorbed into the true score estimate for persons. Second, neither term in the right side of the equation is directly observable. This means that the error is not a residual term in the traditional sense.

The assumption can be made that the expected value of the observed score is equal to zero, which fully constrains the error term to have an expected value of zero. Conversely, one can assume the error score has an expected value of zero, which sets the true score to the expected value of *X*. Thus, these terms derive wholly from the definitions applied. In this sense, error does not indicate lack of fit in the model, and falsification of CTT cannot be made a consideration.

Proportion of True Score Variance

The notion of an unpredictable error term impacting the observed score invites the question as to how stable a given measure is. In other words, how consistent would observed scores be from one measurement occasion to another with a particular assessment? This is the converse to asking the magnitude of random error influence. Essentially, for a given population, it is beneficial to decompose total score variance into true score variance and random error variance. Through this decomposition, the proportion of true score variance is derived and can be used as a proxy for the stability of a measure. In CTT, this metric is referred to as the reliability of a measure.

$$\text{Reliability} = \sigma T \sigma T + \sigma e = \sigma T \sigma X.$$

Where σ*T* is the variance of true scores in the population, σ*e* is the variance of error scores in the population, and σ*X* is the variance of observed scores in the population.

To arrive at this metric, it is necessary to consider the extent to which the true scores in a population covary with the observed scores. To the extent that the correlation between true score and observed score is greater, the error score variance is depreciated. In the extreme case of perfect collinearity with observed score and true score, the error score variance diminishes to zero.

Writing the true score/error score correlation (note: *X* and *T* are written as mean deviation values):

$$\rho XT = xtN\sigma T\sigma X = t + eDtN\sigma T\sigma X$$

$$= t2N\sigma T\sigma X + teDN\sigma T\sigma X.$$

Because the correlation between true and error scores is zero as defined previously:

$$\begin{aligned}\rho XT &= t2N\sigma T\sigma X + 0 = t2N\sigma T\sigma X \\ &= \sigma T2\sigma T\sigma X = \sigma T\sigma X.\end{aligned}$$

Thus, we can interpret the correlation of true score to error score as the ratio of true score variance to total observed score variance. However, the true score is not directly observable. To circumvent this issue, it can be shown that the correlation between observed scores on parallel test forms is also equal to the proportion of true score variance:

$$\rho X1X2 = x1x2N\sigma X1\sigma X2 = \rho XT = \sigma T\sigma X.$$

However, a heavy constraint is placed upon this relationship. For this relation to hold, parallel test forms must satisfy the following conditions: (a) subjects earn the same true score on both measures and (b) there are equal error variances across the two measures. Therefore, much of the focus of CTT has been to develop strategies to develop parallel test forms. Four essential approaches to this end can be discussed: (1) *Test–retest reliability* involves administering two identical assessments to examinees to ensure parallel forms. The time interval between assessments is selected partially based on the purpose of the test. For example, the authors of an occupational interest survey may be interested in the stability of test scores over a long interval (up to several years) and may space assessments accordingly. The reliability coefficient is calculated as the simple Pearson moment correlation between scores on the two assessments and is referred to as the coefficient of stability.

(2) *Alternate form reliability* involves constructing two different forms of an assessment that are thought to be equivalent in terms of item content and difficulty and administering them to a subject pool. Assessments are administered with as small a time interval as practical. This approach may reduce concerns of practice effects interfering with the reliability estimate. Additionally, this approach may be most appropriate when practical concerns require several distinct versions of an assessment to be administered (to reduce cheating or for test security reasons). The reliability coefficient is calculated as the simple Pearson moment correlation between scores on the two assessments and is referred to as the coefficient of equivalence.

(3) *Test–retest with alternate forms* combines the previous two approaches to arrive at a coefficient of stability and equivalence. Two equivalent forms of an

assessment are administered to a subject pool after a particular time interval.

(4) *Internal consistency* considers parallel forms as being derived from two halves of a single assessment. Because only one assessment is administered, the derived reliability coefficient is not affected by maturational or practice effects. The simplest method to derive an internal consistency coefficient is to split the assessment into two halves and calculate the Pearson correlation coefficient between scores on each half. Several splitting methods exist: odd–even split, random assignment, and content matching. In the odd–even split, odd items comprise one half while even items comprise the second half. In random assignment, items are randomly selected to each half. In content matching, the test is split such that items have matching content across halves.

Regardless of the particular method chosen, it is important to recognize that by increasing the number of items in an assessment, the reliability generally increases. Thus, the split-half methods underestimate the reliability of an assessment because the coefficient is based on correlating only half of the test items. Therefore, a correction is usually applied to offer the expected, improved reliability coefficient for the full-length assessment. Several methods exist, but the procedure developed by both Spearman and Brown (identically, but independently) in 1910 gained the most traction. The final Spearman-Brown prophecy formula is defined as follows:

$$\rho_{\text{composite}} = 2\rho X1X21+\rho X1X2.$$

Where $\rho_{\text{composite}}$ is the expected reliability of the composite, and $\rho X1X2$ is the observed correlation between the two halves. It should be noted that the Spearman-Brown prophecy formula is derived under the assumption of parallel test halves. To the extent that this assumption is not met, the corrected reliability coefficient is still likely to be an underestimate of the true value.

However, concern remained about the lack of a unique internal consistency estimate. That is, estimates were not invariant across different methods of test splitting. Three publications led to similar methods to address this issue. In 1937, G. Frederic Kuder and Marion Richardson developed the iconic KR 20 and KR 21 formulas, which offered procedures for developing a universal internal consistency metric. Later, Lee Cronbach developed what may be the most popular procedure to develop a universal internal consistency value:

$$\alpha = kk-1\left(1-\sigma i2\sigma X2\right).$$

Where *k* is the number of items on the assessment, $\sigma i2$ is the variance of item *i*, and $\sigma X2$ is the observed variance in total test score. This formula can accommodate both dichotomously and polytomously scored items.

Standard Error of Measurement

Extending the notions of true score variance and error score variance, a logical question to ask is how much error variability surrounds any given subject’s true test score. Although this parameter is unknown at an individual level, it can be derived from the estimated true score variance. Conceptually speaking, one could administer a test iteratively to a single subject and obtain a distribution of scores. The mean of the distribution would represent the true score, while the standard deviation of the distribution would serve as an indication of the amount of error in measurement. The expected value of this standard deviation, taken over all subjects in the distribution, is the standard error of measurement. To derive this value, consider the CTT decomposition of observed variance into true score variance and error score variance:

$$\sigma X2 = \sigma T2 + \sigma E2.$$

And, dividing through by observed score variance:

$$1 = \sigma T2 \sigma X2 + \sigma E2 \sigma X2.$$

And because the ratio of true score variance to observed score variance defines the reliability coefficient, $\rho XX1$:

$$1 = \rho XX1 + \sigma E2 \sigma X2.$$

And, rearranging terms, we arrive at an estimate for the error variance around any given true score:

$$\begin{aligned} \sigma e2 &= \sigma X1 - \rho XX1 \\ &= \text{Standard error of measurement.} \end{aligned}$$

With the assumption of normally distributed error scores invoked, this statistic allows for calculation of a confidence interval (CI) around the true score:

$$\mathrm{CI} = T \pm Z\alpha 2 \times \sigma e2.$$

Item Analysis

Although CTT primarily focuses on psychometric properties at the global test level, a framework of item analysis statistics has been developed within the CTT paradigm. The goal of these statistics is ultimately to aid in selecting items that provide the most information regarding examinee performance and maximize reliability. To understand item selection procedures, it is first helpful to recall the variance of any composite score. If *Y* is a composite of n subcomponents, then:

$$\sigma Y2 = i = 1n\sigma i2 + 2\rho ij\sigma i\sigma j.$$

Where *i* < *j*. For dichotomously scored items, it is also true that the variance of a single item is equal to:

$$\sigma i2 = pi\left(1 - pi\right),$$

where *p* is the proportion of respondents answering item *i* correctly (also called item difficulty). Variance of scores for any item is maximized then when *pi* = .5. From this theorem, the following corollaries can be developed in relation to testing:

Selecting items that exhibit high covariance values also maximizes test score variance.

Selecting items with difficulty = .5 maximizes variance in respondent total score.

Thus, these items offer the most information for distinguishing examinees. Increased variance of scores also improves the stability and equivalence reliability coefficients because they are based on the correlation coefficient of scores between parallel forms. S. Henryssen recommends a general range of items with difficulty = .5.

The exception to selecting items with difficulty = .5 is when a specific cut score is to be used to distinguish groups of examinees. In these cases, *items with difficulty of .5 for only those examinees whose total score equals the cut score* should be included.

Item discrimination is another important variable to consider. In general, if an item is written well and relates to the trait of interest, individuals who pass the item should also obtain higher test scores. Conversely, those with a lower probability of answering an item should obtain lower test scores. One method to assess this property is to compute a biserial correlation between a single dichotomously scored item and the total test score for a group of subjects. The higher the observed biserial correlation, the better an item is able to distinguish high-performing subjects from low-performing subjects. Negative item total score correlations indicate a very poorly functioning item that is operating in the reverse (those with higher test scores respond incorrectly). It is important to note here that Cronbach's α metric for internal consistency will be maximized when the biserial correlations for all items are maximized.

Validity

The procedures reviewed earlier that describe the structural properties of a measure are helpful in determining the stability of measures and the extent to which items form a homogenous pool and elicit consistent responses from subjects. Validity concerns the inferences and applied utility of the measure. Validity is the extent to which interpretations and applications of test scores are appropriate. At a very basic level, if reliability informs the assessor regarding how consistent the test is, validity is concerned with what the test *actually* measures. Validity is not a property of the assessment itself but a collection of empirical findings and theoretical justifications pointing toward the suitability of the conclusions drawn from test scores. However, validity is partially a function of reliability; stability in test scores is a necessary, but not sufficient, criterion to establish the validity of test interpretation.

Traditionally, the psychometric community had defined validity in terms of several distinct types or aspects: content validity, criterion validity, and construct validity. Recently, however, there has been a growing trend to conceptualize any assessment as an indicator of a particular construct or domain and that multiple forms of evidence exist to establish convergence between test scores and the construct. The components of validity evidence recognized by the prevailing psychometric communities are content validity, response process evidence, internal structure, relationships to external variables, and consequences of test implementation.

Limitations

The weakness of CTT is threefold: First, CTT generally focuses on test-level statistics. Although an approach to item-level diagnostics exists in the CTT framework (as explicated previously), it is not elaborated as with the more contemporary IRT approaches. In CTT, no underlying model is specified to link specific item stimulus features to item difficulty or to the interaction with an examinee's latent trait in effecting a response outcome. Similarly, CTT generally assumes identically distributed error terms for each item. This assumption precludes a more discrete analysis where standard error of measure scores are estimated for each item separately. Generally speaking, the basis for scaling persons rests on the distance of an individual's true score from the true score of the norming population. Conversely, IRT approaches base person scaling on the location of a latent trait score on the item scale.

Second, CTT scaling is grounded on a circular dependency: Person observed scores are dependent upon the distribution of item statistics on the assessment, and the distribution of item statistics is dependent upon the distribution of observed scores. Thus, person true score estimates are not invariant across different item sets, and item property estimates are not invariant across different person samples. This imparts a particular difficulty in comparing true scores across different assessments. Although test equating techniques exist, considerable error can be introduced in the process. Generally speaking, this precludes the ability for CTT to be applied to adaptive testing procedures, where each examinee receives a different set of items conditional on the examinee's performance pattern.

As briefly discussed previously, another circular dependency exists in distinguishing error scores from true scores. True score and error score jointly define the observed score, but neither true score nor error score is directly observed. Thus, the main terms of CTT are defined by the particular assumptions of the theory. In this purest form of CTT, this precludes falsification of the model; it must be true by its own definition.

Despite these limitations, CTT continues to occupy an important place in the field of educational and psychological measurement. In fact, under certain conditions, CTT can return similar person ability estimates as the more computationally demanding and data-intensive IRT approach. Considered together with the benefits enumerated earlier, it behooves anyone involved in psychological or educational testing to have a basic understanding of the principles of CTT.

Clifford E. Hauenstein and Susan E. Embretson

See also Correlation; Equating; Item Response Theory; Psychometrics; Reliability; Standard Error of Measurement; *Standards for Educational and Psychological Testing*; Validity; Validity Coefficients; Variance

Further Readings

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anastasi, A. (1988). Psychological testing. New York, NY: MacMillan.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rhinehart … Winston.

Embretson, S. (1996). The new rules of measurement. Psychological Assessment, 8(4), 341–349.

Jones, L., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics: Vol. 26 (pp. 1–27). Amsterdam, the Netherlands: North-Holland.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Murphy, K. R., & Davidshofer, C. O. (1988). Psychological testing: Principles and Applications. Englewood Cliffs, NJ: Prentice Hall.

Novick, M. R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3(1), 1–18.

April Galyardt April Galyardt Galyardt, April

Classification

Classification

283

285

Classification

Classification refers to a broad set of statistical methods that arise in many different applications. In a classification problem, we have a categorical response variable that we wish to investigate in relationship to one or more input variables. Classification methods can be applied to problems in a wide variety of settings; applications in education include analyzing patterns of responses to standardized exams, inferring which middle school students will benefit from a drug prevention program, and predicting which graduating high school seniors will choose to attend a particular university if they are offered admission.

Common classification methods include logistic regression, support vector machines, decision trees, random forests, neural networks, and k-nearest neighbors. This entry discusses a few general issues in classification that should be considered when choosing a method and the differences between classification and the related problem of clustering.

General Issues in Classification

Classification problems include both *prediction* and *inference*. In an inference problem, the goal is to describe the relationship between the response variable and the explanatory variables, whereas in a prediction problem, the goal is to predict the value of an unobserved response variable for a new data point based on observed predictor variables. For example, if we wish to examine the relationship between a person’s diet and whether the person later gets cancer, this is an inference problem because the question of which foods put a person at risk is of paramount importance. In contrast, if we wished to classify the content of an image based on features extracted from the digital representation of the image, this is a prediction problem because which features are useful for making the classification are not important.

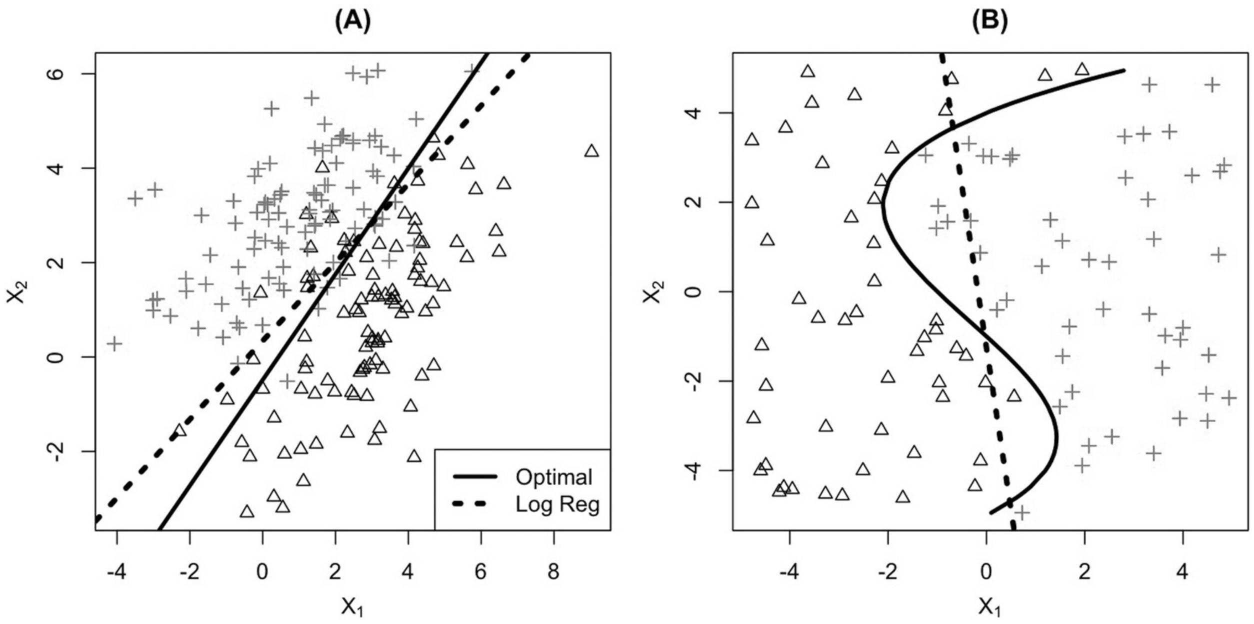
Logistic regression and decision trees are examples of methods that are appropriate for inference because they provide easy to interpret information about the relationship between the response variable and the explanatory variables. Though, as with any statistical methodology, making causal claims based on the results from a classification analysis relies on proper experimental design. K-nearest neighbors, support vector machines, and random forests may provide accurate predictions, but can be more challenging to interpret, and are therefore more appropriate for prediction problems than inference problems.

Any problem with a categorical response variable may be deemed a classification problem, but methods differ based on how many levels the categorical response has. Logistic regression is most often used as a binomial method for a binary response variable; by contrast, multinomial logistic regression, k-nearest neighbors, and linear discriminant analysis can easily handle any number of classes.

Decision Boundaries

Decision boundaries separate the space of input variables into regions labeled according to classification. One of the key elements determining the complexity of a classification problem is the shape of these boundaries. Figure 1 shows two classification problems with two classes (Δ, +) and two predictor variables (*X1, X2*). The solid line shows the Bayes's optimal decision boundary, whereas the dotted line is the decision boundary estimated with logistic regression. Figure 1A shows a case where the Bayes's optimal decision boundary is linear, whereas in Figure 1B, the boundary is nonlinear. If the input variables describe a space best partitioned using a nonlinear decision boundary, it is important to choose a method that can estimate such a boundary, particularly for inference problems.

Figure 1 Two examples of classification problems



Some of the most popular classification methods, including logistic regression, support vector machines, and linear discriminant analysis, will produce boundaries that are linear in the input space; in Figure 1, the dotted lines are the decision boundaries estimated using logistic regression. Other methods, such as k-nearest neighbors, decision trees, and random forests, can find decision boundaries that take more complex shapes.

Nonetheless, even when the optimal decision boundary is nonlinear, linear methods may still have very good *predictive* performance. In Figure 1B, the linear boundary is nonoptimal, but only a small percentage (4%) of points fall on the wrong side of the boundary. In general, this phenomenon is known as the *bias–variance trade-off* and is part of assessing model–data fit.

Relationship Between Classification and Clustering

Clustering is a closely related set of statistical methods. In both classification and clustering problems, we assume that the population consists of subgroups, so that the probability distribution for the population can be expressed as a finite mixture model. The difference is that in classification, we observe the class labels for some or all of the data, whereas in clustering, we do not observe any group labels and must infer both what subgroups exist and which points belong to which group. Classification is part of a larger group of methods often called *supervised learning*, where the response variable is observed, whereas clustering is a subset of *unsupervised learning* methods where the investigator is looking for patterns in data but no specific response variable has been recorded.

In this sense, latent class analysis is a clustering method because the classes are latent and are to be inferred. Similarly, diagnostic classification models are unsupervised models; these models are used to infer which students are in a class possessing a particular set of skills, and the skill profile of each student is unobserved.

April Galyardt

See also Categorical Data Analysis; Cluster Analysis; Data Mining; Diagnostic Classification Models; Latent Class Analysis; Logistic Regression; Model–Data Fit

Further Readings

Agresti, A. (2013). Categorical data analysis (3rd ed.). Hoboken, NJ: Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer-Verlag.

Shmilovici, A. (2005). Support vector machines. In O. Maimon & L. Rokach (Eds.), Data mining and knowledge discovery handbook. (pp. 257–276). doi:10.1007/0–387–25465-X_12

Bruce B. Frey Bruce B. Frey Frey, Bruce B.

Classroom Assessment Classroom assessment

285

288

Classroom Assessment

Classroom assessment refers to student assessments that teachers design and administer themselves. The term usually is distinct from the standardized tests that are given in schools such as intelligence tests or tests used in statewide testing systems. A more formal definition of classroom assessment would be the teacher-directed systematic collection of information about students' learning, traits, and abilities.

Classroom assessment can be used before, during, or after instruction and can be formal, with standardized procedures and predetermined scoring criteria, or informal, consisting of brief observations. This entry first describes the five basic approaches to classroom assessment. It then discusses the purposes of classroom assessment and how assessments are scored and interpreted.

Approaches to Classroom Assessment

There are five basic approaches to classroom assessment. Teachers choose an approach based on different philosophies, different reasons why the data are being collected, and how the data will be used. Different approaches differ in their purpose, in the nature of the data they produce, in their intended audience, and in some cases in the assumptions they make about children and about learning. The five approaches to classroom assessment are traditional paper-and-pencil, performance-based, formative, authentic, and universal design. For organizational purposes, this entry discusses these approaches as independent ways of treating assessment, but there is much overlap among the five approaches.

Paper-and-Pencil Assessment

Paper-and-Pencil Assessment

Paper-and-pencil assessment is a familiar approach featuring items such as multiple-choice questions, matching items, true/false items, and fill-in-the-blank items. These types of tests are typically used for assessing knowledge and basic understanding. In many contexts, paper-and-pencil tests work the best to assess a large number of objectives quickly using a format with which students are familiar. There are traditionally two item formats for paper-and-pencil assessment: selection items and supply items. With selection items, the correct answer is there for students to select (e.g., a multiple-choice item). With supply items, the student must provide the answer, as with a fill-in-the-blank item or an open-ended, short-answer question.

Because paper-and-pencil assessments can be scored objectively without subjective evaluations, there is very little randomness in the scoring. Consequently, paper-and-pencil assessment is the most reliable of classroom assessment approaches. Whether these assessments are the most valid way to assess knowledge or understanding, however, is a different question. Often, approaches that explore understanding or ability more deeply do a better job of tapping into the constructs of interest.

Performance-Based Assessment

Thirty-five years ago, this approach was seen as newfangled, but now is so common Performance-based assessment gained popularity in the 1980s and 1990s and is now so common that many consider it as traditional as a multiple-choice test. The idea is to go beyond the measurement of low-level knowledge and understanding by asking students to perform a skill or create a product and assess student ability then assessing the performance or product. Performance assessment is typically used for assessing skill or ability and is the common approach in the areas of communication (e.g., writing and public speaking), mathematics, science, athletics and physical education, social skills, and the performing arts.

The performance-based assessment framework led by necessity to new scoring options, such as the creation of subjective scoring rubrics or guides that outline what teachers wish to measure, but can lead to scoring difficulties because of the judgments required. Scoring rubrics identify components or criteria of quality for an assignment or assessment and provide a range of scores for each piece.

Often, rubrics provide descriptors for what each score along the range means. Assignments and classroom activities that are often scored with rubrics include:

group projects, in which students work together on a collaborative problem and can be assessed on their discussions or group presentations;

writing assignments, which require written description, analysis, explanation, or summary;

scientific experiments, which allow for observation of how well students can conduct scientific investigations;

demonstrations that students perform, showing their mastery of content or procedures; and

portfolios or collections of students' work. Typically, portfolios are used to evaluate students' development over time.

Formative Assessment

Data collection that occurs during instruction not only can guide the teacher on instructional effectiveness but also can let students know where they are and how they are doing. Formative assessment is typically used to give feedback to students and teachers about how things are going and does not affect grades. More importantly, providing frequent feedback directly to students so they can monitor and control their own learning is just about the only assessment approach that has been found to directly affect learning (and, of particular importance to administrators, to increase standardized test scores). Students who learn in a formative assessment environment become self-directed learners. Self-directed learners are self-managing (they make use of their own experiences), self-monitoring (they use metacognitive strategies), and self-modifying (they alter their approach to learning).

There are many ways that teachers use formative assessment in the classroom, including the following:

frequent quizzes or tests that do not affect grades but merely give feedback to students and teachers;

conferences in which work plans and strategies are discussed;

performance control charts on which students have a numeric criterion for success on an assignment; and

self-reflection work sheets to identify areas of difficulty or strength.

Authentic Assessment

A best practice in the modern classroom is to utilize assessment tasks that match the real-world expectations. Authentic assessment typically requires students to perform in ways that are valued outside the classroom. This approach may increase the meaningfulness of classroom assessment across all ages, preschool through graduate school. Understanding this approach to assessment has difficulties, though, because there is disagreement over what it means to say that an assessment is *authentic*.

The idea that assessment tasks should be intrinsically meaningful and motivating and require skills or knowledge that is valued in the world is a powerful one, though, and can produce powerful assessments. Authenticity in assessment has different meanings across different content areas and at different student ages. Common across these areas, though, are several dimensions that make classroom assessment authentic. Assessment is authentic when the context is realistic and cognitively complex, when students collaborate with each other and use feedback formatively, and when the scoring is interpreted under an expectation of mastery with multiple indicators combined.

Universal Test Design

Universal test design emphasizes accessibility and fairness for all children, regardless of gender, first language, ethnicity, or disability. Basic standards exist that can and should be applied to classroom assessment in all contexts and at all levels. Application of these standards can ensure that testing, whether teacher developed or state mandated, is inclusive and measures what it is supposed to.

Assessments following universal design principles are typically used when the classroom teacher is concerned that irrelevant student characteristics might affect performance. General guidelines for knowing whether the content of an assessment follows universal design principles and allows “access” to all students include making sure that all students would likely have the experiences and prior knowledge necessary to understand the question and that the vocabulary, sentence complexity, and required reasoning ability are appropriate for all students’ developmental levels.

The wording used in assessments can affect accessibility. Sentences should be short, use the jargon of the field and instead of uncommon words, more common short, use the jargon of the field and instead of uncommon words, more common synonyms should be used. Finally, teachers should establish a consistency in format and style across assessments and within each assessment.

Choosing and Combining the Approaches

In the past, nearly all classroom assessment was traditional paper-and-pencil assessment with, especially in the last few decades, some performance-based assessment activity. More recently, the field has begun to recognize the value of modern approaches such as formative assessment, authentic assessment, and universal design of assessment. This is clear from a review of the scholarly literature on classroom assessment. Hundreds of studies have been published examining the effectiveness and usefulness of formative assessment, authentic assessment, and applications of universal design. The traditional multiple-choice test and performance assessment still predominate in the classroom, but these three modern approaches are now part of the conversation around classroom assessment.

The five approaches emphasized in this entry are derived from different theoretical frameworks, emphasizing different purposes of assessment, but they are not mutually exclusive. A teacher does not have to pick one theory or approach over another but can focus on the purpose for a particular assessment and choose different approaches that are consistent with that goal. A teacher can design a traditional paper-and-pencil test that follows universal accessibility guidelines. A single classroom assessment might be performance based, authentic, and formative. Most authentic assessments are probably performance based, but many performance-based assessments are not authentic. A formative assessment might inform both the teacher and the student and parents.

In general, teachers choose an assessment strategy using paper-and-pencil formats when they wish to measure basic knowledge that can be memorized and use performance-based formats when they wish to measure skill, ability, or deeper understanding of concepts. Formative assessment is used to give students and teachers feedback while learning is still happening. Authentic assessment following universal design principles is chosen when teachers wish to evaluate students using the real-world tactics that work equally well for all students.

Purposes of Classroom Assessment

Teachers can choose to use assessment for several different reasons, all of them important. Sometimes a combination of reasons is in play:

> Assessment *for* learning: Teachers gather information about where students are “at” (what they know and can do) and how they are reacting to instruction. The purpose is to design and revise instruction, so that it is the most effective.
> Assessment *as* learning: Data are gathered either *by* or *for* students to help them understand how they learn. This is the formative assessment approach. The purpose is that students develop learning skills and control their own learning.
> Assessment *of* learning: Data are gathered to reach a conclusion about how much students have learned after instruction is done. Until recently, this was the only use of classroom assessment. The purpose is to share with the students and others how much they have achieved.

Scoring and Interpreting Classroom Assessment

The five approaches to classroom assessment consist of a set of three different formats or types of assessment designs (traditional, performance based, and authentic), a relatively new way of thinking about the purpose of assessment (formative), and an overall philosophy about the usefulness of a given assessment (universal test design) for all students. Whatever the format, classroom assessments can be scored and interpreted in two ways—norm-referenced and criterion-referenced.

Norm-referenced scoring means that performance is interpreted by comparing scores to each other; the information in a score comes from referencing what is normal. Criterion-referenced scoring applies external criteria that have nothing to do with how the average person performed. Common criteria for classroom assessment score interpretation using this framework are grading scales that assign scores above 90% correct as an A, for example, or teachers concluding that instructional objectives have been met by applying a standard of mastery.

Classroom teachers use both criterion-referenced score interpretation and norm-referenced interpretations all the time. If everyone can get an A on a test or on a report card by meeting some set of standards, objectives, or criteria, then criterion-referenced interpretation is at work. If the grading is “on a curve” or individual scores have meaning only in comparison to how others performed, then norm-referenced interpretation is being used.

Norm-referenced evaluations are so common in education that one may not even realize that someone (a teacher, policy maker, administrator, and test developer) has chosen that approach over criterion-referenced interpretation. Criterion-referenced assessments are also common and just as much a part of the classroom culture as norm-referenced. The meaning of an assessment differs depending on which interpretation is applied.

Bruce B. Frey

Adapted from Frey, B. B. (2014). *Modern classroom assessment*. Thousand Oaks, CA: Sage.

See also Authentic Assessment; Formative Assessment; Paper-and-Pencil Assessment; Performance-Based Assessment; Universal Design of Assessment

Further Readings

Black, P. J., Buoncristiani, P., & Wiliam, D. (2014). Inside the black box: Raising standards through classroom assessment. Cheltenham, Australia: Hawker Brownlow Education.

Frey, B. B. (2014). Modern classroom assessment. Thousand Oaks, CA: Sage.

Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. Journal of Advanced Academics, 18(3), 402–423.

Mertler, C. A. (2016). Classroom assessment: A practical guide for educators. Abingdon, UK: Routledge.

Drew Gitomer Drew Gitomer Gitomer, Drew

Classroom Observations Classroom observations

289

293

Classroom Observations

Classroom observation represents a measurement approach used to characterize teaching quality through the use of an observation protocol. This entry briefly reviews the history of how classroom observations have been used and then considers research that examines the design, reliability, and validity of classroom observations. This entry focuses on work that has been done in K–12 education only.

The History of Classroom Observations

Classroom observations have long been a primary component of teacher evaluation, a process that was intended to both support employment-related decisions and provide feedback to teachers to improve their instruction. Typically, principals or other school administrators would observe teachers once a year as they taught their class. Such observations were often required and specified by statute and/or labor contracts. The observer would usually make and record evaluative judgments using a form that listed a set of statements about teacher behaviors and classroom characteristics (e.g., *The teacher praised students*; *The students were well-behaved*). Based on the observations, principals would then provide an annual evaluation that included written and/or oral feedback.

Traditional observation approaches have been criticized as not being particularly useful in providing meaningful evaluative information in the majority of school contexts. In fact, the vast majority of teachers received the highest scores available across observation rating scales. Policy makers and many policy researchers have been dissatisfied with evaluations that did not effectively identify teachers who were weak performers. Educators were dissatisfied that evaluations were, in many cases, little more than a bureaucratic necessity that contributed little to the improvement of instruction.

This dissatisfaction with traditional observation methods, together with a significant research interest in understanding and characterizing teaching quality, has led to significant efforts in the design and study of classroom observations in order to assess, evaluate, and improve teaching.

The Structure and Process of Observation Protocols

A *protocol* includes a set of concepts, processes, and procedures that describes the design, training, implementation, scoring, and quality control of a classroom observation measurement system.

Observation systems require the application of a *scoring rubric* to a sample of teaching, typically a class lesson. A rubric typically consists of a set of *dimensions* that describe important aspects of teaching, such as classroom discourse, behavior management, and depth of content. Some protocols focus almost exclusively on teacher actions, while others focus on interactions between teachers and students in the classroom. Dimensions are often organized into *domains* that capture larger constructs of teaching, such as classroom environment and quality of instruction. Each dimension has a scale that includes categories describing different qualities of teaching. The number of points on the scale varies among rubrics. Rubric scales can be binary (i.e., a certain type of evidence is present or absent) or can differentiate performance levels, with between three and seven levels used most often.

Some protocols also include evidence and scales that address aspects of teaching beyond the observed lesson. For example, some protocols may include lesson plans, classroom artifacts such as instructional assignments, and teacher’s oral or written reflections on the lesson.

Although the descriptions in the rubric are typically quite brief, many protocols also include detailed elaborations designed to help observers (and teachers) understand the intended meaning of each protocol dimension. Elaborations may include exemplars and explicit behavioral indicators that suggest different levels of teaching quality.

Protocols also differ in terms of their domain and age/grade specificity. Some of the most frequently used protocols in schools are designed to be applied across all K–12 classrooms by observers who may or may not have particular subject-matter expertise. Others are designed specifically to score observations in certain disciplines such as mathematics, science, or language arts. Domain-specific protocols provide greater detail about the nature of good teaching that is appropriate for a particular domain. For example, while a general protocol may include a dimension that values *analytic reasoning*, domain-specific protocols are more apt to detail the nature of such reasoning in the respective domain. Most often, domain-specific protocols have been used in research contexts.

The protocol design also specifies how teaching is *sampled* to create observation scores. In some protocols, the observer may watch an entire lesson before scoring the teaching. In other protocols, raters may observe lessons for some specified period, often referred to as a *segment* and lasting from 7 to 15 minutes, and then apply scores for the individual segment. Segment scores are then aggregated through averaging to create a lesson score.

Observers (or raters) learn to use the protocol through a *training* process that normally involves a review of each scoring dimension together with sample videos that exemplify different score points. Training includes instructions for taking notes that are thorough, descriptive, and nonevaluative during the observation segment. Once the segment is complete, raters review the notes and assign scores for each dimension. During training, scores are discussed, as the trainer attempts to ensure that observers have learned how to use the protocol to make valid judgments.

At the completion of training, raters typically take and need to pass a *certification* test. Such tests require candidates to assign scores to a small sample of selected video segments. Raters are certified if their assigned scores meet a threshold of agreement with the scores assigned by experts who have rated the videos previously.

Once raters are trained, the observation system may also include the *recalibration* of raters to ensure that they continue to apply the protocol dimensions as designed. Raters will be asked to score videos that have been assigned scores by experts. Raters may have to pass a recertification test and/or enter into discussion if their scores deviate substantially from the expert scores.

Conducting Observations

In educational practice, most observations are done with the observer physically present during the classroom lesson. The record of the observation consists of the observer's notes and scores. In research and assessment contexts, the classroom lesson is often captured using video camera technology. Raters judge the quality of teaching by watching the video and assigning scores as directed by the protocol. The training of observers can also vary and is increasingly being conducted using web-based training tools.

The scheduling of observations can also vary. In most cases, observations are scheduled in advance, but in some teacher evaluation systems, at least some observations are conducted without any prior notice to teachers. The number of times a teacher is observed for the purposes of an evaluation or assessment can vary substantially across and within systems. Within systems, the number of evaluations may depend on teacher seniority and prior evaluation scores. New teachers and those with weaker evaluation scores are observed more frequently in some systems.

In most evaluation systems, principals, assistant principals, and curriculum supervisors who have supervisory authority conduct the observations. After a short time following the observation, there is generally a feedback session in which the observer discusses the observation results with the teacher. Guidance to administrators about carrying out effective feedback sessions varies across protocols.

Research on Classroom Observations

Substantial research has been conducted on the reliability and validity of classroom observations. General patterns of research approaches and findings are described in this section.

Reliability

Reliability of observations considers the consistency of scores obtained under different conditions. Some variation, such as that which results from different raters judging the same lesson, represents measurement error. Other variation in scores, such as that which may be associated with different lessons, may be a function of actual differences in teaching performance. Researchers and evaluators often want to estimate a teacher’s overall teaching effectiveness across different sources of variation. Thus, reliability analyses are used to determine how the number of observations and the conditions under which scores are produced affect the stability of an estimate of teaching quality for a given teacher.

Reliability of Raters

A key measurement issue concerns the degree to which different raters agree on the scores they give to an observation. Research studies often include observations that are independently scored by two raters. Agreement is sometimes indexed by exact agreement rates (i.e., both raters assign the same score), but better estimates that take into account the distribution of scores include κ and intraclass correlations.

Studies have also used generalizability designs to determine the proportion of score variance that is attributable to raters. Across studies, there is substantial error attributable to raters, implying that using multiple raters would improve the quality of scores. It is also more challenging to obtain reliable scores on some dimensions than others, suggesting that observers may have more idiosyncratic understandings of different aspects of teaching.

Reliability of Lessons

Scores vary substantially across individual lessons. Researchers have also found that the quality of lessons can vary substantially over the course of the school year. Thus, it is important not to characterize a teacher’s overall performance on the basis of a single lesson or from only one part of the school year. In order to fairly evaluate teachers, it is important to observe teachers multiple times (research generally recommends four observations to obtain a reliable estimate of teaching quality) that occur at different points during the school year.

Reliability of Mode

Researchers have explored whether it matters if the observation occurs live, takes place with the observer in the classroom, or is video recorded and later scored by a rater who was not physically present during the lesson. Scores based on these two methods are correlated almost perfectly when the correlation is adjusted for measurement error.

Validity

Validity refers to the quality of evidence supporting the interpretation and use of observation scores. The majority of research has focused on the meaning of scores and their relationship to other measures associated with teaching quality.

The Meaning of Scores

A number of studies have examined evidence to support the dimensional and domain structure of observation protocols. Using factor analysis, research has examined whether patterns of scores are consistent with the theoretical structure built into the protocol design. The evidence is mixed. In general, dimensions within a protocol are highly correlated and dominated by one general factor. A number of studies, however, have found modest evidence in support of multiple factors. In general, to the extent that different factors are identified, they seem to be associated with instructional and classroom environment constructs.

Other studies have examined the distribution of scores across dimensions. This work has shown that scores for classroom management dimensions (e.g., behavior, on-task activity) are typically higher on their respective scales than they are for instructional dimensions (e.g., the quality of content and reasoning about subject matter). In fact, across protocols, average scores for instructional dimensions are usually described by lower scale points in the scoring rubric.

Relationships to Other Measures

One question is whether different observation protocols lead to different inferences about teacher quality. In general, relative rankings of teachers across different protocols are very similar. Although different protocols may emphasize different aspects of teaching, they will result in highly similar relative ranking of teachers.

A second question is the extent to which observation protocols are related to other measures of teacher quality. The most widely studied measures are those that estimate teachers’ contributions to student learning (e.g., value-added methods). This work has found consistently significant but modest correlations between these measures. Studies have also explored relationships with other measures including teacher knowledge and student surveys. Findings are mixed, but even when correlations are significant, they tend to be low. Low correlations are, in part, due to the measurement error associated with each of the separate measures contributing to the correlation.

Classroom Observations in Practice

Classroom observations have been and are currently used as part of formal assessment and evaluation systems. They now play a major role in state teacher evaluation systems that have been guided by federal and state policy. They have also contributed to assessments that have been used to license and certify teachers.

Licensure and Certification

In the 1980s, efforts were made to assess teachers by examining actual samples of teaching and not simply relying on paper-and-pencil tests. The National Board for Professional Teaching Standards, as well as a number of state licensure systems, called for a teaching portfolio that would include one or two video segments of teaching. The video was only one part of a portfolio that contained extensive teacher reflection and other teaching artifacts. The portfolio was scored using standardized assessment practices with trained raters who also had subject-matter expertise related to the teaching they were scoring.

Teacher Evaluation

More recently, teacher evaluation has been a cornerstone of educational accountability, formally codified in the Race to the Top federal initiative. States were expected to develop teacher evaluation systems that included both student growth and classroom practice components. All state evaluation systems have relied on classroom observations as the sole or major contributor to the classroom practice component.

States and districts have adopted commercially available observation protocols, some of which have been part of the research described previously. Other states and districts have developed their own protocols, although they have a great deal in common with the commercial products.

Given the constraints and pressures faced by schools, it is not surprising that many research recommendations have not been heeded in practice. For example, training and certification testing is normally much less intensive than it is in research studies. Double scoring of given observations to evaluate reliability is rare. Raters often have limited knowledge of the subject matter that is being taught. Many teachers are observed for fewer lessons than is recommended by research. And perhaps most importantly, observers have a personal stake in and a relationship with the teachers they are observing, which is quite different from the conditions under which research study observations are made. Additionally, policies often specify that teachers receiving mediocre ratings are subject to some type of job-related actions up to and including termination over time.

Therefore, it is also not surprising that scores vary substantially from those found in research studies. Scores are substantially higher in practice than they are in research studies. In some schools, there may be no discrimination at all among teachers. Thus, the meaning of scores that has been considered in research studies is quite different from the scores that are generated by evaluation systems in practice. There are few published studies that examine the quality of scores produced in functioning teacher evaluation systems.

Drew Gitomer

See also Certification; Data-Driven Decision Making; Race to the Top; Teacher Evaluation; Teachers' Associations; Value-Added Models

Further Readings

Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. Educational Assessment, 17, 62–87. doi:10.1080/10627197.2012.715014

Bill and Melinda Gates Foundation. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Washington, DC: Author. Retrieved from http://eric.ed.gov/?id=ED540960

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. Teachers College Record, 116(6). Retrieved from http://www.tcrecord.org/Content.asp?

ContentId=17460

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. Harvard Educational Review, 83(2), 371–384. doi:10.17763/haer.83.2.d11511403715u376

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. Educational Researcher, 38(2), 109–119. doi:10.3102/0013189X09332374

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York, NY: The New Teacher Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Douglas Steinley Douglas Steinley Steinley, Douglas

Cluster Analysis Cluster analysis

293

298

Cluster Analysis

Generally, cluster analysis refers to the goal of identifying or discovering groups within the data, in which the primary caveat is that the groups are not known *a priori*. Prior to discussing methods for identifying clusters, it is helpful to consider the fundamental question: What is a cluster? For an $N \times P$ data matrix $\mathbf{X}$, containing measurements on N observations across P variables, each observation can be thought of as a point in P dimensional space. Clusters then are groups of points in P dimensional space that are similar in some fashion. After furthering the introduction of clusters, this entry lists and then examines the seven steps of cluster analysis. Those steps include determining which observations are to be clustered, which variables are to be used, and whether those variables should be standardized. Subsequent steps include selecting an appropriate measurement, choosing the clustering method, and then determining the number of clusters. The final step focuses on interpreting, testing, and replicating the results of the cluster analysis.

In an early, and still excellent, review of the field of cluster analysis presented to the Royal Statistical Society, Richard Melville Cormack advanced the notion that clusters have to be externally isolated and internally cohesive. Geometrically, internal cohesion indicates that the observations within a cluster are “clumped” together in the multivariate P dimensional space, whereas externally isolated indicates that the observations are well separated from each other. Alternatively, this can be further thought of as regions of the multivariate space that are dense with spaces of “sparseness” separating them, leading to a natural conceptualization of clusters corresponding to multiple modes in the multivariate space. While attempting to capture this dual notion of isolation and cohesion, several different metrics of “clusteriness” and algorithms to uncover these notions have been developed.

The vast majority of methods, whether hierarchical or nonhierarchical, often have the goal of obtaining a clustering solution such that the clusters are mutually exclusive and collectively exhaustive; that is, that each observation is assigned to one and only one cluster and all observations are assigned to at least one cluster. Initially, it seems like the best approach would be to evaluate all possible cluster solutions (e.g., look at all possible assignments of observations to clusters); however, the number of possible solutions (e.g., partitions) is enormous. Specifically, for *N* observations and *K* clusters, the number of possible ways to assign the *N* observations to the *K* clusters is given by the Stirling number of the second kind:

$$S(N,K)=\frac{1}{K!}\sum_{i=0}^{K}(-1)^{i}\binom{K}{i}(K-i)^{N},$$

a quantity that can be approximated by , which increases rapidly for increases in both *N* and *K*. For instance, the number of possible partitions of 20 observations into five clusters is 7.95 × 10^{11}; modestly increasing the sample size to 100 observations results in 6.27 × 10^{81}, resulting in a situation in which it is impossible to evaluate all possible partitions. As such, the goal has been to develop approaches that give good solutions without evaluating all possible partitions. Because all possible solutions are not being evaluated, the ability to definitively state that the best solution (often referred to as the globally optimal solution) has been found is lost. That is, these approaches are heuristic in nature in that a set of rules are established that defines how the approach identifies the resultant clusters and not all possible solutions are evaluating, so it is impossible to know whether the final set of clusters is the best possible of all the *S*(*N*, *K*) partitions. Given the monumental task of selecting a candidate partition as “the best” of all the possibilities, it is necessary to have some guidelines about how to proceed.

Steps of Cluster Analysis

Conducting a cluster analysis requires many decision points. Over his career, Glenn Milligan established the following seven steps as the requisite components of conducting a competent cluster analysis:

1. Determine which observations are to be clustered.

2. Determine which variables should be used in the clustering.
3. Determine whether and how variable standardization should be implemented.
4. Select an appropriate measurement of (dis)similarity for assessing how similar observations are with each other.
5. Choose the clustering method/algorithm.
6. Choose the number of clusters.
7. Validate the cluster solution through interpretation, testing, and replication.

Determining the Observations to Be Clustered

The primary directive of choosing the observations to be clustered is to ensure that the observations should contain the underlying clusters or populations that are being sought. Simply put, the resultant clusters will reflect the cluster structure within the sample. Unlike many methods, cluster analysis is not wholly driven by statistical theory, and the sample size estimates and power analysis that are normally employed with many techniques (e.g., regression) are not available. Beyond making sure that the sample covers the groups being sought, one other concern when choosing the observations is whether a particular observation may or may not be an outlier. Unfortunately, this determination is complicated by the fact that outliers (e.g., an observation that is far from all other observations) could be (a) part of a small cluster if there are other “outliers” in the same vicinity or (b) the only point from a cluster that was undersampled in the original construction of the data set. As such, specific recommendations in the form of thresholds on a metric for determining an outlier are not available; rather, each situation must be evaluated based on the judgment of the analyst and subject matter experts.

Determining the Variables to Be Used

Many statistical methods are somewhat “robust” to the inclusion of irrelevant variables in the analysis. For example, in the case of multiple regression, if a predictor variable that is not related to the dependent variable is included in the regression model, it will appear as a nonsignificant predictor that can be removed on further refinement of the model. Often, users will carry over this logic to variable inclusion in cluster analysis and follow what is deemed as the “everything but the kitchen sink” approach; in other words, researchers often include all possible variables at their disposal in the cluster analysis. This approach is often used so as not to “lose” important information on any of the variables and is likely reinforced due to there being no inherent penalty in terms of algorithmic convergence or estimation by including a large number of variables. Unfortunately, cluster analysis is quite sensitive to the variables that are included. Specifically, each variable in the cluster analysis increases the dimensionality of the space in which the observations reside, and this increased dimensionality results in an overall *de facto* increase in the variation within the system. In the context of cluster analysis, variability can be thought of as good (e.g., originating as a by-product of a multimodal cluster structure in the multivariate space) or bad (e.g., originating as variability due to the nature of how the specific variables are measured and not related to the cluster structure being sought). In an idealized scenario, it would be possible to relate variable importance to the amount of multimodality exhibited by the marginal distribution of the variable; however, it is entirely possible that the multimodality (e.g., clusters) only manifest itself when certain variables are combined and examined jointly. Unfortunately, the inclusion of even one or two variables that comprise predominantly “bad” variability that never results in clusters being identified either in the full space or any of the possible reduced spaces can result in the complete inability to recover the true cluster structure present in the sample. As such, it is imperative that each variable is evaluated with regard to how it should theoretically contribute to uncovering the overall cluster structure being sought.

Determining Whether the Variables Should Be Standardized

One of the primary decisions that data analysts encounter in any analysis is the decision of whether to standardize the variables prior to analysis. Some type of standardization is usually recommended if variables are measured on different scales. Furthermore, the method of standardization that has become the *norm de rigueur* is the *z* score:

$$z_i = \frac{x_i - \bar{x}}{\sigma_X},$$

where is the mean and σ_X is the standard deviation of the variable *x*, with such a standardization resulting in each variable having a mean of 0 and a standard deviation of 1. However, studies have shown that this is not the preferred method for standardization in the context of cluster analysis. Specifically, by setting all the variables to have a variance of 1, variables with cluster structure will be relatively down weighted and variables without cluster structure will be relatively up weighted.

An alternative method to the standard *z* score is standardizing by the range, given by the formula:

$$z_i^r = \frac{x_i}{\max(x) - \min(x)}.$$

This particular standardization seems to alleviate many of the problems related to standardization using *z* scores. Further, several studies have shown that clustering on range-standardized data generally leads to better recovery of cluster structure than clustering on raw data, whereas clustering *z* score standardized data can actually lead to worse recovery than clustering the raw, unstandardized data.

Selecting a Proximity Measure

To determine whether clusters are present within the data, it is necessary to define how similarities and differences between observations are measured. This is usually accomplished with some type of distance measurement, with the most common being Euclidean distance

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{iP} - x_{jP})^2},$$

where x_{ip} is the measurement of the *i*th observation on the *p*th variable and d_{ij} represents the distance between the *i*th and *j*th observations. Some common alternative distance measures include the squared Euclidean distance (obtained by squaring the Euclidean distance) and the city-block distance (e.g., Manhattan or Taxi-Cab metric), which is obtained by taking the sum of absolute differences across all variables. Geometrically, the Euclidean distance is a straight-line distance between two observations in the multivariate space, while the city-block distance is the sum of the distances along each of the axes that define the multivariate space. The former is the most commonly used measure (and its squared counterpart), while the latter is used in situations where it is suspected that there may be outliers.

Although the aforementioned are the most common measures for continuous variables, there is a large suite of possible measures for binary data as well. In fact, the statistical computer package SPSS contains between 20 and 30 different measures of binary similarity. Although it is prohibitive to discuss each in detail, the measures are based on different formulations of four different quantities for each pair of observations: (1) the number of times that each observation exhibits a “1” on the same variable, (2) the number of times that the first observation exhibits a 1 on a variable and the second observation exhibits a “0” on the same variable, (3) the number of times that the first observation exhibits a 0 on a variable and the second observation exhibits a 1 on the same variable, and (4) the number of times both observations exhibit a 0 on the same variable. These four quantities lead to numerous proximity measures. For instance, the simple matching coefficient is $(a + d)/(a + b + c + d)$.

Choosing the Clustering Method/Algorithm

Traditionally, cluster analysis can be broken down into two different types of approaches: hierarchical and nonhierarchical. Although cluster analysis is an innate human process (e.g., assigning observations/items to groups), its formal representation as a mathematical algorithm extends back (at least) to the psychometrician Robert Thorndike (nonhierarchical clustering) and the biologists Robert Sokal and Peter Sneath (hierarchical clustering). Each of these approaches can be thought of as broad classes of approaches for finding clusters, with many variations within each class.

Hierarchical Clustering

Hierarchical clustering itself can be divided into two types: agglomerative and divisive. In the former, all observations begin in their own cluster and, as the algorithm proceeds, each observation is merged into clusters (either existing or creating a new cluster) until all observations are in the same cluster. The reverse is true for divisive clustering—all observations begin in one cluster and they are divided until each observation is in its own cluster. Of the two types, agglomerative is much more popular than divisive, so attention is focused there.

Within agglomerative hierarchical clustering, there are several different approaches for joining the observations, with the four most popular being single linkage, complete linkage, average linkage, and Ward’s method. These techniques differ in how they define the similarity between pairs of observations, pairs of clusters, or an observation (e.g., a singleton cluster) and a cluster; however, the general structure of the agglomeration process is the same:

1. Begin with each observation in their own cluster.
2. Compute the distances between all pairs of clusters. Before any observations are merged into clusters, this is computing the pairwise distance between all observations. The closest pair is then merged together to begin the process—even though the algorithms discussed can lead to quite different results, the first merger is the same for all of them.
3. Merge the pair of clusters that are closest to each other.
4. Repeat Steps 2 and 3 until all observations belong to the same cluster.

As previously stated, the algorithms only differ in the definition of “close” in Step 3. *Single linkage* (e.g., nearest neighbor clustering) defines the distance between two clusters as the minimum distance between any pair of elements within the two clusters, resulting in clusters that can be long and straggly. *Complete linkage* (e.g., furthest neighbor clustering) defines the distance between two clusters as the maximum distance between any pair of elements within the two clusters, resulting in compact clusters of similar size. *Average linkage* defines the distance between two clusters as the average pairwise distance between the observations in one cluster with the observation in the second cluster. *Ward’s method* merges observations such that the increase in the total variance of the clusters is a minimum at each step; as such, this is often done using squared Euclidean distance. Ward’s method often results in spherical clusters of similar size.

Nonhierarchical Clustering

Rather than merging the observations in a sequence, nonhierarchical methods initialize their algorithm with some starting “seed” and then iterate through an estimation process until convergence has been reached. There are several nonhierarchical clustering algorithms of varying degrees of complexity, with the most popular being *K*-means clustering. While there are variations on the *K*-means algorithm, a general structure is the following alternating least-squares approach:

1. Randomly assign each observation to one of the *K* clusters.
2. Compute the mean of each cluster.
3. Compute the distance of each of the *N* observations to each of the *K* cluster means.
4. Assign observations to the cluster with the closest mean.
5. Repeat Steps 2–4 until no observation changes cluster membership.

The *K*-means clustering algorithm attempts to minimize within-cluster variability and can be thought of as the nonhierarchical counterpart to Ward’s method. This type of algorithm (and nonhierarchical clustering in general) is prone to converging to a locally optimal solution, meaning that the final solution will depend heavily on the initialization in Step 1. As such, it is often recommended to initialize the *K*-means clustering algorithm several thousand times (say 5,000) and choose as the final solution the one with the smallest within-cluster variability. Other nonhierarchical clustering approaches include the *P*-median algorithm, swarm optimization approaches, and many more.

Choosing the Number of Clusters

Similar to all of the decisions in cluster analysis, there are numerous methods to choose the number of clusters; however, the favored technique is evaluating several different cluster solutions and comparing them with an index that takes into account the number of clusters being fit to the data. Although there are many, the most popular index that seems to compare favorably under many different scenarios is the Calinski–Harabasz index:

$$CH(K) = \frac{\mathrm{SSB} / (K-1)}{\mathrm{SSW} / (N-K)},$$

where SSB and SSW are the sum of squares between clusters and the sum of squares within clusters, respectively. The general approach is to fit estimate the number of clusters for several different values of *K* and then choose the solution that maximizes *CH(K)*. This index favors well-separated, compact clusters—and, all else being equal, a fewer number of clusters is better.

Interpreting, Testing, and Replicating

The final step in the cluster analysis is to assess the quality of the clustering solution. Interpreting the results relies on substantive knowledge about the content domain, and specific recommendations are not given beyond evaluating the means and variances of the clusters and determining how much they comport with theoretical expectations.

Statistical testing for cluster analysis is a little more delicate. Critically, tests to determine whether the clusters are significantly different from each other can only be conducted on variables that were not used to create the clusters in the first place. These tests are often conducted using analysis of variance in which the clusters serve as the groups. If the variables that were used to construct the clusters are tested, the results will almost always be significant; however, these results are invalid because the groups were designed to be maximally separate on those variables, and as such, any associated *p* values are meaningless. This type of analysis is analogous to externally validating the cluster structure.

Internal validation relies on replication, which is usually accomplished by splitting the sample into two (or more) subsamples and determining whether the same cluster structure is extracted from each subsample. The comparison of cluster structure can be done either informally by assessing whether the parameters (e.g., means and variances) are consistent across solutions or formally by comparing the solutions via an appropriate index (e.g., adjusted Rand index). Unfortunately, it is possible to consistently extract similar cluster structure from the data without actually proving that clusters truly exist. Even if that is the case, each of the decisions discussed in this entry can have a dramatic impact on the final solution, and being able to replicate that solution is required to have a semblance of confidence in the final result. While replication is a necessary but not sufficient condition for determining whether true clusters have been identified, it is also the first step in determining whether a cluster structure exists.

Douglas Steinley

See also Cluster Sampling; Exploratory Factor Analysis

Further Readings

Cormack, R. M. (1971). A review of classification. Journal of the Royal Statistical Society, A, 134, 321–367.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis (5th ed.). New York, NY: Wiley.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50, 159–179.

Sneath, P. H. A., & Sokal, R. R. (1973). Numerical taxonomy. San Francisco: W. H. Freeman.

Steinley, D. (2006). K-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology, 59, 1–34.

Breanna A. Wakar Breanna A. Wakar Wakar, Breanna A.

Dmitriy Poznyak Dmitriy Poznyak Poznyak, Dmitriy

Cluster Sampling Cluster sampling

298

299

Cluster Sampling

Cluster sampling is a probability sampling technique in which all population elements are categorized into mutually exclusive and exhaustive groups called clusters. Clusters are selected for sampling, and all or some elements from selected clusters comprise the sample. This method is typically used when natural groups exist in the population (e.g., schools or counties) or when obtaining a list of all population elements is impossible or impractically costly. As compared to simple random sampling, cluster sampling can reduce travel cost for in-person data collection by using geographically concentrated clusters. At the same time, cluster sampling is generally less precise than simple random or stratified sampling; therefore, it is typically used when it is economically justified (i.e., when a dispersed population would be expensive to survey). This entry discusses selecting clusters with equal and unequal probability and provides a comparison of cluster sampling to other sampling methods.

Selecting Clusters With Equal Probability

Cluster sampling can be applied in one or more stages but, regardless of the number of stages, the first step is to select the clusters (primary sampling units) from which sample elements (secondary sampling units) will be drawn. A basic one-stage design takes a simple random sample of clusters and selects for sampling all elements within those clusters, although this design is rarely used in practice. A researcher could select schools and collect data about every student in the selected schools. Because elements within a cluster are often similar—a phenomenon called a cluster effect—it may be redundant and inefficient to sample a large proportion of the elements within a cluster.

Large-scale studies typically use a multistage cluster sampling method. A basic implementation of this type of sample is a two-stage cluster sample selecting clusters via simple random sample and independently subsampling elements within each cluster, using the same sampling fraction across clusters. The downside of this simple approach is that it results in differing sample sizes per cluster, making it less attractive than other designs. Designs with more than two stages may also be useful; a three-stage statewide survey, for example, could sample school districts, then schools within selected districts, then teachers within selected schools.

In multistage sampling, the variance of the estimated quantities depends on within-cluster and between-cluster variance. Within-cluster variance is related to the intraclass correlation coefficient (ICC), which measures the degree of homogeneity of the variable of interest for elements within a cluster. ICC is typically interpreted as the correlation between the responses of individuals in the same cluster. Using schools as clusters and students' test scores as an outcome, an ICC of 0.2 would mean that 20% of the variation in the student test scores is accounted for by the school a student attends, and 80% is accounted for by variation across students within schools.

Selecting Clusters With Unequal Probability

Clusters may also be selected with probability proportional to size. This means that clusters containing a greater size measure (e.g., the number of population elements) are more likely to be included in the sample than clusters with fewer elements. Such a sampling scheme would, for instance, be more likely to select a college dormitory where 100 students live than one where 20 live. Specifically, the probability of selection for cluster c is $m \times N_c N$ where m clusters are selected from the population of clusters, N_c is the measure of size (e.g., the number of secondary sampling units) in cluster c, and N is the sum of the measure of size across all clusters (e.g., the number of elements in the population). If the cluster sample is stratified, then the numbers in this probability formula reflect those in a particular stratum. The second stage sample could select an equal number of elements from each cluster. This creates an equal workload in each cluster, which is preferable if data collection involves face-to-face communication, and in expectation results in a self-weighting sample. If the probability of selection within cluster c is nN_c, then the cumulative probability of selection for each secondary sampling unit reduces to *mnN*, a constant. The variance of an estimator of the population mean is a function of the number of clusters selected, the sample size within each cluster, and the ICC. Higher ICC values increase the variance for a given sample size, and increasing the ratio of the number of clusters selected to the sample size within a cluster reduces the overall variance and increases the precision of the final estimates.

Comparison to Other Sampling Methods

Cluster sampling has some parallels to stratified sampling, in that both divide the population into groups (clusters or strata) and make selections from those groups. However, cluster sampling samples elements from only selected clusters, where stratified sampling selects a sample of at least one element from every stratum. As a result, in cluster sampling, the selection process is subject to two levels of chance variation, one for the selection of clusters and one for the selection of elements within a cluster. In contrast, the stratified sampling process (and its simpler form as a simple random sample) is only subject to one level of chance from the selection of elements within each stratum, which will often improve the precision of an estimate made from the sample. There are exceptions to this general rule depending on how the sample size is allocated across the strata to produce subgroup estimates. But stratified and simple random sampling, in general, produce a more precise estimate than cluster sampling for the same sample size.

Breanna A. Wakar and Dmitriy Poznyak

See also Representativeness; Sample Size; Simple Random Sampling; Stratified Random Sampling; Variance

Further Readings

Bethlehem, J. G. (2009). Applied survey methods: A statistical perspective. Hoboken, NJ: Wiley.

Kish, L. (1965). Survey sampling. New York, NY: Wiley.

Levy, P. S., & Lemeshow, S. (2008). Sampling of populations: Methods and applications. Hoboken, NJ: Wiley.

Lohr, S. L. (2010). Sampling: Design and analysis (2nd ed.). Boston, MA: Brooks/COLE.

David C. Hoaglin David C. Hoaglin Hoaglin, David C.

Cochran *Q* Test Cochran *Q* test

299

302

Cochran *Q* Test

In conducting meta-analyses of estimates from a set of studies, researchers often use a statistic denoted by *Q* to assess the homogeneity of the estimates. When used as the basis of a formal statistical test of homogeneity, *Q* is commonly referred to a chi-square distribution, and the test is called the Cochran *Q* test. The name of the test, however, embodies a misunderstanding: Although William G. Cochran wrote about the statistic *Q*, he did not propose a test based on it. Also, the test generally uses an incorrect null distribution. This entry describes the *Q* statistic, examines its statistical behavior, and discusses implications for the heterogeneity measure I^2 and for a popular method of random-effects meta-analysis.

The *Q* Statistic

In the main paper in which *Q* appears, Cochran was concerned with combining estimates from *k* separate experiments. The types of experiments included determinations of a physical or astronomical constant, bioassays, and agricultural field experiments; typical estimates were a simple mean, a difference between two means, a median lethal dose, and a regression coefficient. Each experiment provided an estimate, y_i, and an estimate, , of the variance of y_i. Also, each had a number of degrees of freedom, n_i, which would ordinarily come from a mean square (e.g., the sample variance or the residual mean square of a regression). Thus, the setting differed from most meta-analyses. Ideally, all experiments were estimating the same quantity, μ, but it could vary among experiments (i.e., the quantities could be μ_i instead of a single μ). Thus, *Q* summarized the variation among the estimates in the form of a weighted sum of the squared deviations of the y_i from the weighted mean with weights ,

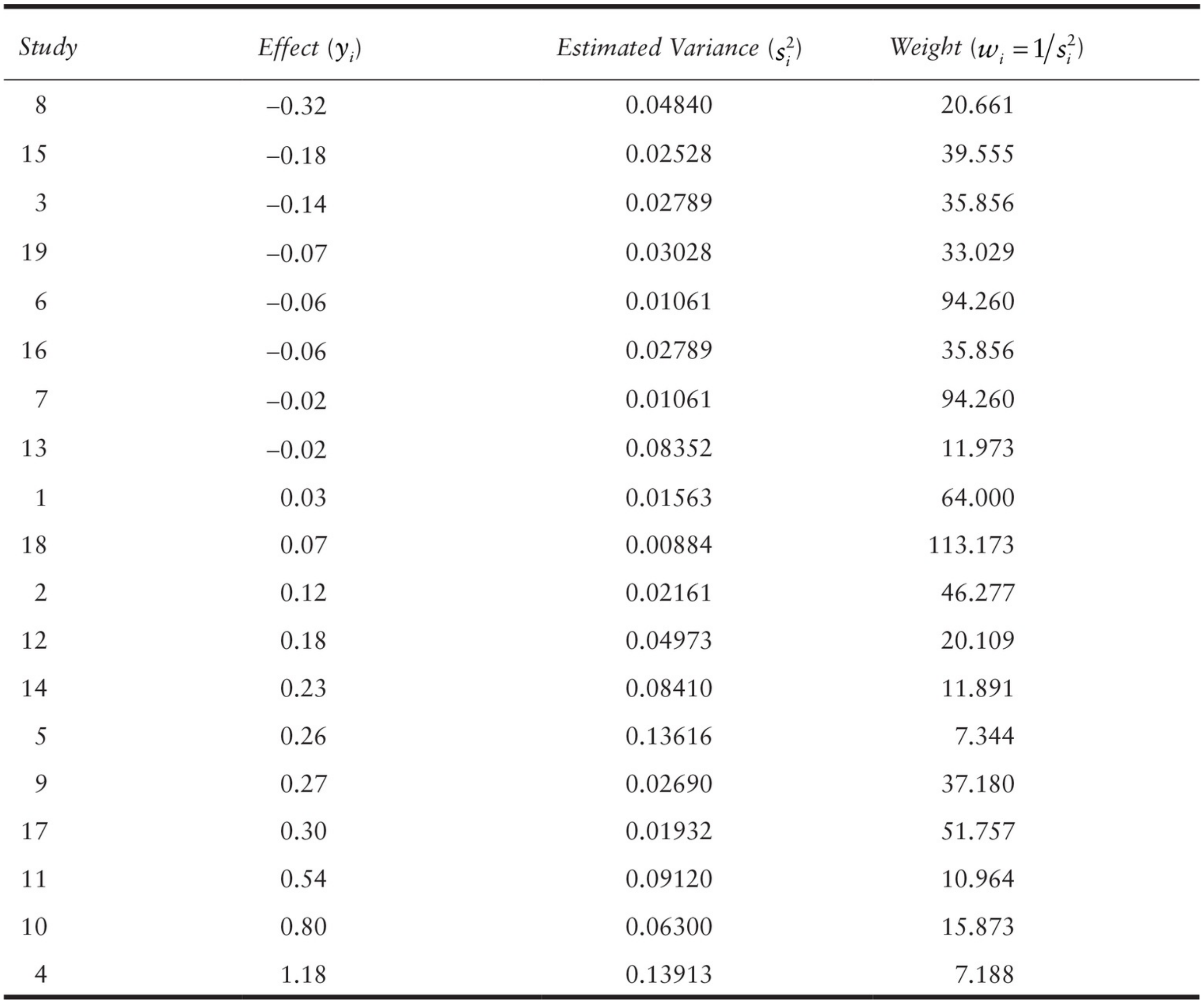
$$Q = \sum_{i=1}^{k} w_i (y_i - \bar{y}_w)^2 .$$

A large value of Q indicates heterogeneity among the y_i. The degrees of freedom, n_i, are not used in calculating Q; but they do appear, for example, in an approximate formula for the standard error of .

When the degrees of freedom (n_i) are “large” and the μ_i are equal, the distribution of Q approaches the chi-square distribution on $k-1$ degrees of freedom. The literature, however, provides little information on a quantitative definition of “large.”

In a meta-analysis, y_i is usually the estimate of the effect in Study i (e.g., the standardized mean difference or a difference of proportions), and is an estimate of its (within study) variance. The test of homogeneity refers Q to the chi-square distribution on $k-1$ degrees of freedom and rejects the null hypothesis if $p < .10$ (the criterion $p < .05$ is less common because the test is considered to have low power). Many authors routinely use chi-square on $k-1$ *df* as the null distribution. This procedure is understandable because the in a meta-analysis are rarely, if ever, accompanied by numbers of degrees of freedom. Indeed, some common measures of effect, such as a difference in proportions, do not have a natural way to define a number of *df*. When the n_i are available, they are often closely related to the total sample size of the two groups in the study. In many meta-analyses, the total sample size is not “large,” so the test of homogeneity is unreliable.

To illustrate the application of the Q statistic, Table 1 lists the values of the standardized mean difference, its estimated variance, and the corresponding weight for 19 studies of the effects of teacher expectancy on pupil IQ. These data yield and Q=35.83. The usual test of homogeneity would have $p < .01$, but (as discussed in the next section) that test uses an incorrect null distribution. Also, a single number, such as Q (or, for that matter,), seldom gives an adequate summary of a set of data. Examination of the values of y_i in Table 1 provides an informal, but more useful, indication of possible heterogeneity. The bulk of the estimates range from about −0.3 to +0.3, and three high values stand out (1.18, 0.80, and 0.54). Thus, Studies 4, 10, and 11 would merit some scrutiny. Their relatively small weights suggest that those studies may have had relatively small samples, but they may have other distinctive characteristics. The authors who published the data in Table 1 used those data in a limited methodological example and did not discuss the studies’ sample sizes or other characteristics.



Study	*Effect* (y_i)	*Estimated Variance* (s_i^2)	*Weight* ($w_i = 1/s_i^2$)
8	−0.32	0.04840	20.661
15	−0.18	0.02528	39.555
3	−0.14	0.02789	35.856
19	−0.07	0.03028	33.029
6	−0.06	0.01061	94.260
16	−0.06	0.02789	35.856
7	−0.02	0.01061	94.260
13	−0.02	0.08352	11.973
1	0.03	0.01563	64.000
18	0.07	0.00884	113.173
2	0.12	0.02161	46.277
12	0.18	0.04973	20.109
14	0.23	0.08410	11.891
5	0.26	0.13616	7.344
9	0.27	0.02690	37.180
17	0.30	0.01932	51.757
11	0.54	0.09120	10.964
10	0.80	0.06300	15.873
4	1.18	0.13913	7.188

Source: Shadish, W. R., … Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, … J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., p. 264). New York, NY: Russell Sage Foundation.

A warning: Researchers sometimes choose between a fixed-effect meta-analysis and a random-effects meta-analysis on the basis of the test of homogeneity. That is a mistake. Even if the studies’ sample sizes are “large” and the test uses the correct null distribution, the resulting two-step procedure has very little statistical support.

Statistical Behavior of *Q*

A key ingredient in the statistical behavior of Q (i.e., the distribution of Q in the presence of homogeneity or various patterns of heterogeneity) is the weights, w_i. If each weight was the reciprocal of the true within-study variance of y_i (i.e.,), straightforward applications of statistical theory would yield the null and non-null distributions of *Q*. The are usually unknown, however, and that theoretical approach is not compatible with the actual weights in the definition of *Q*, . Many authors ignore this distinction, relying on to converge to as the sample size becomes large. Accurate estimate of a variance, however, can require surprisingly large samples.

In the setting discussed by Cochran, is usually statistically independent of y_i. For some measures of effect, however, and y_i are associated because of their relation to sample estimates. If y_i is the difference between the rates of an event in the two groups in Study *i*, $y_i = p_{i1} - p_{i2}$, and the groups' sample sizes are n_{i1} and n_{i2}, the usual estimate of the variance of y_i is .

Studies of the behavior of *Q* that take into account the variability of the and their relation to y_i involve substantial complexity; they generally use asymptotic expansions to obtain approximate formulas for the mean and variance of *Q* and then verify those formulas by simulation. The results for standardized mean difference, risk difference, and odds ratio yield three different null distributions for *Q*, none of which is the chi-square distribution on $k-1$ degrees of freedom. Thus, the usual test of homogeneity gives invalid *p* values.

Implications for I^2

The heterogeneity measure I^2 is usually calculated from *Q:*

$$I^2 = \max\left\{0, \frac{Q-(k-1)}{Q}\right\},$$

and interpreted as the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies. The development of I^2 treated *Q* as having, in the absence of heterogeneity, the chi-square distribution on $k-1$ degrees of freedom. Thus, I^2 is positive when *Q* exceeds the mean of that distribution ($k-1$). Such a value of *Q* has a substantial probability: .392 when $k = 4$, increasing to .465 when $k = 30$ and approaching .5 as *k* becomes large. These probabilities are large enough to rule out interpreting I^2 as "the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies," but they do not indicate the probability of larger values of I^2. The probability that $I^2 > 25\%$ decreases from .261 when $k = 4$ to .108 when $k = 30$, and the corresponding probabilities that $I^2 > 50\%$ are .112 and .001.

These probabilities are hypothetical because the null distribution of Q is generally not chi-square on $k-1$ degrees of freedom. If one wanted to continue using the mean of the null distribution as the threshold for positive values of I^2 (not necessarily a useful choice), the formula for that mean would depend on the measure of effect size and would generally have to be estimated from the data in each meta-analysis. These complications sharply limit the usefulness of I^2 as a measure of heterogeneity.

Role of *Q* in Random-Effects Meta-Analysis

The popular DerSimonian-Laird method for random-effects meta-analysis uses Q as the basis for its estimate of the between-study variance (i.e., the variance of the μ_i in the random-effects model). The method does not assume a particular distribution for Q, but it does use the as if they were the . The consequences include bias in estimating the between-study variance, bias in the overall estimate, and confidence intervals with inadequate coverage. The relation between y_i and is an underappreciated source of bias.

David C. Hoaglin

See also Meta-Analysis; *p* Value; Power

Further Readings

Cochran, W. G. (1954). The combination of estimates from different experiments. Biometrics, 10, 101–129. doi:10.2307/3001666

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7, 177–188. doi:10.1016/0197-2456(86)90046-2

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21, 1539–1558. doi:10.1002/sim.1186

Hoaglin, D. C. (2016). Misunderstandings about *Q* and “Cochran’s *Q* test” in meta-analysis. Statistics in Medicine, 35, 485–495. doi:10.1002/sim.6632

Kulinskaya, E., & Dollinger, M. B. (2015). An accurate test for homogeneity of odds ratios based on Cochran’s *Q*-statistic. BMC Medical Research Methodology, 15, 49. doi:10.1186/s12874-015-0034-x

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011). On the moments of Cochran’s *Q* statistic under the null hypothesis, with application to the meta-analysis of risk difference. Research Synthesis Methods, 2, 254–270. doi:10.1002/jrsm.54

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011). Testing for homogeneity in meta-analysis I. The one-parameter case: Standardized mean difference. Biometrics, 67, 203–212. doi:10.1111/j.1541-0420.2010.01442.x

Mohsen Tavakol Mohsen Tavakol Tavakol, Mohsen

Coefficient Alpha

Coefficient alpha

302

306

Coefficient Alpha

Social science measurement requires scores that are both valid and reliable. Reliability refers to the amount of random fluctuation in a set of scores. One of the most popular methods for estimating the reliability of test scores is coefficient α, which estimates the proportion of observed score variance that is due to true score variance. The coefficient α is widely used in the education and psychosocial literature as an index of reliability of test scores.

Coefficient α was first described by Lee Cronbach in 1951, so it is sometimes termed Cronbach’s α. Coefficient α is used for establishing reliability of test scores from a single administration. However, in spite of the common use of coefficient α in the literature as an index of reliability of test scores, there is a lack of the proper use of coefficient α and its interpretation.

To achieve a greater understanding of coefficient α, this entry discusses these issues. First, the entry offers a more complete definition of reliability and then provides a detailed explanation of coefficient α. The entry then looks at assumptions of coefficient α as well as conceptual issues of this measure of reliability. Finally, the entry demonstrates when the use of alternative methods of reliability, such as ω, may be more appropriate.

Reliability

In 2011, in a study by Tavakol and Dennick, the reliability of a test is concerned with the capability of a test to consistently measure an attribute. A consistent test generates more or less the same results when administered on different occasions. Indeed, reliability refers to the trustworthiness of test scores (McDonald, 2014).

Reliability is based on the classical test theory model. Under the classical test theory model, the observed score is equal to the true score plus the error score (observed score = true score + error score). Researchers are interested in gaining knowledge about the true score and the discrepancy between an observed score and true score (i.e., how strongly the observed score is related to the true score). In other words, researchers want to know how much the differences in observed scores can be directly determined (explained) by the differences in true scores? By answering this question, researchers are in a position to identify the reliability index. Indeed, the reliability index is a correlation between the true score and the observed score (Raykov and Marcoulides, 2011). Given this, reliability is the ratio of the individual differences (variance or variability) of the true score to the individual differences of the observed score. Thus, the greater the ratio of the true score differences to the observed score differences, the more reliable the test.

The Source of Measurement Error and Reliability

As previously pointed out, reliability is the ability of a test to consistently measure an attribute (e.g., students' competencies). For example, if a student is tested 1,000 times using the same assessment questions, these tests will likely have different scores. The mean of all 1,000 scores produces the true score of the student. However, practically researchers are unable to conduct that many tests in order to obtain the true score of the student. Using a single test administration, the observed score can be cofounded by random or systematic errors, as proposed by the classical test theory.

The random and systematic errors attached to the observed score decrease the function of a particular test. No test is immune from random or systematic errors. Indeed, reliability refers to the degree to which test results are free from random errors. Thus, if researchers minimize the measurement errors associated with the obtained scores in a test, the test scores will be more consistent, and researchers can have enough confidence to make a decision about the test results. In the following section, the association between reliability and the coefficient α is discussed.

What Is Coefficient α?

What Is Coefficient α?

Error variance is an integral part of any measurement. If researchers are interested in measuring the error variance attached to the different items of a particular test, coefficient α is calculated to obtain the error variance that is due to the interaction between item scores and persons. The following formula is used for coefficient α:

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_j^2}{\sigma_x^2}\right),$$

in which *n* is the number of items in a test, is the sum of the variances of the individual items, and is the variance of the whole test. Based on this formula, we need to have two or more items in a test to calculate coefficient α. Coefficient α is mainly affected by two factors: the number of test items and the correlation coefficients between items (the interrelatedness of the items in a test). As the number of items in a test increases, the value of coefficient α will increase. If the items within a test are highly correlated to each other, the coefficient α increases. Thus, to boost internal consistency of test scores, the test should include more items testing the same concept (Tavakol and Dennick, 2011a). It is also noteworthy to mention that coefficient α is not concerned with the reliability of a test itself. It is concerned with the reliability of the test scores from a specific sample of testees. Because the characteristics of the testees being assessed vary from one sample to another, coefficient α will vary from one cohort of testees to another. This suggests that the coefficient α should be calculated each time the test is administered. Coefficient α is addressed as a number between 0 (uncorrelated items) and 1 (perfectly correlated items). The closer the coefficient α value is to 1, the more reliable the test results are. If testees vary in relation to their performances, that is, when there is a wide dispersion of skills or performances, the coefficient α will approach 1.

If a test produces poorly reliable results (i.e., a low value of coefficient α), it is difficult to make fair assumptions on the test and its results. However, caution should be taken in interpreting the value of coefficient α, either high or low. For example, in educational assessments, students’ abilities may be more or less the same (low variation in performance), or educators may be faced with a low value of coefficient α that is attributed to a low dispersion of scores on assessment questions. In addition, if a small sample of testees is selected for the test of questions. In addition, if a small sample of testees is selected for the test of interest, a low coefficient α may be produced. What is more, moderately difficult items (from 25% to 75% correct answers) that differentiate between who know the content and who do not contribute enormously to coefficient α values.

Given the Cronbach’s α formula, if the items of a test are statistically independent (e.g., when items are randomly scored), would be equal to , then α = 0. This indicates that if the test items are correlated to each other, the coefficient of α is increased. If items in a test provide the same information (i.e., item responses of the test are the same), is equal to , then α = 1. This indicates that a very high value of α (e.g., greater than .90) does not necessarily mean you have better test scores. This may suggest redundancies and show the test length should be reduced by removing the items that repeat the information provided by other items. The generally acceptable value range for coefficient α is from .70 to .95. However, in educational research with groups, a coefficient α greater than .60 is allowable.

Robust Estimation of the Coefficient α

When it is stated that a test is reliable, it should be reliable for the bulk of the examinees. However, the traditional coefficient α does not reflect the bulk of examinees. It has been shown that test scores are often skewed with “very heavy tails.” Because the coefficient α is based on the individual variances and total variance of test scores, a small departure from the normality of test scores can influence the variances, which in turn can provide a misleading value of the coefficient α. Consequently, it is necessary to have a coefficient α that is resistant to extremes. To overcome this issue and make a robust estimation of the coefficient α, the midvariance and midcovariance estimates are used instead of the variances and covariances in the coefficient α formula. Such a robust estimation of the coefficient α is resistant to extreme values (outliers) in a distribution of scores.

Assumptions of Coefficient α

Before data are subject to a particular statistical procedure, such as regression analysis, we check a number of assumptions about the data. If these assumptions are violated, our conclusions from the analysis of the data will be misleading. Similarly, before test items are subject to coefficient α, researchers need to assess the assumptions of coefficient α. If the assumptions are not satisfied, the coefficient α estimated results in a misleading estimate of reliability. Coefficient α is grounded in the *essentially tau-equivalent measurement model*, often referred to as the *true score equivalent model*. The essentially tau-equivalent measurement model assumes each test item measures the same latent variable on the same test. Put another way, the items of a test should be unidimensional (i.e., test items reflect one construct). Indeed, the factor loadings (i.e., correlations between items and the factors) under the framework of confirmatory factor analysis are identical in all items on a single-factor solution, but the error variances may differ. This suggests that items are homogenous in a test. It should be emphasized even if an item in a test does not measure the underlying construct equally like other items, the sensitivity of coefficient α may be significantly undermined.

It should also be emphasized that when test items are essential tau equivalent, the reliability is equivalent to coefficient α. To assess the assumption of essential tau equivalence on the reliability of test scores, the sample of test items is subject to confirmatory factor analysis. Under the confirmatory factor analysis approach, test items should be explained by the one-factor solution (the unidimensionality assumption) in order to use coefficient α. In addition, a broad difference in the standard deviations of item responses in a test may be an indication of the violation of the assumption of tau equivalence.

Conceptual Issues Surrounding Coefficient α

The reliability of test scores is equal to coefficient α if and only if all items are essentially equivalent, otherwise coefficient α does not equal the reliability of test scores. If one factor/trait does not underlie the items on a test, the assumption of the essentially tau-equivalent model is violated, and hence, coefficient α estimates a lower bound of the reliability of test scores. Stated another way, the calculated coefficient α in samples will be considerably below average for reliability in the population if the assumption of the essentially tau-equivalent model is not met. The greater the violation of the essentially tau-equivalent model, the more coefficient α underestimates the reliability of test scores. What is more is that it has been documented that when items measure an underlying single construct (i.e., a unidimensional test), the degree of the bias in coefficient α would be trivial. In addition to this, it has also been shown that a well-constructed test with more than four items and a mean factor loading of .60 or higher is more likely to hold the assumption of essential tau equivalency and or higher is more likely to hold the assumption of essential tau equivalency and hence shows less bias.

It seems there is some confusion in the coefficient α literature on the use of the terms *internal consistency, homogeneity,* and *unidimensionality,* which are often used interchangeably. Having a clear understanding of these terms can help to improve the proper use of coefficient α. Internal consistency refers to the interrelatedness of a sample of test items, whereas homogeneity is concerned with the unidimensionality of a set of test items. Coefficient α is a function of internal consistency (the function of interrelatedness), that is, the average interitem correlation. This does not imply that the coefficient α equals the homogeneity or unidimensionality of a set of items. Both a unidimensional test and a multidimensional test can have a high or low value of coefficient α.

Mathematically speaking, the coefficient α is a function of the test length and the average interitem correlation. The coefficient α increases as the number of items in a test increases. A low coefficient α, therefore, may be due to a short test. Therefore, in order to improve the reliability of test scores of a shorter test, the length of the test should be increased. For example, a test of 40 items may have a better reliability compared with a test of 20 items. It should be emphasized that the reliability of test scores is inflated regardless of the consistency of items when large numbers of items are produced. Having a large number of items in a test is not “bad,” but the coefficient α calculated should be cross-checked against other indices of internal consistency (e.g., the average interitem correlation). The higher the average of interitem correlation, the higher the coefficient α, and the more reliable the test scores. To calculate the average interitem correlation, all items that measure the same construct in a test are correlated to each other (i.e., correlation between each pair of items) and then the computed correlation coefficients are averaged. As a guideline, it has been recommended the average interitem correlation lies in a range of .15 – .50. A higher average interitem correlation is required if a test measures a narrower construct, such as “talkativeness.”

Although the average interitem correlation is more suitable than the coefficient α, little attention has been paid to the average interitem correlation as an index of internal consistency. More importantly, if either the average interitem correlation or correlation between each pair of items is not presented, it is difficult to assess the coefficient α reported. In addition, the coefficient α is practically useless as an index of internal consistency when the number of items in a test is more than 40.

The Use of ω

Ignoring the fundamental assumptions of the coefficient ω can provide a misleading picture of the reliability of test scores. It has been well documented that these assumptions are hard to meet, such as, for example, the assumption of essential tau equivalence, which assumes all test items measure the same underlying construct (i.e., the test is unidimensional), or under factor analysis, all items have the same factor loadings. If these assumptions are violated, alternative reliability approaches to the coefficient α, such as ω (McDonald, 1999), should be used. ω has been strongly supported by the psychometric literature as an index of internal consistency, particularly as an alternative to α. Using the ω approach, the issues attached to the coefficient α, such as inflation or attenuation of internal consistency, are less likely to happen to a test.

Alternative Methods of Reliability

It has been argued that the coefficient α is not resistant to the resulting measures of reliability. Recommended robust alternatives approaches along with bootstrapping can better estimate the coefficient α. However, psychometricians believe that the coefficient α has been extensively misapplied. If the assumptions of the coefficient α are violated (e.g., the essentially tau-equivalent model), the resulting measures of reliability will be less decisive. Given the restriction assumptions that have been placed on the coefficient α (e.g., all items should have identical factor loadings), it is very likely that these assumptions are violated. If this is the case, the coefficient α should not be applied. Alternative approaches to the coefficient α, such as the ω approach, should be used.

The ω approach outperforms the coefficient α, if the assumptions of the coefficient α are violated. However, the ω would perform the same as the coefficient α, if the assumptions of the coefficient α are met. By applying the ω approach along with bootstrapping, researchers can obtain a better estimation of the reliability of test scores.

Mohsen Tavakol

See also Omega; Reliability; Split-Half Reliability; Validity

Further Readings

Cronbach, L. (1951). Coefficient alpha and the internal structure of testes. Psychometrika, 16, 297–333.

Dunn, T., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A pratical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105, 399–412.

Ebel, R. (1972). Essentials of educational measurment. London, England: Prentice-Hall International.

Graham, J. (2006). Congeneric and (Essentially) tau-equivalent estimates of score reliability. Educational and Psychological Measurement, 66, 930–944.

Lord, F., & Novick, M. (1968). Statistical theories of mental test scores. London, England: Addison-Wesley.

Nunnally, J., & Bernstein, I. (1994). Psychometric theory. New York, NY: McGraw-Hill Higher Education.

Meagan M. Patterson Meagan M. Patterson Patterson, Meagan M.

Cognitive Development, Theory of Cognitive development, theory of

306

311

Cognitive Development, Theory of

The theory of cognitive development developed by Swiss psychologist Jean Piaget (1896–1980) is one of the most influential theories in the fields of educational and developmental psychology. Piaget described his theoretical orientation as one of “genetic epistemology” focused on the emergence, growth, and evolution of knowledge. Piaget’s theory of cognitive development is premised on the notion that thinking and learning are adaptive; our cognitions allow us to adapt to and function effectively within our environments.

Starting from this fundamental premise, Piaget explored the ways in which scientific thinking and reasoning develop, and how our interactions with the physical and social world shape our thinking. Piaget, along with other early psychologists such as William James and B. F. Skinner, contributed to the evolution of psychology as an empirical, rather than a purely theoretical, science. This entry discusses the development of Piaget’s theory of cognitive development, its major constructs, the stages of cognitive development, and how the theory has been evaluated, applied in education, and built on by other researchers.

Piaget’s theory is a general theory, based on the premise that disparate aspects of cognition develop together, undergoing similar changes. In his research, Piaget explored many aspects of children’s thinking, including beliefs about the physical, biological, and social worlds. Piaget argued that cognitive development consists of a set of discrete stages and that thinking in different stages is qualitatively different. Piaget viewed cognitive development as driven by four critical factors: maturation, the physical environment, social interaction, and equilibration.

Piaget argued that children learn and understand through action. For young

Piaget argued that children learn and understand through action. For young children, this action is generally physical (e.g., grasping and manipulating objects), whereas for older children, the action may be physical or mental (performing logical cognitive actions, termed operations). In contrast to earlier theorists who generally viewed children as relatively passive recipients of instruction, Piaget thought of children as “little scientists” who were constantly developing and testing hypotheses about the world around them. Piaget believed knowledge was a process and not a state and wanted to learn not just what children knew but how they knew it. Thus, he employed research methods that allowed him to study children’s cognitive processes, including responses that might be considered incorrect or mistaken from an adult point of view.

Research Methods

Piaget used a variety of research methods in developing and testing his theory. One of the first methods he used was naturalistic observations of children (primarily his own children and the children of his friends). These observations were carefully recorded in diary entries, which included detailed descriptions of how the children interacted with the world around them. For example, in his book *The Construction of Reality in the Child*, Piaget describes an interaction with his 7-month-old son Laurent in which Laurent appears to lose interest in a desired object as soon as the object is hidden from view. Piaget used this interaction as support for his view that young infants lack an understanding of object permanence.

Piaget also frequently used clinical interviewing to study children’s thinking and reasoning. These clinical interviews typically begin with the interviewer posing a question or problem for the child to address. The interviewer then observes the child’s behavioral or verbal responses and asks follow-up questions to probe for additional information that elaborates on thought processes. For example, in *The Child’s Conception of the World*, Piaget reports on clinical interviews conducted with a series of children that focused on what it means to be alive. In these interviews, Piaget would ask children whether various organisms and objects (e.g., trees, rivers, the sun) were alive and to explain the reasoning for their conclusions. Many children viewed motion or action as key elements of being alive (e.g., concluding that rivers are alive because the water in them moves from place to place). These responses contributed to Piaget’s view of animism as a key component of young children’s thinking. Clinical interviews sometimes included specially designed tasks, such as the “three mountains task” and conservation tasks described later in this entry.

Major Constructs

Piaget’s theory was complex and far-reaching and evolved meaningfully over the course of Piaget’s long scholarly career. Major constructs of his theory are a constructivist view of knowledge, schemas, equilibration, assimilation, accommodation, operations, décalage, and stages of cognitive development.

Construction of Knowledge

Piaget’s theory of cognitive development is a constructivist theory, based on the belief that learners mentally construct an understanding of the world. These cognitive constructions are based on previous experiences as well as the learner’s current level of cognitive development. In Piaget’s view, children are not passive recipients of lessons from parents and teachers but active seekers of information and explanations for the phenomena that they observe.

Schemas

In Piaget’s theory, schemas are the fundamental building blocks of knowledge. A schema is a basic cognitive structure, an organized way of making sense of experience. Schemas provide a general way of knowing the world used for processing information and analyzing experience. For example, the schema for “chair” would include what chairs generally look like (i.e., four legs, a seat, a back) and ways in which chairs can be used (i.e., for sitting, for standing on to reach things). Our knowledge of schemas underlies all of our behaviors.

Equilibration

Equilibration is the process by which learners create a stable understanding of a construct. The desire for cognitive equilibrium is an internal motivator that drives cognitive growth and change. Cognitive structures and abilities arise from the equilibration process and are the result of the learner’s efforts to organize experiences into a coherent mental structure.

Equilibration involves movement between stages of equilibrium and disequilibrium. Typically, a learner begins in a state of equilibrium regarding a particular concept or schema. If the learner encounters new information that reveals that the learner’s understanding is inadequate, this will push the learners into a state of disequilibrium. Eventually, the learner will revise the relevant schema to incorporate this new information and reach a new state of equilibrium.

Assimilation and Accommodation

Equilibration involves balancing assimilation and accommodation. The process of assimilation involves using existing schemas to interpret the external world, incorporating new objects or experiences into the existing schema. A child who sees a guinea pig for the first time and labels it “kitty” is using an assimilation process to fit this new animal into an existing cognitive structure. Piaget viewed play as a form of assimilation, in which the child engages in familiar actions or routines, using existing behavioral and cognitive schemas, and derives pleasure from this process. In contrast, accommodation involves making adjustments in one’s cognitive organization of the world as a result of the demands of the world. In accommodation, the learner adjusts existing schemas or creates new ones when existing schemas do not capture the environment completely. The child who labels a guinea pig “kitty” is corrected, adds a new term to an existing understanding of animal categories, and is engaging in accommodation. When learners are in a state of equilibrium, they tend to assimilate more often than they accommodate, whereas when they are in a state of disequilibrium, accommodation tends to predominate.

Operations

Operations are mental representations of actions that are based on symbols and obey logical rules. With development, operations become more cognitive (less physically based) and more abstract. Operations exist in an organized mental structure in which all operations are linked together. All operations follow key logical principles, such as reversibility.

Décalage

In cognitive developmental theory, *décalage* refers to the invariant order in which cognitive skills develop. The theory differentiates between horizontal and vertical décalage. Horizontal décalage describes situations in which the child’s thinking appears to be at different cognitive levels at a given point in time. Horizontal décalage is typically demonstrated through different tasks that tap the same underlying cognitive structure. For example, if a child displays conservation ability on tasks of conservation of number and volume, but not on tasks measuring conservation of mass, this would be an example of horizontal décalage. Horizontal décalage typically refers to the ordering of cognitive accomplishments within a given developmental stage.

Vertical décalage, in contrast to horizontal décalage, describes ways in which individuals approach the same task with increasingly complex approaches over the course of development. For example, a child may develop a sensorimotor knowledge of location (e.g., knowing how to move from one room in the house to another) long before a representational knowledge of location develops (e.g., being able to draw a map of the house or give verbal directions indicating how to get from one room to another). Vertical décalage typically refers to the ordering of cognitive accomplishments across developmental stages (i.e., how a given problem is approached and solved differently across stages).

Stages of Development

Piaget’s cognitive developmental theory is a stage theory, in which the developing child progresses through multiple stages of cognitive development. Piaget proposed that these stages were universal and invariant (i.e., that one must move through each stage in order and that stages cannot be skipped) and that thinking was qualitatively different across stages. That is, it is not simply that thinking is faster, better, or more logical in later stages, but that learners in each stage approach, interact with, and conceptualize the world in meaningfully different ways. Thinking in each stage is characterized by certain cognitive accomplishments and limitations. Although Piaget indicated approximate age ranges for each stage, it is important to note that individuals of various ages may think in various ways, depending on factors such as environmental supports and task demands.

Sensorimotor Stage

In the sensorimotor stage, which lasts from birth to approximately 2 years of age, intelligence is expressed through sensory and motor capabilities. Thinking in this stage is limited, but over the course of the sensorimotor stage, children reach several key cognitive milestones. These include object permanence (the awareness that an object continues to exist even when it is not in view), the beginning of representational thought (the capacity to form mental images, as evidenced in the child's increasing language capabilities and capacity for deferred imitation), and understanding of causality.

The sensorimotor stage is divided into six substages. Movement through these substages is characterized by several overarching trends in development: movement from reflexes to goal-directed activity, transition from acting on the body to acting on the outside world, and increasing ability to coordinate multiple actions to achieve a goal. In substage 1, the infant's behaviors are limited to innate reflexes, such as sucking and grasping. In substage 2, the infant gains greater control over these reflex responses and engages in repeated physical actions (such as kicking the legs or sucking a thumb) for the enjoyment of these behaviors.

In substage 3, the infant moves beyond the bounds of the body and engages with objects in the environment. For example, a child in this substage might repeatedly bang a rattle against the floor to hear the sound that it makes. In substage 4, the infant is able to coordinate actions to reach a goal. For example, a child in this substage might crawl across a room, reach for and grasp a desired object, and use the arms and hands to place the object into the mouth. In substage 5, the infant begins to explore new possibilities with objects. In this substage, we begin to see the emergence of the "little scientist" who understands the world through trial and error. In the final substage, the beginnings of representational thought are evident. The child now has a basic understanding of using symbols (including words) to stand for objects.

Preoperational Stage

The preoperational stage, which Piaget posited as lasting from approximately 2 to 7 years of age, is characterized by increasing the use of mental representation (particularly in the realm of language) compared to the sensorimotor stage. The increasing capacity for mental representation at this stage also allows for thinking about the past and future and the development of pretend play. However, thinking in this stage tends to be rigid, inflexible, and illogical.

Preoperational stage children tend to be egocentric in their thinking (i.e., they have difficulty taking the perspective of others and often do not recognize that others do not necessarily see what they see or know what they know). Piaget demonstrated this egocentric thinking with the “three mountains task,” in which children were seated on one side of a model landscape that included a variety of objects arranged around three central mountains. The mountains blocked certain objects from view, depending on one’s location and visual perspective. Children in the preoperational stage tended to state that an experimenter seated opposite to them would see what they themselves saw, reflecting egocentric thinking. Piaget viewed this egocentrism as contributing to the lack of logical thinking at this stage—if one cannot take others’ perspectives, one cannot use those perspectives to correct logical fallacies or otherwise revise one’s reasoning.

Thinking in the preoperational stage also tends to be characterized by centration—a focus on one particular element of an object or situation and a tendency to ignore other relevant features. This cognitive centration is evident in children’s performance on conservation tasks, which measure the ability to recognize that a quantity remains the same despite a change in appearance. For example, in a task intended to measure understanding of conservation of liquid volume, a child is first shown two short, wide glasses, each containing the same amount of water. The water from one glass is then poured into a taller, thinner glass. Children in the preoperational stage will often say that the thin glass now contains more water than the wide glass, because the level of liquid in the thin glass is higher. This reflects the fact that children’s thinking was centered on one salient element (the level of the water) and did not include other relevant elements (e.g., the shapes of the two glasses).

Concrete Operational Stage

In the concrete operational stage, which includes children from approximately 7 to 11 years of age, thought is logical, flexible, and organized in its application to concrete information. Children are also able to cognitively manipulate their mental representations. Children in the concrete operational stage understand concepts such as identity, reversibility, and decentration, which allow them to perform successfully on conservation tasks. In relation to conservation tasks, identity refers to the idea that the appearance of an item can change without it changing the item’s basic nature. Reversibility refers to the concept that the effects of actions can be reversed, whereas decentration refers to the idea that a change in one aspect of an item can compensate for a change in another aspect. However, the capacity for abstract thinking is not yet present. This means that

However, the capacity for abstract thinking is not yet present. This means that children can solve a variety of logical problems, including conservation tasks, with reference to the real-world objects or situations, but are not able to solve the same types of problems in a purely logical or abstract context.

Formal Operational Stage

The last of Piaget’s cognitive developmental stages is the formal operational stage, which begins around age 11–12 years and continues into adulthood (although it is important to note that most research on the formal operational stage was conducted with participants between the ages of 11 and 15 years). This stage is characterized by the capacity for abstract, scientific thinking. In this stage, individuals can engage in mental actions applied to hypothetical properties of objects or events. Achievements of this stage include hypothetical reasoning and logical, systematic hypothesis testing. Individuals in the formal operational stage are also able to reflect on their own thinking in a logical manner. Although Piaget generally viewed progress through the cognitive developmental stages as the result of biological maturation, he did acknowledge that formal education might be necessary in order to reach the formal operational stage.

Evaluation of Theory

Piaget’s cognitive developmental theory has been influential due to a number of major strengths. First, Piaget’s theory was one of the first theories of child development to focus on cognition and cognitive processes and to examine children’s reasoning for its own sake. Piaget’s view of the child as an active seeker of knowledge, rather than a passive recipient of environmental conditioning, was appealing and influential to many researchers and educators. Piaget’s theory also had a wide scope, examining many important aspects of thinking and reasoning. Piaget also used ecologically valid methods, examining how children solved a variety of academic and social problems.

There are, however, meaningful critiques of Piaget’s approach as well. The first is that Piaget’s methods may have led to the underestimation of the capabilities of infants and young children. For example, to test for object permanence, Piaget required infants to reach for a hidden object. Later research that required only looking, rather than reaching, indicated that children may have a mental understanding of object permanence substantially earlier than Piaget’s findings suggested. Similarly, Piaget’s theory may have overestimated the cognitive abilities of those in the formal operational stage, particularly younger adolescents. Piaget’s theory is also somewhat vague regarding the mechanisms of cognitive development (e.g., there is little discussion of what is happening in a child’s mind during the process of equilibration). In addition, Piaget has been criticized for ignoring individual and cultural differences in development.

Educational Applications

The tenets of Piaget’s theory have inspired a range of educational applications and practices. These include the use of developmentally appropriate practice, in which curriculum is tailored to students’ level of cognitive and social development. Another application of Piagetian theory is the use of “hands-on” teaching methods, such as inquiry learning and the use of manipulatives.

As with Piaget’s theory in general, there have been criticisms of attempts to apply cognitive developmental theory to educational practice. For example, educational psychologist Jerome Bruner opposed Piaget’s notion of readiness. Bruner argued that teachers held students back by waiting for students to be cognitively ready for certain subject matter.

Neo-Piagetian Theories

A number of researchers have built on Piaget’s foundations to advance research and theory in the field of cognitive development. These researchers have often integrated Piagetian theory and methods with constructs from other psychological theories, such as information processing theory or dynamic systems theory. Neo-Piagetians have built on cognitive developmental theory to explore areas such as the mechanisms of change within cognitive developmental stages. This includes research on how cognitive skills and capabilities such as working memory and executive function contribute to cognitive development. Neo-Piagetian researchers continued Piaget’s tradition of exploring cognitive development across a range of domains, including thinking about the physical, biological, and social world.

Meagan M. Patterson

See also Active Learning; Adolescence; Childhood; Constructivist Approach; Educational Psychology; Kohlberg’s Stages of Moral Development; Montessori

Schools; Problem Solving

Further Readings

Beilin, H. (1992). Piaget’s enduring contribution to developmental psychology. Developmental Psychology, 28, 191–204.

Birney, D. P., & Sternber, R. J. (2011). The development of cognitive abilities. In M. H. Bornstein & M. E. Lamb (Eds.), Cognitive development: An advanced textbook (pp. 369–404). New York, NY: Taylor … Francis.

Feldman, D. H. (2004). Piaget’s stages: The unfinished symphony of cognitive development. New Ideas in Psychology, 22, 175–231.

Gruber, H. E., & Vonèche, J. J. (1995). The essential Piaget: An interpretive reference and guide. Northvale, NJ: Jason Aronson.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures. New York, NY: Psychology Press.

Lourenço, O. (2012). Piaget and Vygotsky: Many resemblances, and a crucial difference. New Ideas in Psychology, 30, 281–295.

Morra, S., Gobbo, G., Marini, Z., & Sheese, R. (2008). Cognitive development: Neo-Piagetian perspectives. New York, NY: Erlbaum.

Piaget, J. (1954). The construction of reality in the child. New York, NY: Routledge.

Piaget, J. (1959). The language and thought of the child. New York, NY: Psychology Press.

Smith, L. (1996). Critical readings on Piaget. New York, NY: Routledge.

Jingchen Liu Jingchen Liu Liu, Jingchen

Gongjun Xu Gongjun Xu Xu, Gongjun

Cognitive Diagnosis

Cognitive diagnosis

312

314

Cognitive Diagnosis

Cognitive diagnosis is a type of assessment or measurement used to identify the taxonomic group that an individual belongs to based on the individual's observed behaviors. In educational measurement (such as a test), an examinee's solutions to test problems are observed, and cognitive diagnosis provides assessment of the mastery status of the set of skills required by the test problems. This type of assessment or measurement is referred to as *diagnostic assessment* or *diagnostic measurement*. Instead of overall scores provided in traditional tests, cognitive diagnosis presents detailed assessments of the specific strengths and weaknesses in subcategory skills, based on which teachers may direct students to focused and effective future studies. After further explaining the basic concepts of cognitive diagnosis, this entry explores the components of diagnostic classification models (DCMs), examines two specific DCMs, and finally reviews empirical validation and construction of the *Q* matrix.

Concepts in Cognitive Diagnosis

There are several important concepts in cognitive diagnosis. The observed behaviors (for instance, solutions to exam problems) are often referred to as *responses* that are collected by instruments (exam problems). These instruments are called *items*. Responses to items depend either deterministically or statistically on certain characteristics of the individual (mastery status of various skills) that are often not observable. This dependence is the foundation based on which one assesses the unobserved characteristics through responses to items.

The underlying characteristics are often referred to as *attributes*. These concepts are common for most measurement models such as classical test theory, item response theory, and so on.

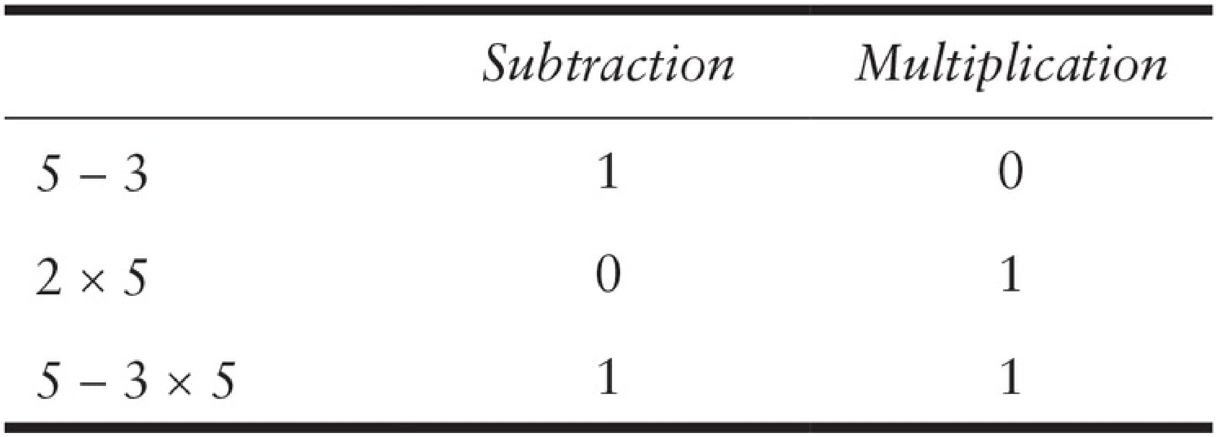
There are several features that make cognitive diagnosis distinct from other measurements. Cognitive diagnosis features a set of discrete attributes for the diagnostic purpose. More precisely, each attribute has only finitely many possible states and in fact most of the time two states: “1” for *mastery of a skill* and “0” for *nonmastery*. In case of more than two states, it is often ordinal, such as, 0, 1, 2, … standing for different skill levels. The entire attribute profile is referred to as the *knowledge state*. Cognitive diagnosis measures the attributes by means of the responses to items and casts each individual onto a knowledge state according to the assessment of individual attributes. For instance, in an arithmetic test measuring two attributes, subtraction and multiplication, each of which has two states, there are potentially four knowledge states. Moreover, cognitive diagnosis aims to provide a fine-grained assessment of the subcategory skills instead of an overall score used to rank the students from the top to the bottom. There are multiple attributes associated with a set of items, and multidimensionality is another important feature of, though not unique to, cognitive diagnosis.

DCMs

The psychometrics models for cognitive diagnosis are referred to as DCMs. Their main task is to specify the relationship between the responses to J items denoted by $r = (r_1, \ldots, r_J)$ that are directly observed and the knowledge state α that is unobserved. The knowledge state is a multidimensional vector $\alpha = (\alpha_1, \ldots, \alpha_K)$. Each α_i is discrete and takes finitely many possible values. In educational measurement, each attribute α_i often corresponds to a subcategory skill. For instance, a list of skills measured by a test of fraction subtraction may contain converting a whole number to a fraction, separating a whole number from a fraction, simplifying before subtracting, finding a common denominator, borrowing from a whole number part, column borrowing to subtract the second numerator from the first, subtracting numerators, and reducing answers to simplest form.

The responses $r_1, \ldots, r_J$ connect to the knowledge state α through the so-called Q

matrix. The *Q* matrix is a *j* by *k* matrix $Q = q_{jk}$. Each row corresponds to an item and each column corresponds to an attribute. Each entry q_{jk} takes two possible values: “1” means that *attribute k is associated to the response to Item j* and “0” *otherwise*. The following matrix is a simple and self-explanatory example of *Q* matrix for three arithmetic problems and two attributes.



	Subtraction	*Multiplication*
5 – 3	1	0
2 × 5	0	1
5 – 3 × 5	1	1

The *Q* matrix provides a qualitative description of the item–attribute relationship. The precise quantitative relationship depends on specific model parameterizations. DCMs are confirmatory in nature and the *Q* matrix specifies which items load onto which attributes.

Empirically, individuals admitting the same knowledge state may respond differently to items. Thus, the relationship between the item responses and the knowledge state is nondeterministic. DCMs specify a statistical relationship by characterizing the statistical law of the responses on each knowledge state. Technically, the model provides the conditional distribution frα.

A measurement of the knowledge state α based on the response vector **r** is obtained by the posterior distribution of α given the observed *r* via the Bayes’s rule. Measurement errors often exist due to the statistical relationship between the responses and the knowledge state and can be gradually removed as more responses are collected from the same individual.

Two DCMs

A variety of cognitive diagnostic models have been developed in the literature. The main variation among them lies in their loading structures. We present two such models and list the names of others.

Deterministic Inputs, Noisy “and” Gate (DINA) Model

The DINA model is one of the most popular DCMs, especially in educational measurement. It considers the simple case that both the responses and the attributes are binary. In particular, $r_j = 1$ represents the correct response to an exam problem and $\alpha_k = 1$ represents mastery of a skill. According to a Q matrix, suppose that Item j is associated to a number of attributes. Define the ideal response to Item j for a knowledge state α, denoted by ξ_j, as whether α has all the required attributes, equivalently, $\xi_j = 1$ if $\alpha_k \geq q_{jk}$ for all $k = 1, \ldots, K$ and $\xi_j = 0$ otherwise. The ideal response labels whether an individual on knowledge state α is capable of solving a problem and it depends deterministically on the knowledge state. The ideal response does not necessarily equal the actual response in that students may not solve the problem correctly even if they are capable of doing so and vice versa. Therefore, two additional concepts are introduced, slipping and guessing. If $\xi_j = 1$, an individual responds correctly to Item j with probability $1 - s_j$ and thus s_j is the probability of slipping; if $\xi_j = 0$, the correct response probability is g_j and that is the guessing probability.

The DINA model assumes a conjunctive relationship among multiple attributes. One needs to master all required skills to correctly solve a problem. Such a situation frequently appears in educational testing.

DINO Model

The DINO model is mathematically considered as the dual of the DINA model. Its specification is very similar to the DINA model. The difference is that the ideal response is defined as $\xi_j = 1$ if $\alpha_k \geq q_{jk}$ for at least one $k = 1, \ldots, K$ and $\xi_j = 0$ otherwise. Thus, the ideal response is constructed based on a disjunctive relationship among the attributes. Given the ideal response, the response admits the same structure as that of the DINA model. The DINO model is mathematically equivalent to the DINA model if one applies the negation operator (in Boolean algebra) to both the responses and the attributes.

Other DCMs

There are several other DCMs. An incomplete list includes the noisy input, deterministic output “and” gate model, the noisy input, deterministic output “or” gate model, the reduced reparameterized unified model, the compensatory reparameterized unified model, the additive cognitive diagnostic model, the rule space method, the attribute hierarchy method, the nonparametric clustering method, the general diagnostic model, the log-linear cognitive diagnostic model, and the generalized-DINA model.

Empirical Validation and Construction of the *Q* Matrix

Cognitive diagnosis is confirmatory in nature. The *Q* matrix specifies the set of attributes each item measures. It is customary to have a prespecified *Q* matrix based on knowledge of the items and the attributes. For instance, a teacher specifies the set of skills tested by each problem in a test. In practice, such a subjective specification may not be entirely accurate. The misspecification of the *Q* matrix could possibly lead to inaccurate and biased assessments of the knowledge state and further misleading diagnostic results. Several methods have been developed in the literature to identify possible misspecifications in the *Q* matrix, to provide means for corrections, and to empirically reconstruct the *Q* matrix based on the responses alone. The results include fundamental theories, methods, and numerical algorithms.

Jingchen Liu and Gongjun Xu

See also Classical Test Theory; Classification; Computerized Adaptive Testing; Item Response Theory; Latent Class Analysis

Further Readings

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. Psychometrika, 77(2), 201–222.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. Journal of the American Statistical Association, 110, 850–866.

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. Applied Psychological Measurement, 39(1), 5–15.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. Psychometrika, 74, 633–665.

de la Torre, J. (2011). The generalized DINA model framework. Psychometrika, 76, 179–199.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 361–390). Hillsdale, NJ: Erlbaum.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika, 74, 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement, 25, 258–272.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka’s rule-space approach. Journal of Educational Measurement, 41, 205–237.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-Matrix. Applied Psychological Measurement, 36, 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. Bernoulli, 19(5A), 1790–1817.

Liu, J., Ying, Z., & Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. Psychometrika, 80(2), 468–490.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010) Diagnostic measurement: Theory, methods, and applications. New York, NY: Guilford Press.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1), 143–157.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Journal of Educational Statistics, 12(1), 55–73.

Tatsuoka, K. K. (2009). Cognitive assessment: An introduction to the rule space method. New York, NY: Routledge.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. Psychological Methods, 11, 287–305.

von Davier, M. (2008). A general diagnostic model applied to language testing data. British Journal of Mathematical and Statistical Psychology, 61, 287–307.

Xu, G., & Zhang, S. (2015). Identifiability of diagnostic classification models. Psychometrika. doi:10.1007/s11336-015-9471-z

Michael I. Posner Michael I. Posner Posner, Michael I.

Cognitive Neuroscience Cognitive neuroscience

314

317

Cognitive Neuroscience

The advent of the ability to image the human brain has given rise to a new subdiscipline of cognitive neuroscience at the interface of neuroscience and psychology. Although neuroscience seeks to understand all brains and often chooses the simplest organism to gain the most general principles, cognitive neuroscience is centered on understanding the function of the human brain. Among the most important topics of the new field are attention, memory, and learning, all of obvious interest to the field of education.

Studies using the methods and theory of cognitive neuroscience applied to education have used various terms such as *educational neuroscience* or *brain and education,* and the journals *Mind, Brain, and Education* and *Trends in Neuroscience and Education* are central to the field. This entry outlines the methods used in cognitive neuroscience studies and gives some examples of the brain networks that have been examined and their application to education. It also examines new studies showing changes in the brain with development, looks at how exercises might enhance the brain, and considers individual differences in network efficiency.

Methods

The methods used by the field include those used to visualize the brain’s activity. The most prominent are functional magnetic resonance imaging and electroencephalography and magnetoencephalography. Functional magnetic resonance imaging has been used to localize activity in brain areas and to trace connectivity between brain areas. Electroencephalography and magnetoencephalography can measure the time course of brain activity. Transcortical magnetic stimulation has been used to disrupt localized brain circuits. Direct current stimulation of frontal areas has been used in experiments designed to increase efficiency of new learning.

Methods that rely on changes in blood flow and oxygenation are best at localization of the activity of neurons within a cubic millimeter or so, however, this still involves many neurons. Methods that record electrical or magnetic signals from outside the brain are best for the study of temporal factors for which they may indicate changes in the range of a millisecond. Combining these methods results in moderately high levels of spatial and temporal accuracy, thus allowing us to trace the location of mental computations and their time of occurrence.

Brain stimulation by transcortical magnetic stimulation can be used to temporarily inhibit the activity of a particular area, which is valuable in determining whether that area is critical to achieving the expected performance. Recently, DC stimulation across the frontal lobes has been shown to enhance some forms of learning, but questions about the limits and safety of stimulation methods remain a subject of active investigation.

Brain Networks

A quarter century of work in neuroimaging has shown that brain networks involved in cognitive tasks must often involve a small number of brain areas restricted in size and often widely scattered over cortical and subcortical brain areas.

These active brain areas are connected by bundles of axons that form pathways between them forming a network. The brain areas in the network need to be orchestrated over a few hundred milliseconds to carry out such cognitive tasks as reading words or solving arithmetic problems. Networks studied by brain imaging related to cognitive, emotional, and social processes, along with subjects learned in school, include networks involving the following:

- Attention
- Autobiographical memory
- Facial recognition
- Fear
- Music
- Object perception

- Reading and listening
- Reward
- Self-reference
- Spatial navigation
- Working memory

Networks are thought to improve in the efficiency of activation with use, thus providing a source of change or plasticity with learning.

Enhanced Teaching

Applications of cognitive neuroscience to classroom instruction have been most obvious in the fields of early reading and arithmetic instruction. For example, imaging studies of word reading have provided evidence of two important regions of activation in posterior brain areas. The first, in the superior temporal lobe, appears to underlie the phonological interpretation of visual words and is obviously related to the problem of decoding. The second lies more fully in the visual system in the left fusiform gyrus and appears to relate to “chunking” of letters into a word. The first develops early in childhood with efforts to decode written words, whereas the second appears to develop slowly with exposure to visual words of one’s language. There remains a lack of understanding of why successful decoding often fails to lead to fluent reading and debate continues within the education field over the effectiveness of phonological methods compared to the so-called look–say method of reading instruction. Understanding the brain systems involved may help teachers design an appropriate curriculum.

Imaging studies involving the processing of numbers are also important for early instruction. Studies have revealed a number line in the posterior part of the brain that allows even infants to appreciate the idea of quantity. The importance of the elaboration and connection of the number line to language and exact calculation systems provide goals for early instruction. As in reading, the brain imaging studies do not provide support for any particular curriculum, but knowledge of them may help the teacher in designing student learning.

Most of the cognitive neuroscience applications to education relate to elementary school learning. One approach to the study of higher level cognitive tasks learned in secondary school involves studies of high school algebra. Using principles of cognitive science, John Anderson and his colleagues have developed an intelligent tutoring system that has been used in 1,000 schools in the United States involving more than 500,000 students. They also conducted imaging studies to connect brain areas with some of the functions performed by the tutor.

In one study by Anderson and his colleagues, functional magnetic resonance imaging was used to study changes in brain areas following 6 days of training. The study examined six brain regions that previous studies identified as important in carrying out algebra problems. One of these areas was the anterior cingulate, which was found to be active early in problem solution and was identified as holding the subgoal used in solving the problem. The anterior cingulate operates in combination with the lateral prefrontal cortex in the storage and retrieval of declarative memories. Thus imaging may be useful in understanding how intelligent tutoring changes brain activity.

More generally, the study of expertise within cognitive neuroscience can be applied to many advanced fields. Neuroimaging has also allowed us to understand mechanisms of expertise in particular domains. Some of these domains, such as perceiving faces, are common to most or all humans, whereas others, such as reading words, are of critical importance for school. Both of these skills depend, in part, on highly specialized mechanisms within the brain's visual system.

The efficient perception of faces depends on the right fusiform gyrus (fusiform face area). In the case of words, there is an important computation involved in word recognition that depends upon the left fusiform gyrus as was discussed earlier. The visual perception of words is clearly learned, and while face perception has innate components, there is evidence that face perception differs greatly with the familiarity of the face. Moreover, improved recognition due to expertise in birds or dogs appears to modify posterior visual brain areas used for faces. The involvement of posterior brain areas in high levels of expertise may underlie the observation that the experienced person perceives the world in a different way than the novice.

Brain Development

The use of magnetic resonance to image brain networks during the resting state has provided a new window on human brain development. Although it is difficult to design tasks that can be performed at all ages, it is possible to get people of all ages, even infants, to lie passively in a scanner. Brain scanning studies indicate that several brain networks are already present in infancy.

During the resting state, a default network not usually engaged in tasks alternates in activation with portions of the brain’s attention system, which is engaged in most tasks. A protolanguage system in the left hemisphere is present even early in infancy, and the infants can show recognition of their own language and the phonemes that are constituents of all spoken languages, but in the month prior to the emergence of speech, the phonemes of the language(s) the infants hear become stronger and those of other languages weaker.

Network and State Training

Of importance to education are the claims that special kinds of training can improve brain function at all ages. The training often involves specific networks, for example, working memory, activated by cognitive tasks or computer games. No one doubts that the trained network can improve in efficiency, but the degree of transfer to other tasks is less clear. The utility of these methods may rest on better cognitive theory, which could help specify what set of tasks might help a child or adult to improve performance in daily life.

State training involves the use of very general practices such as physical exercise, mindfulness meditation, or exposure to nature that may allow a person to achieve a better overall physical and mental state that could improve performance in many situations. Certainly, physical exercise has the largest and most consistent evidence, but meditation has also been shown to produce improved attention and mood and can be applied in the school setting.

Individuality

While much of cognitive neuroscience deals with brain networks common to humans, there is recognition that individuals differ in the efficiency of these networks. Some teachers have applied the theory of multiple intelligences to deal with individual differences in the classroom. Cognitive neuroscience studies have revealed some of the mechanisms behind various form of intelligence. However, there is also evidence of brain structures that appear to underlie general intelligence, which is common among linguistic, musical, mathematical, and other forms of more specific ability.

Longitudinal studies have shown the importance of effortful control as measured from questionnaires in early childhood on the success of adults in measures of well-being, including health, income, and social interactions. Brain scanning studies have provided some evidence on the role of brain structures involved in effortful control. Moreover, exercises of the type discussed in the previous section have sometimes been useful in improving these networks, but how lasting the changes are and whether they effect daily life remains unknown. Considering the importance of temperamental characteristics such as effortful control on later life, teachers would be well advised to keep abreast of developments to better understand and help students improve effortful control.

Temperament can change in development but it is relatively fixed compared to attitudes. Research shows the importance of attitudes in school achievement. Brief interventions to make students understand that school performance depends on hard work, and not on some immutable innate ability, have been shown to improve school performance. It is important for teachers to know that mind-set or attitudes toward the educational experience can be critical in fostering achievement.

Michael I. Posner

See also *g* Theory of Intelligence; Multiple Intelligences, Theory of; Reading Comprehension

Further Readings

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York, NY: Oxford University Press.

Lambertz-Dehaene, G., & Spelke, E. (2015). The infancy of the human brain. Neuron, 88, 93–109.

Marceschal, D., Butetrworth, B., & Tomie, A. (2014). Educational neuroscience. Oxford, UK: Wiley Blackwell.

Paunesku, D., Walton, G., Romero, R., Smith, E., Yeager, D., & Dweck, C.S. (2015). Mindset interventions are a scalable treatment for academic underachievement. Psychological Science. doi:10.1177/0956797615571017

Poldrack, R. A., & Farah, M. J. (2015). Progress and challenges in probing the human brain. Nature, 526, 371–379.

Posner, M. I., & Rothbart, M. K. (2014). Attention to learning of school subjects. Trends in Neuroscience and Education. 3(1), 14–17. doi:10.1016/j.tine.2014.02.003

Rothbart, M. K. (2011). Becoming who we are. New York, NY: Guilford.

Tang, Y. Y, & Posner, M. I. (2014). Training brain networks and states. Trends in Cognitive Science, 18(7), 345–350. Retrieved from http://dx.org/10.1016/j.tics.2014.04.002

Tokuhama-Esponosa, T. (2014). Making classrooms better. New York, NY: Norton.

Brittany Murray Brittany Murray Murray, Brittany

Thurston A. Domina Thurston A. Domina Domina, Thurston A.

Andrew McEachin Andrew McEachin McEachin, Andrew

Coleman Report

Coleman report

318

320

Coleman Report

The *Equality of Educational Opportunity* (EEO) report, known as “the Coleman report” after principal investigator James S. Coleman, is widely seen as one of the most significant contributions of the 20th century to education policy and research as well as the field of sociology. The national study, commissioned by Congress under a provision in the Civil Rights Act of 1964, documented school inequality on an unprecedented scale. Attempting to survey every principal, student, and teacher in a nationally representative sample of 4,000 schools, the EEO represented a historic data collection effort in American education. The project collected cross-sectional survey and test score data from approximately 600,000 students and 50,000 teachers. This entry further discusses how the researchers conducted the study and their findings. It then looks at critiques of the study and at research that reexamined its findings.

The report sought to assess the quality of educational opportunities available to racial and ethnic student subgroups across the nation. In defining school quality, Coleman and his colleagues measured school resources such as the number of textbooks and laboratories available, curricular offerings, and academic practices such as tracking systems. Authors also took into account student and teacher characteristics. Student characteristics included measures of socioeconomic status, parent education, and peer attitudes and aspirations, while teachers were evaluated on their education, experience, salary, attitudes, and aptitude.

Student achievement measures were constructed using scores on standardized tests of pupils' verbal and nonverbal skills at the end of Grades 1, 3, 6, 9, and 12 as well as results of more traditional achievement tests in reading and mathematics at Grades 3, 6, 9, and 12 and tests of students' command of science, social studies, and other general information administered at Grades 9 and 12.

Coleman and colleagues demonstrated that American public schools remained highly racially segregated 12 years after *Brown v. Board of Education*. Further, they documented large racial achievement inequalities among U.S. public school students. The Coleman study found that the Black–White test score gap among first graders was equivalent to approximately one grade level. By the time students reached 12th grade, this achievement gap had widened to 3.5 grade levels. However, after accounting for regional differences in educational resources, Coleman showed that racial gaps in school resources were relatively small. Further, and perhaps most notably, the Coleman report cast new doubts on the extent to which schools contribute to educational inequality, showing that within-school variation in achievement was larger than between-school variation. In short, students' achievement was impacted far more by their out-of-school experiences than the instructional practices and policies of their schools.

The Coleman study found that family background played a large role in student achievement while school practices such as per pupil expenditures and teacher quality had little effect. Summarizing these findings, the report goes on to say:

> One implication stands out above all: That schools bring little influence to bear on a child's achievement that is independent of his background and general social context; and that this very lack of an independent effect means that the inequalities imposed on children by their home, neighborhood, and peer environment are carried along to become the inequalities with which they confront adult life at the end of school (Coleman et al., 1966, p. 325).

Consistent with this interpretation, recent research on summer learning loss such as that by Allison Atteberry and Andrew McEachin indicates that class inequalities narrow during the months in which schools are in session and broaden during the summer months. The Coleman report is often credited with directing educational researchers' attention toward achievement inequalities that exist among youth prior to school entry as well as the role of family practices and resources in producing educational inequality. However, evidence is mixed regarding summer learning loss and the role of schools in producing Black–White achievement gaps. Further, several studies indicate that highly effective schools can have large enough effects on student achievement to offset achievement gaps.

Reexamining the Coleman Findings

Despite its considerable influence on the field of educational research, the Coleman report remains controversial, and its findings have been closely scrutinized. Scholars have raised questions about the extent to which survey nonresponse undermined the reports’ claim to having assembled nationally representative data as well as the adequacy of the report’s regression-based methods to identify the unique consequences of educational resources. However, reanalyses largely replicate the report’s major findings.

Geoffrey Borman and Maritza Dowling reanalyzed the report’s data using multilevel modeling; they found approximately 40% of achievement variation is between schools (net of student controls), which is 3–4 times higher than the variation that Coleman and colleagues reported. This is largely because Coleman and colleagues did all analyses within individual regions and thus failed to recognize remarkably large differences between schools in different regions and particularly low performance in the South. This finding suggests bigger school effects than the report originally found. However, consistent with the report’s main finding, the Borman and Dowling reanalysis indicates that neither equalizing the school resources to which students of color are exposed nor desegregating schools are sufficient to erase Black–White test score gaps.

Other critiques focus less on the study’s analytic techniques than its interpretation. Because Coleman and colleagues collected observational data at one time point, the report’s findings are purely correlational. As such, it is impossible to separate the effects of the school resources at the center of the report’s analyses from the potentially confounding effects of student characteristics that are associated with these resources.

Because the report focuses primarily on between-school educational inequalities, it provides limited information on the extent to which educational resources are allocated unequally within schools and the extent to which these resource inequalities matter for student achievement. In particular, although Coleman and colleagues found that teachers were the most influential school input on student achievement, recent work suggests that they underestimate the magnitude of teacher effects. The study measured teacher quality in terms of teacher education and experience—two factors that loosely correlate with value-added measures—and failed to capture the sizeable degree of variation in teacher quality within schools.

These critiques notwithstanding, more than 50 years after its publication, the Coleman report remains a central touchstone in American educational research.

Brittany Murray, Thurston A. Domina, and Andrew McEachin

See also Achievement Tests; African Americans and Testing; Applied Research; *Brown v. Board of Education*

Further Readings

Atteberry, A., & McEachin, A. (2016). School's out: Summer learning loss across grade levels and school contexts in the U.S. today. In Alexander, K., Pitcock, S., & Boulay, M. (Eds.), The summer slide: What we know and can do about summer learning loss. New York, NY: Teachers College Press.

Borman, G. D., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's Equality of Educational Opportunity data. Teacher's College Record, 112(5), 1201–1247.

Bowles, S., & Levin, H. M. (1968). The determinants of scholastic achievement—An appraisal of some recent evidence. Journal of Human Resources, 3(1), 3–24.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. L. (1966). Equality of educational opportunity. Washington, DC: U.S. Government Printing Offices. (ERIC No. ED012275) Downey, D. B., & Condron, D. J. (2016). Fifty years since the Coleman report: Rethinking the relationship between schools and inequality. Sociology of Education, 89(3), 207–220.

Hoxby, C. (2016). The immensity of the Coleman data project. Education Next, 16(2). Retrieved from http://educationnext.org/the-immensity-of-the-coleman-data-project/

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., …, … Michelson, S. (1972). Inequality: A reassessment of the effect of family and schooling in America. New York, NY: Harper … Row.

Mosteller, F., Moynihan, D. P., … Harvard University. (1972). In F. Mosteller & D. P. Moynihan (Eds.), On equality of educational opportunity: Papers deriving from the Harvard University faculty seminar on the Coleman report. New York, NY: Random House.

Rebecca H. Woodland Rebecca H. Woodland Woodland, Rebecca H.

Collaboration, Evaluation of Collaboration, evaluation of

320

323

Collaboration, Evaluation of

Organizational collaboration is embraced across all sectors of society as a primary strategy for cultivating innovation, conserving economic resources, building relationships, addressing complex problems, and reaching essential outcomes. It is through collaboration that PK–12 educators address issues of teaching and student learning and accomplish organizational goals that fall outside the grasp of any individual teacher, principal, or school working independently. Although educational collaboration is widely recognized as having the capacity to connect fragmented educational systems and cultivate instructional innovation, effective collaboration does not emerge spontaneously and cannot be sustained without thoughtful attention to its development. The systematic examination and improvement of interorganizational and professional collaboration in educational settings has become imperative. This entry begins by providing a clear explanation of the two primary types of collaboration in an educational setting: interorganizational and professional. Next, the need for researchers to operationalize the concept of collaboration is discussed. Finally, the entry examines five approaches to evaluating collaboration.

What Is Collaboration?

Collaboration has become a fundamental school improvement strategy that denotes two or more people, groups, or organizations working together to reach goals that could not be accomplished by individual entities working independently. To attain essential educational outcomes, schools and school personnel increasingly work in strategic partnership with one another and with people and organizations from across all sectors of society. Interorganizational collaboration entails partnerships among schools and between schools and other agencies. An example of federally sponsored school-based interorganizational collaboration is the Safe Schools/Health Students (SS/HS) initiative. The SS/HS initiative was launched through the joint efforts of the U.S. Departments of Education, Justice, and Health and Human Services. The SS/HS initiative is based on evidence that an integrated, community-wide, collaborative approach is the most effective way to promote healthy childhood development and to address the problems of school violence and alcohol and other drug abuse. Through the SS/HS initiative, more than $2 billion in resources have been allocated to 365 communities in nearly all 50 U.S. states.

Professional collaboration, on the other hand, generally refers to the work of a single group of individuals such as a grade-level or subject-area team made up of individual teachers or other educators. These teams or groups are often referred to as professional learning communities (PLCs). PLCs have gained widespread popularity among educators over the past decade and have taken root in schools across the country because their benefits are numerous and profound. Studies show that PLCs enhance everything from teacher satisfaction to student achievement, positively impact school culture, improve teacher self-efficacy, reduce teacher isolation, boost an organization’s overall capacity, and build a shared culture of high-quality instructional practice. Collaboration, whether interorganizational or professional, is widely considered the lever through which student-, school-, and community-level outcomes will be obtained. Significant sums of federal, state, and foundation money are allocated to support the development of both interorganizational and professional collaboration. Hence, measuring, assessing, and evaluating the quality, value, and impact of organizational collaboration have become imperative.

Operationalizing the Construct of Collaboration

Although the literature in support of organizational collaboration is vast, cross-sectoral, and replete with case studies, collaboration persists as an underempiricized, misunderstood construct. Hence, evaluators who seek to examine organizational collaboration as a dependent and/or independent variable will confront the need to operationalize the concept. Take, for example, the following evaluation research questions:

- Do increases in collaboration between the county’s early intervention program and the nurse home visitation program lead to reductions in referrals to special education?

- To what extent does collaboration between teachers and school psychologists lead to better instruction and improved outcomes for student learning?
- What is the optimal level of linkage between regional educational collaboratives?
- What is the relationship between teacher collaboration, instructional quality, and student achievement?

The construct of collaboration is the central evaluand in each of the previous questions; it must be operationalized so that its development, quantity, quality, and/or effects can be measured and observed.

A synthesis of systems theory and organizational learning literature suggests that there are several observable attributes through which the construct of collaboration can be operationalized. These attributes include: (a) the *sine qua non* of collaboration is a shared purpose, partnerships form in order to improve a shared practice and to address a shared problem or issue; (b) collaboration is a nested phenomenon that takes place within complex open systems, people collaborate within teams, within organizations, and across organizational boundaries; (c) collaboration is developmental and evolves in stages over time, groups form, storm, norm, and perform; (d) collaboration within and between organizations will vary by level of integration, “more” collaboration is not necessarily better; and (e) the process of professional collaboration entails cycles of inquiry, team members dialogue, make decisions, take action, and evaluate their actions.

These five principles can be used to operationalize the construct of collaboration and guide methodological approaches in what and how to evaluate collaboration. Obtaining a clear and theoretically grounded understanding of organizational collaboration can help researchers, evaluators, and practitioners to determine and isolate the most appropriate variables and phenomena to study related to the process and outcomes of educational collaboration.

Methods of Evaluating Collaboration

The following sections will highlight five important approaches that researchers and practitioners can take to examine and improve interorganizational and professional collaboration in education.

Approach 1: Inventory and Map Teams Within a School and/or That Link the School and Its Partner Organizations

Because teams are the predominant unit for decision making in any organization, it is important to ascertain a clear and accurate picture of all the groups at work within an educational alliance. In this approach, evaluators seek answers to the following questions: What teams/groups exist in this school/organization? Who is on each team? For what purpose do the teams meet? How often and with what frequency do the teams meet?

Data about who works with whom on what in an educational setting/partnership can be gathered through a review of archival data/organizational charts, interviews with key personnel, and/or survey methods. Regardless of whether the data are collected and analyzed using simple spreadsheets and pictures or through more complex mathematical processes such as social network analyses, a systematic and thorough inventory of groups within an organization is essential and the foundation for future steps in the evaluation of collaboration. The inventory and mapping process will reveal patterns of interactions between educators within and between organizations. Findings can then be used to determine which individuals or groups might be over-and/or underextended and which teams might be too big or too small and how to target next steps in the collaboration evaluation process.

Importantly, a thorough team identification and mapping process will bring to the surface high-leverage teams (i.e., those groups that appear to focus on substantive issues related to the central goals of the partnership with the greatest capacity to precipitate positive educational change). Educational leaders can use these findings to make informed and strategic decisions about where to channel resources and offer targeted support for collaboration within schools. When conducted over time, inventory and mapping data can be correlated with other measures to determine what patterns of collaboration yield the essential outcomes. For example, inventory data can be compared to longitudinal measures of teacher retention, school climate, and/or student learning.

Approach 2: Monitor Stages of Development

A key attribute of organizational collaboration is that it will go through the

A key attribute of organizational collaboration is that it will go through the predictable stages of development: educators and their partners assemble to form groups, they work to set working norms and establish order, enact activities and perform their shared tasks, and will transform (or adjourn) their group as team goals are met and/or new members or leaders come on board. One stage may go by faster than another, a group may find itself stuck in a stage for a long time, or a team may find itself moving in and out of more than one phase at a time—but inevitably, collaboration requires its members to navigate and emerge from each stage of development in order to successfully implement tasks and reach educational outcomes.

Monitoring collaboration can stimulate groups' successful movement through the stages of development and will promote organizational performance. One effective monitoring strategy is to conduct interviews with members of high-leverage teams that have been identified through the inventory and mapping process in Approach 1. Interviews can be used to identify issues about collaboration quality related to each stage of development to be isolated for special attention, constructive criticism, and improvement. In the formation stage, team members might be asked to discuss their level of shared clarity about the purpose, structures, strategies, leadership, and key tasks of their interorganizational and/or professional collaboration. In transitioning from ordering to performing, evaluators could ask stakeholders how they will move from safeguarding resources and activities from external interference to strengthening the group's creative energy in pursuit of the accomplishment of its goals.

Approach 3: Assess Levels of Integration

One of the operational principles of collaboration is that there are levels of integration that can exist between and within organizations. More integration between organizations is not necessarily better—levels of integration between school and partner agencies should vary according to the purpose of their partnership. If the purpose is to share relatively simple or routine information with one another, then a fairly low level of organizational integration is necessary. On the other hand, if groups want to pool financial resources to create a new and semiautonomous organization to address a complex problem, a high degree of integration is warranted. Data about degrees of integration between organizations can inform decisions about what's working and the appropriate allocation of resources.

Evaluators can use a rubric or survey instrument to generate data about the degree of organizations' shared purposes, integrated leadership structures, and intensity of communications. Instruments such as the Levels of Organizational Integration Rubric or the Levels of Collaboration Survey describe multiple levels of organizational integration and the purposes, strategies/tasks, leadership/decision making, and communication characteristics that tend to be present at each level of integration. Evaluators can use tools such as these to facilitate a process (e.g., online and in person) through which partnership members discuss, determine, and record current and ideal levels of integration. This process can be repeated and data can be collected over time. Longitudinal data about levels of integration can be correlated with other important outcome measures such as school safety, school climate, and student learning.

Approach 4: Assess Quality of Team Process

Evaluating quality of collaboration within group such as PK–12 PLCs is paramount to their success. The instructional improvement process necessitates successful teacher team movement through a series of stages, including recognizing reality, owning the problem, determining a solution, implementing actions, and monitoring the outcomes of those actions. Assessment of teacher team cycles of inquiry can generate findings that can be used to increase team efficacy, efficiency, and effectiveness. Evaluation of professional collaboration can help educators avoid "collaboration lite," make meetings more meaningful, strengthen the collaboration skills of group members, and maximize group performance. Unless the scale and scope of a particular partnership is very limited, it is usually not feasible to evaluate the quality of collaboration in every team within or across a school or educational partnership.

Evaluators can use the inventory results (generated through Approach 1) to make decisions about which groups are high leverage and warrant an in-depth examination of their cycles of inquiry. Educational researchers and practitioners can evaluate the quality of team functioning through interviews and/or through the use of rubrics and survey tools that measure the behavioral and observable attributes of rigorous school-based educator teams. One such instrument is the Teacher Collaboration Assessment Survey, which explicates the characteristics of dialogue, decision making, action, and evaluation at three levels of quality. The Teacher Collaboration Assessment Survey and other valid and reliable measurement instruments for assessing the effectiveness of PLCs can be found through the Center for Effective School Practices at Rutgers University's through the Center for Effective School Practices at Rutgers University's forthcoming searchable online database.

Approach 5: Measure Outcomes of Collaboration

Educational researchers can employ a range of qualitative and/or quantitative methods to investigate the extent to which or ways in which interorganizational or professional collaboration leads to substantive improvements in teaching and learning and/or the attainment of school goals. For example, correlational and multiple regression analyses could be used to investigate the relationship between the quality of collaboration, changes in teachers' instructional practice, and student achievement on annual standardized assessments. In-depth case studies could be used to examine how community partnerships influence the social–emotional and behavioral health of children. Social network analyses could be used to ascertain how strength of ties between administrators affect school level of implementation of a district-wide curricular initiative.

Educational collaboration has emerged as one of the nation's most widely implemented strategies for improving instruction and PK–12 student learning outcomes. Studies suggest that effective interorganizational and professional collaboration can enhance everything from school safety and teacher satisfaction to student engagement and performance. Hence, the systematic examination of educational collaboration is an important undertaking for educational researchers and practitioners.

Rebecca H. Woodland

See also Applied Research; Surveys

Further Readings

Carolan, B. V. (2014). Social network analysis and education. Thousand Oaks, CA: Sage.

Daly, A. J. (Ed.) (2010). Social network theory and educational change. Cambridge, MA: Harvard Education Press.

Frey, B., Lohmeier, J., Lee, S. W., & Tollefson, N. (2006). Measuring collaboration among grant partners. American Journal of Evaluation, 27(3), 383–392.

Gajda, R., & Koliba, C. (2007). Evaluating the imperative of inter-personal collaboration: A school improvement perspective. American Journal of Evaluation, 28(1), 26–44.

Gajda, R., & Koliba, C. (2008). Evaluating and improving the quality of teacher collaboration: A field-tested framework for school leaders. NASSP Bulletin, 92(2), 133–154.

Woodland, R. H. (2016). Evaluating PK-12 professional learning communities: An improvement science perspective. American Journal of Evaluation, 37(4), 505–521.

Woodland, R., & Hutton, M. (2012). Evaluating organizational collaborations: Suggested entry points and strategies. American Journal of Evaluation, 33(3), 366–383.

Liliana Rodríguez-Campos Liliana Rodríguez-Campos Rodríguez-Campos, Liliana

Rigoberto Rincones-Gómez Rigoberto Rincones-Gómez Rincones-Gómez, Rigoberto

Collaborative Evaluation

Collaborative evaluation

323

326

Collaborative Evaluation

Collaborative evaluation is a type of evaluation in which there is a substantial degree of collaboration between evaluators and stakeholders in the evaluation process to the extent that they are willing to be and capable of being involved. Collaborative evaluators are in charge of the evaluation, and they create an ongoing engagement between evaluators and stakeholders, contributing to stronger evaluation designs, enhanced data collection and analysis, and results that stakeholders understand and use. This entry further defines collaborative evaluation, presents a model for collaborative evaluations (MCEs), and discusses considerations when performing collaborative evaluation.

Collaborative evaluation is an approach that offers many advantages, including access to information, quality of information gathered, opportunities for creative problem solving, receptivity to findings, and the use of evaluation results. From a broad perspective, collaborative evaluation belongs to the *use* branch of the evaluation theory tree described by Marvin Alkin, concerned with enhancing evaluation use through stakeholder involvement. Through collaborative evaluation, it is possible to achieve a holistic learning environment by understanding and creating collaborative opportunities. In such an environment, stakeholders better understand the evaluation process and are therefore more likely to use its findings.

Collaborative evaluation has grown in popularity, bringing together evaluators and stakeholders from different sectors, disciplines, and cultures to exchange knowledge on how collaboration can be used as a strategic tool for fostering and strengthening evaluation practice. The literature about collaborative evaluation has increased in both quantity and quality, providing an opportunity for others to gain insights about this approach. One of the first related journal articles was “Researcher as Participant: Collaborative Evaluation in a Primary School” by Edward Booth, published in 1987. At the time, this entry was written, databases key word search with “collaborative evaluation,” either in the title or in the abstract of the journal article, yielded a wide variety of titles appearing in evaluation journals, such as the *American Journal of Evaluation, International Journal of Assessment and Evaluation, Journal of Evaluation and Program Planning, Journal of MultiDisciplinary Evaluation, New Directions for Evaluation,* and *Studies in Educational Evaluation*. In addition, a number of books have made original contributions to the development of collaborative evaluations; some of these are listed in the Further Readings section at the end of this entry.

The steady maturation of collaborative evaluation has been shown as well by the contributions of national and international evaluation associations. For example, in 1995, the American Evaluation Association created the Collaborative, Participatory, and Empowerment Topical Interest Group. Since then, interest in collaborative evaluation has grown, as evidenced by the increasing number of presentations made every year at the American Evaluation Association conference. There has also been an increase in collaborative evaluation presentations at conferences around the world. This evaluation approach also has a growing number of supporters and has benefited immensely from feedback.

In an effort to facilitate understanding of this approach, some authors have structured a collection of comprehensive frameworks that outline the elements of collaborative evaluation while being grounded in the American Evaluation Association’s *Guiding Principles for Evaluators* and within the evaluation literature. These conceptual frameworks have emerged from those authors’ working experience and have been especially useful for novice evaluation practitioners in trying to understand how others view and apply collaborative efforts in a variety of settings.

MCEs

The MCEs, created by Liliana Rodríguez-Campos and Rigoberto Rincones-Gómez, revolves around a set of six interactive components specific to conducting a collaborative evaluation. Additionally, each of the MCE subcomponents includes a set of 10 steps suggested to support the proper understanding and use of the model (e.g., when and how the various elements need to be used). Even though the MCE could create an expectation of a sequential process, it is a system that incorporates continuous feedback for redefinition and improvement in which changes in one element affect other parts of the model.

The six MCE components are (1) identify the situation (the combination of formal and informal circumstances determined by the relationships that surround the evaluation), (2) clarify the expectations (the assumption, belief, or idea about the evaluation and the people involved), (3) establish a collective commitment (compromise to jointly meet the evaluation obligations without continuous external supervision), (4) ensure open communication (process of social interaction used to convey information and exchange ideas to influence specific evaluation actions), (5) encourage effective practices (sound established procedures or systems for producing a desired effect in the evaluation), and (6) follow specific guidelines (principles that direct the design, use, and assessment of the evaluation, evaluators, and collaborators). The MCE has contributed to greater conceptual clarity of the collaborative evaluation approach and has been used as a part of a wide variety of efforts, both in single-and mutiple-site evaluations, across several sectors, and for both formative and summative purposes.

Considerations When Using Collaborative Evaluation

The optimal use of collaborative evaluation requires awareness of its strengths and weaknesses and any potential opportunities and threats along the path of implementation.

The objectivity of collaborative evaluation and other stakeholder approaches has occasionally been questioned because evaluators and stakeholders bring their own experiences and views, which may affect the evaluation, and because some individuals could potentially bias findings in order to secure positive (or negative) evaluation results. In order to protect the credibility of the evaluation, care must be taken when determining what role everyone will play in the effort. In any case, the benefits gained (e.g., in staff cooperation, quality of information gathered, and receptivity to findings) can outweigh the potential difficulties that may ensue.

Individuals usually assume responsibility only for the distinct part of a project on which they work. However, effective groups assume responsibility for the entire project and develop an appreciation of the nuances in all aspects of their work. With a collaborative approach, evaluators can help understand and account for the nature of the work and the full range of stakeholders in the evaluation process.

A collaborative evaluation facilitates the engagement of key stakeholders and improves the odds that the evaluation results will provide a useful basis for guiding a decision-making process that takes into account the evaluand and its interactions within its total system. Thus, the evaluation results are able to provide a useful basis for guiding the decision-making process because people work collaboratively while respecting the evaluand and its interactions within its total system. Collaborative evaluation can lead to an increased quality of information and receptivity to findings.

Liliana Rodríguez-Campos and Rigoberto Rincones-Gómez

See also Action Research; Collaborative Evaluation; Data; Data Mining; Evaluation; Focus Groups; Interviews; Member Check; Participatory Evaluation; Qualitative Data Analysis; Qualitative Research Methods; Quantitative Research Methods; Trustworthiness; Validity

Further Readings

Alkin, M. (2012). Evaluation roots: A wider perspective of theorists’ views and influences. (2nd ed.). Thousand Oaks, CA: Sage.

Arnold, M. E. (2006). Developing evaluation capacity in extension 4-H field faculty: A framework for success. American Journal of Evaluation, 27, 257–269.

Bledsoe, K. L., & Graham, J. A. (2005). The use of multiple evaluation approaches in program evaluation. American Journal of Evaluation, 26, 302–319.

Cousins, J. B., Donohue, J. J., & Bloom, G. A. (1996). Collaborative evaluation in North America: Evaluator's self-reported opinions, practices, and consequences. Evaluation Practice, 17(3), 207–226.

Cousins, J. B., Whitmore, E., & Shulha, L. (2013). Arguments for a common set of principles for collaborative inquiry in evaluation. American Journal of Evaluation, 34, 7–22.

Davies, A., Cameron, C., Politano, C., & Gregory, K. (1992). Together is better: Collaborative assessment, evaluation … reporting. Winnipeg, CA: Peguis.

Fetterman, D. M., Rodríguez-Campos, L., Wandersman, A., & O'Sullivan, R. (2014). Collaborative, participatory and empowerment evaluation: Building a strong conceptual foundation for stakeholder involvement approaches to evaluation. American Journal of Evaluation, 35(1), 144–148.

Fetterman, D. M., & Wandersman, A. (2007). Empowerment evaluation: Yesterday, today, and tomorrow. American Journal of Evaluation, 28, 179–198.

Gajda, R. (2004). Utilizing collaboration theory to evaluate strategic alliances. American Journal of Evaluation, 25, 65–77.

Gibson, J. L., Ivancevich, J. M., & Donnelly, J. H. (2008). Organizations: Behavior, structure, processes (13th ed.). Burr Ridge, IL: McGraw-Hill.

Green, B. L., Mulvey, L., Fisher, H. A., & Woratschek, F. (1996). Integrating program and evaluation values: A family support approach to program evaluation. American Journal of Evaluation, 17, 261–272.

Jurmo, P., & Folinsbee, S. (1994). Collaborative evaluation: A handbook for workplace development planners. Don Mills, CA: ABC Canada.

Morabito, S. M. (2002). Evaluator roles and strategies for expanding evaluation process influence. American Journal of Evaluation, 23, 321–330.

O’Sullivan, R. G. (2004). Practicing evaluation: A collaborative approach. Thousand Oaks, CA: Sage.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. American Journal of Evaluation, 29, 443–459.

Rodríguez-Campos, L. (2005). Collaborative evaluations: A step-by-step model for the evaluator. Tamarac, FL: Llumina Press.

Rodríguez-Campos, L. (2008). Evaluaciones colaborativas: Un modelo paso a paso para el evaluador [Collaborative evaluations: A step-by-step model for the evaluator]. Tamarac, FL: Llumina Press.

Rodríguez-Campos, L. (2012a). Advances in collaborative evaluations. Journal of Evaluation and Program Planning, 35(4), 523–528.

Rodríguez-Campos, L. (2012b). Stakeholder involvement in evaluation: Three decades of the *American Journal of Evaluation*. Journal of MultiDisciplinary Evaluation, 8(17), 57–79.

Rodríguez-Campos, L. (2015). Collaborative evaluations in practice: Insights from business, nonprofit, and education sectors. Scottsdale, AZ: Information Age.

Rodríguez-Campos, L., & O’Sullivan, R. (2010, November). Collaborative evaluation Essentials: Highlighting the essential features of collaborative evaluation. Paper presented at the American Evaluation Association Conference, San Antonio, TX.

Rodríguez-Campos, L., & Rincones-Gómez, R. (2013). Collaborative evaluations: Step-by-step (2nd ed.). Stanford, CA: Stanford University Press.

Ryan, K., Greene, J., Lincoln, Y., Mathison, S., & Mertens, D. M. (1998). Advantages and challenges of using inclusive evaluation approaches in evaluation practice. American Journal of Evaluation, 19, 101–122.

Sanders, J. (2005). *Foreword*. In L. Rodríguez-Campos (2005). Collaborative evaluations: A step-by-step model for the evaluator. Tamarac, FL: Llumina Press.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). Evaluation theory, models, and applications. San Francisco, CA: Wiley.

Veale, J., Morley, R., & Erickson, C. (2001). Practical evaluation for collaborative services: Goals, processes, tools, and reporting systems for school-based programs. Thousand Oaks, CA: Corwin Press.

Yeh, S. S. (2000). Improving educational and social programs: A planned variation cross-validation model. American Journal of Evaluation, 21, 171–184.

Clive R. Boddy Clive R. Boddy Boddy, Clive R.

Collage Technique

Collage technique

326

327

Collage Technique

In art, a collage is a collection or combination of artwork. In research, the generation of a collage is used as an enabling or projective technique to facilitate the discussion of, and therefore the understanding of, a research topic. Collage construction has also been used as a teaching technique in order to teach the organizational culture topic to business students.

When the technique is used in research, participants would typically be asked to think about a topic prior to an interview or focus group discussion and to collect images or objects that express their feelings and attitudes toward the topic of discussion. Participants bring these images to the research exercise and use them to form a collage that is then discussed. Research participants come to the research with considered opinions because they have been thinking about the topic beforehand in their collection of visual stimuli. This facilitates discussion.

Alternatively, research participants can be given a set of magazines and newspapers when they arrive at the research exercise and then be asked to look through them to choose images or articles that correspond with how they view or understand a topic. Research participants could also be shown previously made collages and asked to what extent the images still represented their understanding of the topic under research.

The advantage of the collage technique is that it stimulates the nonrational areas of the brains of research participants because it entails the use of visual and often emotionally meaningful imagery. The technique is thought to access a deeper and broader understanding than rational questions alone would generate.

Furthermore, the use of exciting visual imagery can be different and enjoyable for research participants, and so can stimulate more animated and insightful discussion than might otherwise occur. For example, if an educational institution wanted to discern how its alumni viewed it, then previous students of the institution could be asked to attend a research session and to bring any pictures they found in magazines or newspapers that exemplify how they feel about the institution. For some establishments, the images collected may be of warmth (open fires), friendliness (smiling people), and enjoyment. For other institutions, the images presented may be of coldness (fridges, icebergs), avariciousness (open, empty wallets), and indifference. The researcher would use the images as a stimulus to generating discussion about the institution concerned and what the images collected meant to the research participants.

Clive R. Boddy

See also Bubble Drawing

Further Readings

Boddy, C. R. (2005). Projective techniques in market research: Valueless subjectivity or insightful reality? A look at the evidence for the usefulness, reliability and validity of projective techniques in market research. International Journal of Market Research, 47(3), 239–254.

Colakoglu, S., & Littlefield, J. (2011). Teaching organizational culture using a projective technique: Collage construction. Journal of Management Education, 35(4), 564–585.

Haire, M. (1950). Projective techniques in marketing research. Journal of Marketing, 14(5), 649–656.

Malchiodi, C. A. (Ed.). (2011). Handbook of art therapy. New York, NY: Guilford Press.

Powell, L., & Faherty, S. L. (1990). Treating sexually abused latency aged girls: A 20 session treatment plan utilizing group process and the creative arts therapies. The Arts in Psychotherapy, 17, 35–47.

White, M., & Epston, D. (1990). Narrative means to therapeutic ends (1st ed.). New York, NY: Norton.

Eugene T. Parker Eugene T. Parker Parker, Eugene T.

College Success

College success

327

329

College Success

Factors such as retention, persistence, degree attainment, and grade point average are common measures of college success; however, in the field of higher education, there exists ambiguity regarding the meaning and definition of college success and how to operationalize and measure college success. There exists a lack of a standard and clear definition of college success in higher education because the research on college students is principally an amalgamation of scholarship across several fields, such as education, psychology, and other social sciences. Many theoretical perspectives, frameworks, and conceptualizations have contributed to present-day notions of college success. This entry describes the conceptualization of college success and the outcomes that have traditionally been considered to represent college success. It also describes the factors that higher education professionals, faculty, and researchers have typically contended are predictors of college success.

Conceptualization of College Success

The ways that student success have been conceptualized and measured have varied across higher education, however higher education professionals and scholars have largely focused on persistence and retention. Higher education scholars, faculty, and professionals utilize a wide variety and combination of measures to assess college success and the choice of a measure typically is at the discretion of the researcher. For instance, some studies may simply examine students who return for their second year of college as a proxy for student success. Degree attainment is commonly utilized as another measure to assess student success. Finally, academic grades (i.e., grade point averages) are typically considered to be a suitable measure of college success.

College success is often associated with retention and persistence from the first year to the second and subsequently to the fourth (or sixth) year of college. There are several seminal theories related to how and why students persist in college. Vincent Tinto theorized that students who socially integrate into the campus community become more committed to the institution and are more likely to graduate. Tinto’s student integration model indicates that students’ background characteristics are associated with their capacity to be socially or academically integrated into (or engaged with) their institution’s environment and culture. Other models, such as John Bean’s model of attrition and Alexander Astin’s framework of involvement, are similar to prior notions of persistence but also account for matters such as the influence of the campus environment and peer interactions.

The first year of college serves as a critical transition period for college students. Many in the higher education field consider this to be a pivotal time for students transitioning from secondary schooling to advanced forms of learning. Intellectual and cognitive developments are critical elements of this period in their college experience. Additionally, success during students’ first year functions as a crucial indicator of their ensuring college career. Therefore, many in higher education utilize first-year retention as a measure of college success.

Academic performance during college, as measured by cumulative grade point average, traditionally serves as a representation of college student success. College grades are usually correlated with retention and subsequent degree completion. Academic performance is also recognized as a proxy for college success because grade point average has been positively linked to postcollege outcomes, such as increased aspirations to obtain an advanced degree and academic achievement in graduate and professional programs. Grades during the latter years of college serve as a predictor of college success.

Student success may represent all of the other experiences that students have in college as well as the skills and competencies that students attain on campus that positively impact students’ postcollege existence. Student success represents the collegiate experiences and competencies that subsequently can influence students’ career and professional decisions. It also represents the experiences that students encounter in college that might influence nonacademic or career outcomes, such as civic engagement or volunteerism. Some of these skills and competencies might be participation in internship programs, publication of competencies might be participation in internship programs, publication of articles through undergraduate institutional research programs, completion of supplemental and voluntary certificates, or participation in cultural or leadership programs.

Predictors of College Success

Grade point average and scores on standardized tests such as the ACT and SAT during high school or during the early years of college historically have been deemed to be strong predictors of student success, primarily regarding early performance. High school grade point average has, however, been considered to be an inadequate predictor of college success because of disparate grading systems across secondary schools. The differing characteristics of high schools make it problematic to compare students from varying high schools, and thus grade point average is an ineffectual predictor of subsequent college success. There also has been criticism of the notion that grade point averages during the first or second year of college are a predictor of subsequent college success because of potential grading differences by faculty in individual courses and the differences in grading practices across institutions of higher education.

Student success is associated with students’ academic and social integration while in college. There are psychological and behavioral factors that promote student success in college. Often, the psychological factors embody constructs such as students’ self-reported satisfaction, motivation, self-confidence, and stress management. Students also demonstrate behavioral attributes that can impact their college success. These behavioral aspects are represented by students’ academic and social engagements such as setting aside time to study, joining student organizations and clubs, attending cultural workshops, or participating in institutional programs and activities. While there is mixed evidence for most of these predictors of student success, effective time engagement has been shown to be positively associated with college success.

Demographic characteristics and personal and family background have traditionally been an area of focus for higher education professionals and researchers regarding college success. Factors such as race, gender, socioeconomic status, and first generation status are often considered to be significant influences on the capacity of students to persist, attain a degree, or have higher academic outcomes. Typically, these factors are suitable because of the availability of data. For instance, student affairs professionals can utilize data from students’ admissions applications that represent race or gender to examine the association between demographic characteristics and college success.

Conversely, what to do with the information and results can be a challenge for college administrators primarily because they have little or no control of students’ precollege and background characteristics that might consequently affect college success. For instance, college administrators sometimes consider the link between students’ socioeconomic status and college access and academic preparedness. Yet, there exists a modest body of research that focuses on the specific experiences that promote college success for this particular group of students. Higher education professionals can only influence the college experiences of these students, not their socioeconomic status.

The organizational and institutional contexts are noteworthy themes regarding student success. Environmental conditions are important considerations regarding students’ academic achievement, persistence, and subsequent degree completion. The organizational context comprises the practices, polices, and inherent structure of students’ respective institutions. Elements of the organization or environmental context might be key organizational characteristics (e.g., prestige, size, or geographical location) or institutional resources (e.g., availability and allocation of financial support). These elements of the institution might influence students’ success.

Related to the organizational context, perspective is the notion that student engagement and activities are predictors of college success. These types of experiences can involve opportunities for peer and faculty interaction, mentoring, fraternities and sororities, and study abroad. Generally, involvement in cocurricular activities is positively correlated with college success. College student engagement also embodies academic activities such as interactions within the classroom and nonclassroom experiences that are related to the curriculum or learning growth. Some examples of these activities are speaker series, workshops, and specialized training opportunities.

The research is largely inconclusive regarding the validity of any of these predictors as true indicators of eventual college success. Scholarship focusing on what determines college success is a large, mixed, and disparate amalgamation of research studies with varying evidence about what predicts college success. This is perhaps a result of the indistinct meaning and conceptualization of the term. Still, some higher education professionals and scholars contend that higher levels of predictors, such as student engagement or academic excellence, are levels of predictors, such as student engagement or academic excellence, are positively associated with higher levels of college success.

Eugene T. Parker

See also ACT; Admissions Tests; SAT

Further Readings

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in Higher Education, 12, 155–187.

Krumrei-Mancuso, E. J., Newton, F. B., Kim, E., & Wilcox, D. (2013). Psychosocial factors predicting first-year college student success. Journal of College Student Development, 54(3), 247–266.

Pascarella, E. T., Pierson, C. T., Wolniak, G. C., & Terenzini, P. T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. Journal of Higher Education, 75(3), 249–284.

Pascarella, E. T., & Terenzini, P. T. (2005). How college affects students. In K. A. Feldman (Ed.) (Vol. 2). San Francisco, CA: Jossey-Bass.

Reason, R. D., Terenzini, P. T., & Domingo, R. J. (2006). First things first: Developing academic competence in the first year of college. Research in Higher Education, 47(2), 149–175.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Dopsychosocial and study factors predict college outcomes? A meta-analysis. Psychological Bulletin, 130, 261–288.

Martha L. Thurlow Martha L. Thurlow Thurlow, Martha L.

Common Core State Standards Common core state standards

329

334

Common Core State Standards

The Common Core State Standards (CCSS) were developed in 2010 to provide a set of consistent targets across the United States for English language arts (ELA) and mathematics learning for public school students in Grades K–12. Described as college-and career-ready standards, they represented a shift from previous individual state-defined content standards that were often deemed to be low-level standards that did not meet the needs of students in a global economy. This entry provides an overview of the development of the CCSS and then describes the ELA and the mathematics standards. It details the history of adoption and rejection of the standards by states. In addition, the assessments developed to measure achievement of the CCSS are described.

Development of the CCSS

The CCSS, developed with leadership from the Council of Chief State School Officers and the National Governors Association, were promoted as being (a) research and evidence based, (b) aligned with college and work expectations, (c) rigorous in content and application of knowledge through higher order thinking skills, and (d) internationally benchmarked. They were also described as “fewer, clearer, higher” standards that were built on the strengths and lessons of existing state standards and that could provide a roadmap for K–12 curriculum, instruction, and assessment.

Standards-based education reform began in the early 1990s, spurred by models of reform that rested on a clear designation of what students needed to know and be able to do. Driven by a quest for an education system that promoted both equity and excellence, the assumption was that all students should be taught the same content and be held to the same performance standards. States were to define the content and performance standards and then develop assessments to evaluate how students were doing in relation to the standards.

The need for something other than content standards of ELA and math defined by individual states grew out of numerous discussions among states and national organizations. The reasons behind the push for new standards were numerous. A primary concern was that the United States was falling behind other countries in terms of the knowledge and skills demonstrated by students on international assessments such as the Programme for International Student Assessment developed by the Organisation for Economic Co-operation and Development. On the 2009 Programme for International Student Assessment, 15-year-olds in more than 30 countries outperformed 15-year-olds in the United States. Beyond that was the finding that many students entering postsecondary institutions had to take remedial classes because they had not obtained the skills that they needed, even though they had earned a high school diploma. Supporters of the need for new standards cited the finding that 20% of students entering 4-year colleges and 40% of students entering 2-year community colleges had to take remedial courses.

Another reason frequently cited was that each state had set different criteria for what students needed to know and be able to do, and when students moved from one state to another, they were suddenly either way behind or way ahead of where they needed to be. Evidence of the differences across states was the performance of students in Grades 4 and 8 on the National Assessment of Educational Progress, the one measure used across all of the states. Further, states' content standards were often deemed to represent minimal skills inconsistent with the skills that would be needed in jobs of the future, where individuals would need more than basic math and ELA skills, such as communication, technical reading and writing, literacy across disciplines, and more complex mathematics.

The push for new, more rigorous standards was realized in 2010 when the Council of Chief State School Officers and National Governors Association assembled experts and practitioners in mathematics and ELA to generate a set of common standards. Development teams worked quickly to produce a set of standards that were different in many ways from the standards for mathematics and ELA that states had at that time. Rounds of feedback on the standards were conducted, and a validation team was formed to confirm that the standards were evidence based.

In the introductory materials to both the ELA and the mathematics standards, several points were made about what the standards were not intended to do. For example, it was clarified that the standards were about what students were expected to know and be able to do and not about how teachers should teach. Further, the standards did not indicate how to support students who are well below or well above grade-level expectations. Similarly, they did not define the supports appropriate for English language learners or students with disabilities, although the introductory materials made clear that the same high standards must be met by these students so that they are ready for college and careers. In addition, the standards were described as being what was most essential, not all that could or should be taught. Further, it was recognized that students needed many other skills besides ELA and mathematics skills to be ready for college and career. These other skills included, to name a few, social, emotional, and physical development, as well as strong approaches to learning.

The CCSS were portrayed from the beginning as being for all students in U.S. schools. Attention was given to both students with disabilities and English language learners, two groups that advocates had suggested may not have been considered when states developed their own standards. It was recognized that these students and likely others as well would need instructional supports to ensure that they were appropriately held to the same standards. Nevertheless, there was a commitment to the importance of these students having the opportunity to learn and meet the same high standards as other students, so that they also could access the knowledge and skills needed for their post-school lives.

CCSS of ELA

The CCSS for ELA addressed reading, writing, speaking, and listening. The change in emphasis of these standards from a narrow focus on fiction and writing about personal experiences was reflected in the title of the standards —*Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*.

When describing these standards, most professionals and teachers noted that they reflected three major instructional shifts. One shift was the inclusion of greater focus on nonfiction. The new standards represented a balance of literary and informational texts in Grades K–5 and more emphasis on nonfiction and social studies and science content in the texts in Grades 6–12. Another shift was the broadened focus on writing and speaking as well as reading. Students were expected to use evidence from one or multiple texts to support their responses in writing or verbally. The third major shift was to more complex texts and academic language from a variety of content areas.

The ELA standards were organized to reflect an interdisciplinary approach. They were presented by grade level for grades K–8 and then in grade bands for Grades 9–10 and 11–12. Further, they were divided into strands for reading, writing, speaking and listening, and language. The focus of the reading strand was on text complexity and the growth of comprehension. The focus of the writing strand was on text types, responding to reading, and research. The focus of the speaking and listening strand was on flexible communication and collaboration. The focus of the language strand was on conventions, effective use, and vocabulary. The 10 standards are (CCSI, English language arts standards, n.d., n.p.):

Standard 1

Read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusion drawn from text.

Standard 2

Determine central ideas or themes of a text and analyze their development; summarize the key supporting details and ideas.

Standard 3

Analyze how and why individuals, events, or ideas develop and interact over the course of a text.

Standard 4

Interpret words and phrases as they are used in a text, including determining technical, connotative, and figurative meanings, and analyze how specific word choices shape meaning or tone.

Standard 5

Standard 5

Analyze the structure of texts, including how specific sentences, paragraphs, and larger portions of the text (e.g., a section, chapter, scene, or stanza) relate to each other and the whole.

Standard 6

Assess how point of view or purpose shapes the content and style of a text.

Standard 7

Integrate and evaluate content presented in diverse media and formats, including visually and quantitatively, as well as in words.

Standard 8

Delineate and evaluate the argument and specific claims in a text, including the validity of the reasoning as well as the relevance and sufficiency of the evidence.

Standard 9

Analyze how two or more texts address similar themes or topics in order to build knowledge or to compare the approaches the authors take.

Standard 10

Read and comprehend complex literary and informational texts independently and proficiently.

Research and media use skills were embedded throughout the standards.

The ELA standards for Grades 6–12 were further divided into two sections to reflect the roles of educators in secondary settings. One section was for ELA and the other was for history/social studies, science, and technical subjects. The history/social studies, science, and technical subjects’ standards were not intended to replace the content standards in those areas but rather to support ELA skills related to those content areas.

The ELA standards also described the characteristics of the “literate individual” who had mastered the CCSS in reading, writing, speaking, listening, and language. These characteristics were demonstrating independence; building strong content knowledge; responding to varying demands of audience, task, purpose, and discipline; comprehending as well as critiquing; valuing evidence; using technology and digital media strategically and capably; and coming to understand other perspectives and cultures.

CCSS of Mathematics

Standards for mathematics addressed both content and practices. Three major shifts also were identified for mathematics by professionals and practitioners. A first shift was to narrow and deepen the focus in each grade to ensure that students gained conceptual understanding, skills, and fluency in procedures, and were able to apply their understanding to a wide range of problems. For example, the major focus for each grade or grade band prior to high school was as follows (CCSI, Key shifts in mathematics, n.d., n.p.):

> *K–2*: Concepts, skills, and problem solving related to addition and subtraction.
> *3*–5: Concepts, skills, and problem solving related to multiplication and division of whole numbers and fractions.
> *6:* Ratios and proportional relationships and early algebraic expressions and equations.
> *7:* Ratios and proportional relationships and arithmetic of rational numbers.
> *8:* Linear algebra and linear functions.

In high school (Grades 9–12), the standards were presented by conceptual categories rather than by grades. The conceptual categories included number and quantity, algebra, functions, modeling, geometry, and statistics and probability. According to the standards document, these categories crossed typical high school course boundaries.

The second major shift was to emphasize the need to connect learning within and across grades, so that students build on information from previous years without repeating all the instruction from the previous grade. The third shift was to increase rigor to emphasize conceptual understanding, procedural skills and fluency, and application.

The mathematics standards included a set of eight standards for mathematical practice followed by grade-specific standards from kindergarten through eighth grade and then for high school overall by topic (number and quantity, algebra, functions, modeling, geometry, and statistics and probability). The eight practice standards were (CCSI, Standards for mathematical practice, n.d., n.p.):

1. make sense of problems and persevere in solving them;
2. reason abstractly and quantitatively;
3. construct viable arguments and critique the reasoning of others;
4. model with mathematics;
5. use appropriate tools strategically;
6. attend to precision;
7. look for and make use of structure; and
8. look for and express regularity in repeated reasoning.

Connecting the mathematical practices to the mathematical content standards was presented as a goal for curricula, assessments, and professional development. The introduction to the mathematical standards suggested that the expectations that began with the word “understand” provided the opportunity to connect the practices to the content.

History of Adoption of the CCSS

Chief state school officers adopted the CCSS relatively quickly. By January 2011, fewer than 10 of the 50 states had not yet adopted the standards. In addition, several of the U.S. territories had signed on to the standards, including the District of Columbia and the U.S. Virgin Islands. By 2012, only five states had not adopted both the ELA and mathematics CCSS (Alaska, Minnesota, Nebraska, Texas, and Virginia).

The rapid adoption of the CCSS was prompted, in part, by the incentive to do so provided by a U.S. Department of Education grant program known as Race to the Top. It made available $4.3 billion in grants to states that had adopted standards that were internationally benchmarked and that prepared students for college and careers. The funds were to be used by states to transform instructional practices in line with the more rigorous college-and career-ready standards, to support teachers and leaders, to leverage data systems, and to turn around the lowest performing schools. Over time, however, with changes in chief state school officers and the intervention of state legislatures, the CCSS were rejected in several of the states that had previously adopted them. They were replaced by the states’ own standards for ELA and mathematics, albeit standards that were deemed to be consistent with college and career readiness.

Assessments of the CCSS

States that adopted the CCSS had to develop new assessments based on those standards. In another effort to provide funds to support rapid changes in the ways that states would assess the new, more rigorous standards, the U.S. Department of Education provided grants to consortia of states that would work together to develop common innovative, technology-based assessments. Two consortia of states received funds to develop comprehensive assessment systems. One consortium was called the Partnership for the Assessment of Readiness for College and Careers and the other was called the Smarter Balanced Assessment Consortium (Smarter Balanced).

With the realization that these two consortia did not cover all the assessments taken by students in a comprehensive assessment system, additional funding was provided to support other consortia. First, the U.S. Department of Education provided funds for consortia of states to develop alternate assessments based on alternate achievement standards. The two funded alternate assessment consortia were the Dynamic Learning Maps Alternate Assessment Consortium and the National Center and State Collaborative. These consortia developed assessments for students with the most significant cognitive disabilities.

Additional funding was made available to support English language proficiency assessments that were aligned to the CCSS. Two consortia received these funds: Assessment Services Supporting ELs through Technology Systems and English Language Proficiency for the 21st Century. These consortia based their assessments on standards for English language development that were aligned to the CCSS.

All of the assessments that were developed to be aligned to the CCSS were technology-based assessments that used innovative item types. Some also included classroom-based performance assessments. These new assessments were regarded by many as being more rigorous than previous state assessments, yet many of the states that initially planned to use assessments developed by one of the assessment consortia ultimately rejected them. Only 20 states and the District of Columbia planned to administer Partnership for the Assessment of Readiness for College and Careers or Smarter Balanced tests in 2016–2017, the same number as during the previous year, according to *Education Week*.

CCSS Into the Future

Despite the ups and downs of adoption and rejection of both the CCSS and the assessments based on them, these standards have had a considerable impact on U.S. public schools. When the Elementary and Secondary Education Act was reauthorized in 2015, an emphasis on college-and career-ready standards was evident. States had to adopt academic standards for mathematics, reading or language arts, and science that were aligned with entrance requirements for credit-bearing coursework in each state’s higher education system and with relevant career and technical education standards. Consistent with this emphasis, national organizations pushed for high school graduation diplomas to be based on rigorous college-and career-ready standards rather than the minimal standards that had been used in the past. The CCSS continued to influence standards even in states that avoided using the name in referring to their standards.

Martha L. Thurlow

See also Achievement Tests; Every Student Succeeds Act; Formative Assessment; Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium; Summative Assessment

Further Readings

Calfee, R. C., & Wilson, K. M. (2016). Assessing the common core: What’s gone wrong—and how to get back on track. New York, NY: Guilford.

Common Core State Standards Initiative. (n.d.). English language arts standards » anchor standards » college and career readiness anchor standards for reading. Retrieved from http://www.corestandards.org/ELA-Literacy/CCRA/R/

Common Core State Standards Initiative. (n.d.). Key shifts in mathematics. Retrieved from http://www.corestandards.org/other-resources/key-shifts-in-mathematics/

Common Core State Standards Initiative. (n.d.). Standards for mathematical practice. Retrieved from http://www.corestandards.org/Math/Practice/

Gewertz, C. (2017, February 15). National testing landscape continues to shift. Education Week. Retrieved from http://www.edweek.org/ew/articles/2017/02/15/state-solidarity-still-eroding-on-common-core-tests.html

Hess, F. M., & McShane, M. Q. (2014). Common core meets education reform: What it means for politics, policy, and the future of education. New York, NY: Teachers College Press.

Rothman, R. (2011). Something in common: The common core standards and the next chapter in American education. Cambridge, MA: Harvard University Press.

Thurlow, M. L. (2012, summer). Common Core State Standards: The promise and the peril for students with disabilities. The Special Edge, 25(3), 1, 6–8.

U.S. Department of Education. (2015). Fundamental change: Innovation in America’s schools under Race to the Top. Washington, DC: Author.

Robert D. Ridge Robert D. Ridge Ridge, Robert D.

Compliance

Compliance

334

335

Compliance

Compliance refers to investigators' obligation to abide by federal, state, and local requirements when seeking approval to conduct research with human subjects and when conducting research. Investigators are accountable to ensure that all applicable laws and regulations are adhered to, so that participants and the institution are protected from harm. Failure to comply with the terms and conditions of an approved protocol can result in the suspension of the investigator's research and possibly the suspension of all human subjects research at the investigator's institution.

Federal regulations require that federally funded research involving human subjects undergo a review for ethical propriety by an institutional review board (IRB). Most institutions where research is conducted require that all human subjects research be reviewed by an IRB, whether or not it is federally funded. An investigator is required to provide complete and accurate information regarding the study's aims, the proposed methodology, and any potential risks to participants. The investigator's qualifications are submitted as truthful evidence to conduct the research. When the IRB approves the protocol, the investigator agrees to comply with federal and IRB conditions.

Data collection cannot begin until the IRB has approved the research. Even if the research qualifies for an exemption because it involves certain categories of people (e.g., political officials), it must still be reviewed and approved by the IRB. Once the research has been approved, investigators agree to several stipulations. For example, they agree to obtain and maintain informed consent from all participants.

Consent may be provided on paper or in an electronic format and must be available for inspection by the institution or by federal regulatory agencies. In addition, investigators agree to conduct the research as proposed and to not change anything without IRB review and approval. Changes to an approved protocol require that the investigator submit a request for an amendment or modification to the study, and when the request has been evaluated and approved by the IRB, changes may be made.

During the course of the research, investigators agree to provide progress reports to the IRB and to submit periodic (typically annual) applications to obtain renewed approval for the study. If adverse events occur, such as participant injury or a breach of security or confidentiality, investigators should report the event to the IRB immediately. The IRB then evaluates the event to decide if adjustments to the protocol must be made or if the research must be suspended or ended. The investigator agrees to comply with all IRB decisions in this regard. Finally, the investigator agrees to cease data collection when the approved period for data collection has expired.

Failure to comply with any of these requirements may result in a suspension of the investigator’s authorization to conduct future research with human subjects. Because the consequences of noncompliance are serious and potentially severe, it behooves investigators to understand and comply with all research regulations.

Robert D. Ridge

See also Belmont Report; Ethical Issues in Educational Research; 45 CFR Part 46; Human Subjects Protections; Human Subjects Research, Definition of; Institutional Review Boards

Further Readings

American Psychological Association. (2010). Ethical principles of psychologists and code of conduct, including 2010 amendments. Retrieved July 11, 2016, from http://www.apa.org/ethics/code/index.aspx

Code of Federal Regulations, Title 45, Part 46: Protection of Human Subjects. Retrieved July 11, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/

Carl Francis Falk Carl Francis Falk Falk, Carl Francis

Joshua N. Pritikin Joshua N. Pritikin Pritikin, Joshua N.

Computer Programming in Quantitative Analysis Computer programming in quantitative analysis

335

340

Computer Programming in Quantitative Analysis

Computer programming in quantitative analysis refers to the process of creating computer “code”—or instructions that a computer can interpret—to automate quantitative summaries of data. Due to the advent of powerful personal computers, high-level programming languages, and the increasing availability of high-performance computing clusters, such programming is becoming increasingly used and important in both publicly funded research and private and commercial settings. Such computer programming may take many different forms depending on the purpose of the quantitative analysis. This entry provides an overview of particular use cases for computer programming—progressing from the most basic to the most complex—integrated with the introduction of programming concepts.

Basic Use Cases

Applications of computer programming vary widely depending on the purpose of the analyses and on the experience of the programmer. Writing or recording the code used for quantitative analyses can be important for the ability to reproduce an analysis, automate a series of analyses or a simulation study, or develop and test a new quantitative analysis technique.

Reproducible Analyses

Suppose a researcher generates histograms for each of two variables (*X* and *Y*) and performs a regression analysis (*Y* regressed on *X*). If questions arise about the analysis, the researcher may not always remember the bin size used to construct the histograms, whether the predictor variable was standardized when they performed the regression analysis, and so on. A record of the analysis will make it possible to recall exactly how the analysis was conducted without relying on fallible human memory. Analyses that are recorded are called *reproducible analyses*.

Statistical software designed for novice analysts, typically operated by a point-and-click user interface, do not necessarily retain an exact record of the analyses performed. However, it is often possible to persuade such software to produce the underlying code. For example, SPSS can produce “syntax,” which consists of code that can be saved as a plain text file record. This record allows one to see exactly which options were enabled or selected when the analyses were run, even if not immediately discernible to the untrained eye. The act of creating a record of instructions that can be replayed is a rudimentary example of programming. Once an analysis script is available, it is a small step for the user to edit this code by copying and pasting or changing a few variable names. Other popular all-purpose statistical packages, such as R, SAS, and STATA, are also sometimes capable of generating an analysis script from a point-and-click interface.

A typical next step is to copy results into a manuscript or report. Although the user may copy software output manually, it is increasingly possible to integrate the analysis code and narrative text. This approach is known as *literate programming* and has long been advocated by Donald Knuth, an early computer science visionary. For example, suppose the results of the aforementioned regression analysis are to become part of a publication. With authoring formats such as R Markdown or packages such as knitr, it is possible to combine the R code and manuscript text in the same file. Within an R editor such as Rstudio, a button click will run the R code, combine it with the narrative text, and generate a report that automatically displays the output of the R code and can automate generation of tables, figures, and so on. Reports can be generated in a wide variety of formats including portable document format (in conjunction with LaTeX), Microsoft Word, presentations (e.g., with Beamer), and web pages (HTML). Use of such an approach can reduce transcription errors and mislabeling of output and avoid loss of documentation regarding how the results for a figure or table were generated. The code used to run analyses resides in the same place as the text of the report, and the owner of the file can see exactly what code generated each table, figure, or other in-text values reported, while optionally hiding such code from the report for esthetic reasons. Preparation of such integrated documents requires programming investment but pays off by resulting in reproducible analysis code.

Basic Programming Concepts

At a bare minimum, writing code may require understanding how to use *variables*, *functions*, and sometimes *loops*. Although there are many different terms for these concepts, usually they exist in some form regardless of the software or programming language.

Conceptually, a *variable* is a container that can store some type of data or intermediate result. For instance, a variable may contain an integer (5), some text ("Yay for statistics!"), or something more complex such as the data set that the researcher wishes to analyze or the results of a statistical analysis. The types of variables that software or a programming language can support generally depend on the allowable data types and data structures.

A *function* (or *macro* or *subroutine*) performs a specific task. The user gives the function some input, it does something with the input, and then gives back some output. Functions generally are written for performing complex tasks that would usually take many lines of code to write, and the output of a function can be stored in a variable. It is usually good practice to write functions in as general a way as possible such that the code may be reused.

More complex use cases may require a greater level of automation. For example, suppose that the researcher wishes to perform the exact same regression analysis, but for 100 different data sets. One option would be to copy/paste existing code 100 times, each time changing the name of the data set or to use the graphical interface to painstakingly perform all 100 analyses. Such an approach is not scalable to situations in which many data sets are to be analyzed. If the researcher has carefully named the variables ($X1$, $X2$, and Y within all data sets) and carefully named the data sets (e.g., *data001.txt, data002.txt,* through *data100.txt*), it is possible to write a program to perform the same task 100 times, each time changing the name of the data set and saving the result. This task is often accomplished by writing a *loop* and can usually be done in a compact and concise way with only a few lines of code.

Finally, basic code writing may at least entail use of a style guide and the ability to *debug*. A style guide is a series of conventions that dictates how the programmer should format code, name variables or functions, and provide comments or documentation such that others can more easily understand it. If multiple programmers follow similar conventions, it facilitates the ability to quickly read another person’s code. The ability to debug is also useful for fixing mistakes or “bugs” in code. Put simply, debugging approaches let a programmer see what happens inside a function or help the programmer narrow down the various possible causes to an anomalous result.

Software Ownership, Source Code, and Programming Languages

One consideration relevant to reproducibility and more advanced use cases is the continuously evolving nature of statistical software and the software ecosystem on which it is built. New versions of software usually maintain backward compatibility, meaning that analysis scripts created with an older version of the software will continue to work properly with the new version. Occasionally, recorded analysis scripts that are intended to reproduce an analysis fail because of changes to the software ecosystem. When discrepancies arise across programs or across versions of the same program, ownership and accessibility to the source code can be important to allow the user to diagnose the cause of the discrepancy.

One approach to software stewardship is corporate ownership. For example, SPSS, SAS, and STATA are owned by corporations that exercise complete control over how the software evolves, including maintenance of compatibility. Another approach to software stewardship is known as copyleft, in which no group of people exercise exclusive control. Copyleft is a legal license that leverages copyright law to ensure that users can run, copy, distribute, study, change, and improve software themselves. For example, The R Foundation is a steward for the R statistical software but has no special legally enforceable rights over the R software. In addition to these two approaches to software ownership, there is a middle approach known as open-source software that can often be regarded as a compromise between copyleft and corporate ownership. The differences among software ownership models and which is ideal for any given situation is a controversial topic.

situation is a controversial topic.

In theory, the ability to reproduce an analysis is facilitated by copyleft or open-source software models, as users have access to the underlying code and may pinpoint the cause of a discrepancy or determine how to implement the same analysis using a different programming language or software. On the other hand, adaptation to breaks in compatibility may require expertise that is out of the reach of novice users. Commercial software developers may be more vigilant about maintaining compatibility because of paid support contracts. However, backward compatibility is not necessarily guaranteed regardless of the ownership model.

A variety of software and programming languages, each with a particular software ownership model, can be used to conduct programming in quantitative analysis. All-purpose statistical packages such as R, SAS, SPSS, and STATA offer their own idiomatic language to create macros or programs. All of these constitute high-level languages that make it relatively easy to write programs but may result in programs that do not run particularly fast. For example, although core R functions are written in C/C++ (fast, low-level languages), any new functions or code that a user types is interpreted by R rather than compiled into machine code. Low-level programming languages such as C/C++, Fortran, and Rust are typically compiled and may run faster but can require much more programming effort compared to high-level languages. Other programming languages include Java, Python, Perl, and GAUSS. It is increasingly possible for different programming languages or software to work together, such as using C/C++ or Fortran to implement functions to call from within R, execute R code from a proprietary program such as SAS or SPSS, or run and process output from proprietary programs using R.

Advanced Use Cases

Implementation of a Statistical Method

Sometimes the analysis that a researcher wishes to perform is not readily available in existing software. This happens frequently among those inventing new statistical methods but can also happen when methodological researchers publish the mathematical details of a new statistical method but do not provide suitable code. It is possible for the researcher to write code to perform the statistical analysis.

An interesting challenge can arise because of differences in the typical presentation of display mathematics in scientific papers and efficient implementation of the same mathematical idea. For example, in display mathematics, it is typical to represent a permutation using a permutation matrix:

a b c d0100100000010010=(b a d c)

However, matrix multiplication is not the best way to implement this idea in a computer, as it requires operations proportional to the square of the number of items to be permuted. The same permutation can be accomplished in linear time using a mapping (old position → new position). Programmers tasked with implementing mathematical algorithms are advised to become familiar with opportunities to reorganize equations such as the commutative and associative laws.

In other cases, a researcher uses a different programming language to implement a method in existing software. There are trade-offs to building such work on either commercial or copyleft licensed software. For example, it is possible to investigate the source code for copyleft licensed software to determine exactly what the program is doing. Such source code is typically not available for commercial software and it is not possible to directly inspect the accuracy of the mathematical formulae. On the other hand, inspecting source code is a difficult task and some may place more confidence in the correctness of particular software based on the track record or reputation of a particular developer/company or the resources that a company has for technical development and programming. Regardless, there is a potential for mistakes to be present in any software, and one should be more confident in replicating a method when independently developed programs converge on the same result.

Monte Carlo Studies and Parallel Computing

Researchers who study advanced statistical methods often do so through Monte Carlo simulation studies. For example, if data violate one of the regression analysis assumptions (e.g., the residuals are not normally distributed), it is possible to investigate the consequences of this through a simulation study. The researcher may write a program that generates a large number of data sets from a specified population with nonnormally distributed residuals, analyze each data set, and then save the results. In its simplest form, such programming may entail set, and then save the results. In its simplest form, such programming may entail little more than writing a loop. In its more complex form, the researcher may study what happens under different sample sizes, different techniques for appropriately handling nonnormal data, different true regression coefficients in the population, and so on.

If many conditions are chosen in such a factorial design and the number of replications is large, or the type of analysis is computationally intensive, additional work may be required such that the simulations can be completed in a reasonable amount of time. This may entail writing some of the code in a faster, but lower level programming language, or use of parallel computing or a computing cluster. Modern computers often have central processing units that have more than one processing core. Separate processing cores can execute instructions at the same time. High-performance computing clusters may have hundreds or thousands of computing nodes, each with multiple cores. *Parallel computing* refers to creating the computer program so that it can utilize multiple processing cores to complete the analyses in a fraction of the time it would take if only one core were used. Such computing is also useful when only a small number of analyses are to be conducted, but using a computationally intensive technique. Although parallel computing may allow fast completion of analyses, developing the code involves additional complexity.

Collaboration and Best Practices

Making newly invented statistical methods available to applied researchers generally entails a complex programming task and sometimes collaboration among many researchers or programmers. For instance, SPSS, SAS, STATA, and R (and its many R packages) were programmed by many people over many years. In addition, programs such as Mplus and flexMIRT are specialized packages that are the result of fewer researcher programmers but are at the cutting edge of latent variable modeling and measurement research in the social sciences. Copyleft-licensed software, such as R, generally welcome contributions from any researchers interested in getting involved in development. In contrast, commercial software is developed mainly by paid employees of the company that sells the software as a product. In any case, such complex programming endeavors can get unwieldy without good design of the program and some best practices that standardize how programming and tests are conducted and how documentation and changes are tracked.

Deciding on a software ownership model and clear conceptual design is essential

Deciding on a software ownership model and clear conceptual design is essential before embarking on a large-scale or complex programming problem. A conceptual design may consist of a list of stories that describe the different functions that the program is to perform (and their desired input and output) and its data structures such that the programmers will understand how the different pieces are supposed to work together. Each component of the project can then be tackled in small, manageable pieces. When the software is developed in a tightly integrated way, it can be very costly in terms of time and effort to change the original conceptual design. For example, a software package that is designed to address statistical questions using maximum likelihood may not be easily changed to address statistical questions with a Bayesian mean posterior approach. However, a modular design can help, which entails careful separation of the program into independent modules with well-defined interfaces. Changes to one module do not typically affect other modules, provided that the input/output format remains the same.

Sometimes the first draft of code for a function or component of a project is done so that it just “works,” but is not particularly efficient or modular. *Refactoring*, or the process of rewriting or restructuring code, is essential for the long-term manageability of the program. The goal of refactoring is to clean up the underlying code without adversely impacting its functionality or behavior. That is, a user’s reproducible analyses should continue to reproduce the same analysis results, while refactoring makes the underlying code more concise, easier to understand, or perform the task faster. In computer science, *technical debt* describes approximately how urgently refactoring is needed. With copyleft-licensed software, the level of technical debt can be evaluated and paid down by developers familiar with the code. In contrast, technical debt in commercial software exists in a quantity only known to the steward of that software and must be paid down by its own developers before it grows out of control. If too much technical debt accumulates, then it can be easier to start rebuilding a software product from scratch.

One way to cope with the challenge of ensuring reproducible analysis scripts is to invest in regression tests. In software engineering, regression is used to indicate something that broke that was previously working. A regression test checks whether an analysis script obtains the same answer that it did originally. For example, the correlation between two variables might have originally been computed as .4. With the release of a new version, the correlation may be computed as −.3. The analysis is no longer reproducible and a suitable regression test will alert the researcher that something is broken. Writing tests are a helpful way to discover lapses in compatibility. Often, these lapses in compatibility can be repaired by examining the release notes for suggestions or by utilization of a software support communications channel.

A final problem involves tracking changes to the underlying code and the ability for all those involved to readily access the latest version. For example, consider the confusion that may arise if a programmer found a mistake in a function, but the fix was not shared with all those involved or was overwritten by another programmer’s changes to the function. Modern version control systems (e.g., git or subversion) are often used by programmers to track changes to the underlying code, and storage of code in a repository can ensure that everyone involved has access to the latest version. Although such systems can handle very complex collaborative projects, they may also be used by individual researchers who wish to document changes to their own code.

Carl Francis Falk and Joshua N. Pritikin

See also BILOG-MG; C Programming Languages; EQS; flexMIRT; HLM; IRTPRO; LISREL; Monte Carlo Simulation Studies; PARSCALE; R; SAS; SPSS; STATA

Further Readings

Eddelbuettel, D. (2013). Seamless R and C++ integration with Rcpp. New York, NY: Springer.

Gandrud, C. (2015). Reproducible research with R and RStudio (2nd ed.). Boca Raton, FL: CRC Press.

Knuth, D. E. (2011). The art of computer programming: Vols. 1–4A. Reading, MA: Addison-Wesley.

Loeliger, J., & McCullough, M. (2012). Version control with Git (2nd ed.). Sebastopol, CA: O’Reilly Media.

Matloff, N. (2011). The art of R programming. San Francisco, CA: No Starch.

Matloff, N. (2016). Parallel computing for data science: With examples in R, C++, and CUDA. Boca Raton, FL: CRC Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical Recipes: The art of scientific computing (3rd ed.). UK: Cambridge University Press.

Wicklin, R. (2010). Statistical programming with SAS/IML software. Cary, NC: SAS Institute.

Zuur, A. F., Ieno, E. N., & Meesters, E. H. W. G. (2009). A beginner’s guide to R. New York, NY: Springer.

Steven L. Wise Steven L. Wise Wise, Steven L.

Computer-Based Testing Computer-based testing

340

344

Computer-Based Testing

In computer-based testing (CBT), computer technology is used in the administration of achievement or ability test items. Such assessments have been gradually supplanting paper-and-pencil tests in educational assessment since their introduction in the 1970s. The attractiveness of CBT lies in its potential to expand, in multiple ways, the way educational assessment is conducted. This entry provides a brief history of its development, as well as an assessment of the advantages and limitations of CBT.

A Brief History

As the capability and preponderance of computers evolved during the latter part of the 20th century and into the 21st century, so has the nature and impact of CBT evolved. In the early 1970s, computerized testing was primarily found in university research institutions, using CBT delivered through mainframe computers. In the late 1970s, advances in item response theory (IRT) led to the first research on computerized adaptive test (CAT), a particular type of CBT that is highly interactive, conducted primarily at the University of Minnesota and Educational Testing Service.

The 1980s and 1990s saw the emergence of the first operational computer-based tests. Most notably, in the 1980s, the U.S. Department of Defense developed the CAT version of the Armed Services Vocational Aptitude Battery and the Northwest Evaluation Association introduced the first adaptive testing program for U.S. school children, Measures of Academic Progress. In the early 1990s, Educational Testing Service began offering CBT versions of the Graduate Record Exam. This time period also saw the development of numerous smaller scale computer-based tests in educational settings, most of which were computerized versions of preexisting paper-and-pencil tests. The 1990s also saw the first major advances in innovative item development, as researchers began to more fully exploit the capabilities of CBT.

In the early 21st century, CBT became more common in education, as computer technology gained increasing sophistication and the availability of computers in schools increased. Moreover, the emergence of the Internet brought with it increasing expectations that CBT would be delivered online. As of 2015, CBT was successfully being used in the majority of U.S. statewide K–12 testing programs, although online large-scale CBT showed uneven reliability.

Advantages

CBT offers a set of important advantages over paper-and-pencil tests. Some of these advantages influence the quality of measurement, while others favorably influence the costs and logistics of test administration.

Innovative Item Types

CBT enables the administration of innovative item types, which have been defined as those that depart from the traditional text-based, multiple-choice format. This is a broad definition whose meaning has expanded over time, as computer technology has evolved and researchers have increasingly understood the role that computers can play in testing. Innovative items can potentially improve measurement in several ways over traditional items. First, innovative items can provide a more direct measurement of some knowledge or skill. As an example, using CBT, a test taker might be asked to identify and correct grammatical errors in a paragraph—a task that would be awkward to do using text-based, multiple-choice items. Second, innovative items can allow measurement of important parts of a content domain that would be logistically very challenging to measure at all using traditional text-based items. For instance, a test taker might listen to a piece of music and be asked to identify its tempo. Both of these examples illustrate how both the types of cognitive skills that can be measured and the ways they are measured can be enhanced when CBT is used.

Innovative items in CBT can be classified along several dimensions. One dimension is *item format*, which refers to the response possibilities of the test taker. Item formats can range from selected response (in which the test taker chooses among a set of highly defined options) to constructed-response formats (in which test takers construct their own answers). There are, however, numerous possibilities in between. For example, a test taker might be asked to drag a number of historical events steps into chronological order—which is a task that could be administered very efficiently using CBT. Another dimension is *response action*, which represents the physical actions a test taker must perform to answer an item. Response actions that have been used include radio buttons or pull-down menus, typed-in responses, joysticks, use of touchscreens, spoken responses, or answers specified through mouse clicks. However, an endless array of options is possible, constrained only by the creativity of the test developer and the capability of the computer hardware. A third dimension is *media inclusion*, which refers to the degree to which innovative CBT items incorporate multimedia elements, such as graphics, sounds, video, or animations. A fourth dimension is the *level of interactivity* between the test and the test taker. Traditional test items are completely noninteractive; the test taker provides a response action, and the item is complete. In CBT, however, the testing software can interact with the test taker by responding to the test taker entered response in some way, such as giving feedback, branching to particular follow-up items, or performing a simulation based on information specified by the test taker. Collectively, the four dimensions could define a virtually unbounded number of innovative item types that can provide tests involving a wide array of task complexity.

Adaptive Testing

Unlike a traditional test, in which a group of test takers all receive a common predetermined set of test items, a CAT selects items individually for each test taker based on the test taker responses to previous items. As a result, each test taker in a group may receive a unique set of items drawn from a larger item pool. Lower achieving test takers will receive less difficult items, while higher achievers will receive more difficult items. This raises the psychometric issue of how to compare the achievement levels of a set of students if they all take tests of different average difficulty. This is accomplished using IRT, which provides the psychometric basis for scoring a CAT. In IRT, a test taker’s score is a function of both the characteristics (e.g., difficulty) of the items administered and how the test taker did on those items. Of particular importance to a CAT is the IRT principle of item invariance, which states that a test taker’s expected score would be invariant across any subset of items administered from a larger score would be invariant across any subset of items administered from a larger set. This implies that the scores of different test takers on a CAT can be compared because they are on the same measurement scale.

Selection of items on a CAT depends on two other features of IRT. First, both the difficulty of test items and the achievement levels of test takers are represented on the same scale. Second, the closer an item’s difficulty lies to a test taker’s achievement level, the more informative it is in measuring that individual. The psychometric goal of a CAT item selection is to match item difficulty to a test taker whose achievement level is initially unknown. It accomplishes this task by calculating a provisional achievement level estimate after each scoring item response and then selecting and administering an item well matched to that provisional estimate.

The use of a CAT has two favorable outcomes. First, testing becomes materially more efficient because test takers have administered items that are tailored to their achievement levels. As a result, a general rule of thumb is that a CAT can attain measurement precision equivalent to a fixed-item test in about half as many items. Put another way, within a given testing time, a CAT can yield scores that are much more precise than those from fixed-item tests. A second outcome is that a CAT can yield scores of similar precision for all test takers. This contrasts with fixed-item tests, for which test takers in the extremes of a test taker achievement distribution are typically measured less precisely than those in the middle.

Several variations in the basic CAT model have been proposed. Notable among them are multistage tests, in which the testing algorithm adapts after item sets rather than after single items, computerized classification tests, which are designed to classify test takers into proficiency categories, and self-adapted tests, in which test takers are permitted to select the difficulty level of each item they receive.

Enhanced Administration Control and Data Collection

It is desirable to exert control over the administration of test items because it helps standardize a test event and reduces the likelihood of construct-irrelevant factors affecting the validity of inferences made on the basis of test scores. In paper-and-pencil testing, a substantial amount of standardization is possible; test instructions, timing, and the physical environment are all under the control of the test administrator. Some aspects of testing, however, are usually not controllable. Test takers can choose how they take the test by skipping items, omitting answers, reviewing items, and possibly changing previously entered answers. Such test-taking behavior is an essentially uncontrollable consequence of the use of group-based paper-and-pencil testing. With CBT, in contrast, much more administrative control is possible, and test givers can decide the degree to which they will allow test takers to control the way they answer the items. For example, it is not uncommon for CBT to require a test taker to answer an item before the user moves on to the next item or to not allow item review. Such a degree of control may or may not be desirable, however, as some test givers prefer highly controlled test administration, while others prefer to provide test takers the same amount of control as with paper-and-pencil tests (particularly if both CBT and paper-and-pencil versions of the same test are being used).

One advantage of CBT is its ability to collect more information about a test event than is available with paper-and-pencil testing. One important example is CBT’s capability to record how long test takers spend responding to individual items. Item response time has been found to be related to a test taker’s achievement level, cognitive processing speed, and test-taking engagement. Information about other behaviors such as whether test takers review their answers and how often answers were changed can provide useful insights about how people take tests, which can guide test development. Numerous other types of measures of test takers are potentially available when CBT is used, including eye tracking (which may indicate degree of test taker engagement) or biometric data (which can be useful in measuring test anxiety).

Accessibility Features

Testing software can provide a variety of accessibility features for test takers with disabilities. Moreover, many of these features can be provided in CBT with less variability than those provided by human educators. Such features include screen magnification/enlargement, text-to-speech, answer masking, and line readers.

Logistical Issues

Use of CBT can provide a number of logistical benefits to a testing program.

When objective test formats are used, CBT can immediately score a test. While this capability is necessary for adaptive testing, the ability to provide immediate scoring is desirable in all types of CBT. Whenever test takers and educators can be provided immediate feedback about test performance, instructional information becomes much more timely and actionable than is typically the case with paper-and-pencil tests. In addition, the use of CBT allows more flexibility in test administration. Standardized test administration can be maintained while testing students at different times, in different locations, and on a variety of computers/devices. Moreover, online (i.e., Internet-based) tests allow test taking to occur at home or at other locations outside of school.

A number of security concerns associated with paper-and-pencil tests are alleviated with CBT. There are no test booklets that must be kept secure and accounted for at all times by test givers. Similarly, there are no concerns about the reliability of shipping test forms. It should be noted, however, that CBT is not inherently a more secure way to test, as CBT brings with it a new set of security concerns.

Limitations

The powerful advantages of CBT should be considered relative to its limitations. In general, CBT requires more planning and resources than paper-and-pencil tests. In addition, its use can threaten the validity of inferences made on the basis of test scores by the introduction of new construct-irrelevant factors.

Higher Costs and Logistical Demands

Relative to paper-and-pencil testing, considerable up-front resources are required to develop and implement a computerized testing program. Software for administering CBT must be developed or purchased, and computerized versions of items must be developed, along with scoring capabilities. For CAT programs, the item needs are usually sizable, as a large IRT-calibrated item pool will be needed for the CAT to operate effectively. There will be an accompanying need for adequate computer hardware for administering CBT. It is likely that computers will need to be purchased, and additional hardware will be needed for connectivity if online testing is used.

After CBT has been developed, there will be an additional set of ongoing challenges associated with maintaining the program. School computer resources challenges associated with maintaining the program. School computer resources typically vary in terms of type of computers (or other computer device), processing speed and memory, operating systems (both in type and version), connectivity, and Internet browser type and version. Although the test giver may be able to specify some minimum necessary hardware and software requirements needed for delivery of the computer-based test, it will typically be the case that such a test will need to be capable of running on a variety of computer configurations. Moreover, these configurations will continue to evolve over time as newer versions of operating systems and browsers are released.

Administration of CBT brings with it a set of logistical challenges. It will often be the case that the number of students to be tested far exceeds the number of available computers, resulting in multiple testing sessions being needed throughout a testing window. This raises additional security concerns, as students already tested may pass information on to students yet to be tested. In addition, when online testing is used, testing capability is dependent on the quality of the Internet connections. These connections can be slowed or disrupted, resulting in online testing typically having an element of risk that is not present in paper-and-pencil testing.

Comparability Issues

Whenever a test is administered in different modalities, it is important that it yield scores with comparable meaning across modality. In the context of CBT, there are several types of comparability that may need to be considered, including across delivery modes (i.e., CBT vs. paper-and-pencil), as well as across different operating systems, browser types, or computer/device types. The general concern is that the way test takers interact with their test items in CBT can be affected by a variety of factors such as screen size, screen resolution, font type and size, and item display time. Different modalities can vary on these factors, and if comparability is not present, test score validity can be threatened.

Comparability poses a continual challenge for CBT programs, which have a responsibility for evaluating and ensuring comparability. Because there are many types of comparability to consider, test givers have to devote substantial attention and resources to the issue. Moreover, the compatibility issue continues to pose an ongoing challenge throughout the life of a CBT program, as new types of devices, operating systems, and connectivity continue to emerge and require new comparability studies.

Construct-Irrelevant Factors

CBT is potentially vulnerable to several additional construct-irrelevant factors that do not affect paper-and-pencil testing. Test takers with limited experience using computers at home may be at a disadvantage when taking a computer-based test. Some test takers may experience anxiety when using computers. Others may have difficulty understanding how to use the features of the computer testing software. Each of these factors may degrade test takers' ability to demonstrate what they know and can do.

Steven L. Wise

See also Achievement Tests; Computerized Adaptive Testing; Diagnostic Tests; Item Banking; Performance-Based Assessment; Technology in Classroom Assessment; Technology-Enhanced Items; Test Security; Validity

Further Readings

Bennett, R. E. (2003, October). Online assessment and the comparability of score meaning. Paper presented at the International Association for Educational Assessment Annual conference, Manchester, UK.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 367–407). New York, NY: American Council on Education.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 471–515). Washington, DC: American Council on Education.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). Innovations in computerized assessment. Mahwah, NJ: Erlbaum.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. Educational Measurement: Issues and Practice, 20(3), 16–25.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). Computer-based testing: Building the foundation for future assessments. Mahwah, NJ: Erlbaum.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical considerations in computer-based testing. New York, NY: Springer-Verlag.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. The Journal of Technology, Learning and Assessment, 4(6).

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). Computer-based testing: Practices and considerations (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. Psicológica, 21, 135–155.

Richard M. Luecht Richard M. Luecht Luecht, Richard M.

Computerized Adaptive Testing Computerized adaptive testing

344

350

Computerized Adaptive Testing

Computerized adaptive testing (CAT) is a method of sequentially selecting test items or larger test units in real time so that the final difficulty of each test form is optimally matched to the proficiency of each examinee. This tailoring of a test form by difficulty to each examinee's proficiency helps ensure an accurate final score using as few test items as possible. The statistical efficiency of a CAT is therefore realized by reduced test lengths and less testing time and/or improved score accuracy relative to a fixed-test form where every examinee is administered the same items.

Various applications of adaptive testing are found in educational achievement testing, placement, and college readiness, for a variety of psychological tests and aptitude tests, and in many types of employment and certification/licensure tests. This entry provides an overview of the assessment technologies underlying CAT as well as discussing different varieties of adaptive testing.

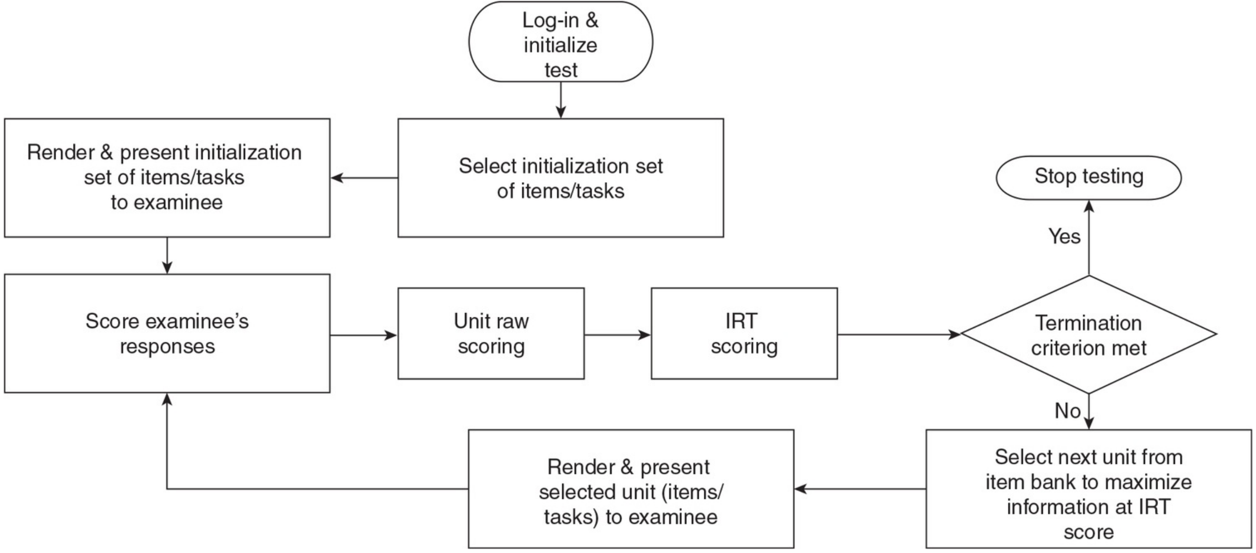
Key Features of Computerized Adaptive Tests

Most paper-and-pencil test forms comprise a fixed number of test items. These fixed test forms are typically constructed by test developers to meet a set of specifications comprising a content *blueprint* that indicates the proportional representation of content on each test form and statistical requirements such as the target test difficulty and minimum reliability (e.g., minimum score precision per form). All examinees assigned to a particular test form see exactly the same test items—usually in the same position.

CAT replaces the paradigm of using *fixed* test forms with one using *variable* test forms that are uniquely customized for each examinee. Theoretically, under

CAT, every individual can receive a test form uniquely designed to match his or her proficiency. One of the earliest adaptive testing paradigms was proposed in 1971 by Frederic Lord, who investigated flexilevel testing as a precursor to CAT. Fixed test booklets contained items arranged in order of difficulty. An examinee would start the flexilevel test in the midrange of difficulty and take easier items each time the examinee answered the current item incorrectly, or conversely, a more difficult item for each correct answer. This same principle underlies most CATs in operational use, today. Figure 1 provides an overview of a slightly more sophisticated (and modern) CAT algorithm.

Figure 1 A basic computerized adaptive testing (CAT) algorithm


Log-in & initialize test
Select initialization set of items/tasks
Render & present initialization set of items/tasks to examinee
Score examinee's responses
Unit raw scoring
IRT scoring
Termination criterion met
Yes
Stop testing
No
Select next unit from item bank to maximize information at IRT score
Render & present selected unit (items/ tasks) to examinee

As shown in Figure 1, the CAT is typically initiated by selecting and administering a small number of preselected test items to provide a preliminary estimate of the examinee's proficiency score. That preliminary proficiency score is then used to select easier or more difficult test items, similar to the logic of a flexilevel test. (The actual maximum information criterion used for optimally selecting the next item is discussed later in this entry.) The selected item is then administered, a new provisional proficiency score is estimated, and another item is selected. The process continues until a fixed test length or some other designated stopping criterion has been reached—a criterion usually related to decision accuracy or to the precision of the estimated proficiency score.

Despite the seeming simplicity of the CAT algorithm shown in Figure 1, it is important to understand that the phrase *computerized adaptive testing* actually refers to many different types of computer-based test (CBT) delivery models and technologies. Two factors tend to distinguish most types of adaptive test: (1) the size and nature of the test units selected and (2) the actual test unit selection/test form assembly and scoring mechanisms employed. In addition, there are a number of subtle variations on the theme of item-level CAT, such as Wim van der Linden’s *shadow testing*, a method of ensuring extended control over the content balance and quality of each customized test form, and *stratified CAT*, a method that attempts to buffer the overexposure of “statistically popular” items in the item bank.

Nature of the Test Units

CBT delivery and modern high-speed, high-bandwidth digital data transmission capabilities have greatly expanded the potential repertoire of item types that can be administered—going beyond multiple-choice, short-answer, and constructed response essays typically encountered on paper-and-pencil test forms. Exploiting the sophistication of modern graphical computer interfaces, multiple input devices (keyboard, mouse, touch screen, and voice recognition), and new response-capturing capabilities, a vast array of CBT item types is possible. This array includes multiple-choice and similar selected response item types, problem-based item sets, various technology-enhanced item types such as “hot spot” and “drag-and-drop” items, open-ended constructed response item types that collect text-based inputs or specific outputs such as the graph of a mathematical function, and complex work simulations involving highly interactive software applications containing drawing tools, programming and design interfaces, spreadsheets, calculators, and search engines.

In addition, almost any of these item discrete types can be combined and preassembled with additional stimulus materials and auxiliary software components to present to each examinee with larger, intact units. For example, two or more items can refer to a reading passage, a graphic or another type of stimulus material. These self-contained units are usually called *testlets*. When the testlets are adaptively selected, they can be called *testlet-based CATs*. Items and CBT performance tasks can be further combined using detailed content and statistical specifications to form preassembled modules that differ in overall difficulty—and that possibly differ in content, as well. The modules are then preassigned as self-adaptive test units called *panels*.

Adaptive Test Assembly and Scoring

As noted earlier, CAT assembles a customized test form for every examinee in real time. The items or tasks are selected from a database called an *item bank*. The item bank typically contains five types of data about each item: (1) the item text and other rendering data that is used by the testing software (called the *CBT driver*) to present the item to the test takers and capture their response; (2) the answer keys or other response scoring mechanisms; (3) item content codes, cognitive codes, and other nonpsychometric data used by the test assembly algorithm; (4) item response theory (IRT) item parameter estimates; and (5) item exposure control parameters or constraints.

The test assembly can occur in real time—that is, while the examinee is actually taking the test—or at least partially in advance of testing where preconstructed testlets or modules are prepared by the test developers and administered in the real time as larger, intact units. An adaptive test is basically an iterative three-step process (see Figure 1). First, an item (or larger test unit) is selected by an assembly algorithm. Second, the item is administered. Third, the examinee’s responses are scored and the new provisional score is used to select the next item or test unit. Key to both adaptive test assembly and scoring is an underlying metric or scale that can link the statistical difficulty and other psychometric characteristics of all of the items to the apparent performance and ultimate scoring of the examinees. That is where IRT comes into play.

Most multiple-choice or selected response items are dichotomously scored (i.e., scored correct or incorrect) and employ the one-, two-, or three-parameter logistic IRT models—often abbreviated as 1PL, 2PL, or 3PL models. These IRT models make it possible for the statistical characteristics of the items such as item difficulty to be calibrated relative to a common scale called theta (and typically represented by the Greek letter θ). For example, the more general 3PL model mathematically expresses the probability of a correct response to an item as:

$$Pr(u_i = 1 \mid \theta; a_i, b, c_i) \equiv P_i(\theta) = c_i$$
$$+(1 - c_i)\{1 - \exp[-a_i(\theta - b)\}^{-1} ,$$

where u_i = 1 is a correct binary scored item response, θ is the proficiency scale, a_i is a discrimination parameter that denotes the sensitivity of each item to the proficiency scale, b_i is the item difficulty parameter that locates each item along the proficiency scale, and c_i is a pseudo-guessing parameter that helps fit inconsistent or noisy response patterns near the lower regions of the θ. That is, the lower asymptote of the IRT probability function is governed by c_i. As the examinee's proficiency, θ, increases, the probability of a correct response, $P_i(\theta)$, increases, approaching 1.0 for examinees with very high proficiency scores. The actual item parameters are estimated using IRT calibration software. More complex IRT models are also available for item types that use polytomous scoring such as integers 0, 1, 2, …, or for testlets.

Once the items are calibrated using IRT, examinees can all be scored on the common θ scale regardless of whether they were administered an easy, moderately difficult, or difficult test form. As noted earlier, this is an essential component of any adaptive test: the capability to actually score the examinees on a common scale regardless of the difficulty of their test form. IRT conveniently provides that capability.

If the 3PL IRT model is used, the item bank will contain a_i, b_i, and, c_i parameter estimates for all $i = 1, \ldots, I$ items in the bank. If the 2PL model is used, only the a_i and b_i parameter estimates are stored (i.e., $c_i = 0.0$ for all items). The most simplistic 1PL model only uses the b_i parameter because $a_i = 1.0$ and $c_i = 0.0$ for all items. In addition to the IRT parameter estimates, the item bank may also contain item exposure control parameters that are used to restrict the overexposure of the best items, as well as various content and other coded item attributes that may be used by the CAT item selection algorithm.

In order to understand how the actual adaptive test assembly takes place, we first need to understand IRT scoring. There are three types of IRT scores: (1) maximum likelihood (ML) scores; (2) Bayes mean scores—often called *expected a posteriori* scores; and (3) Bayes modal scores—usually called modal *a posteriori* scores. For example, modal a posteriori scores can be estimated as:

$$\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}} \equiv \max_{\theta}\left\{g\left(\theta \mid u_{i_1},\ldots,u_{i_{k-1}}\right) : \theta \in \left(-\infty,\infty\right)\right\},$$

where the value of θ at maximum of the posterior likelihood function, , is the model estimate of θ.

Classification-based scores (e.g., nonmaster/master) can also be computed under certain types of IRT latent class models. A comprehensive discussion of classification scoring methods, maximum likelihood, expected a posteriori, and classification scoring methods, maximum likelihood, expected a posteriori, and modal a posteriori scores, is beyond the scope of this entry. Suffice to say, as long as the IRT item parameter estimates used for scoring are calibrated to a common scale, the ensuing estimates of θ for all examinees will be on the same scale, as well.

An adaptive test assembly or item selection algorithm uses the provisional proficiency estimate of θ as the basis for locating the optimal next item to administer to the current examinee. That is, if we let denoted the provisional proficiency score estimate based on accumulated responses for *k* − 1 items administered up to that point in the test (i.e., for raw-scored item responses), then the next item or test unit is then selected from the unused segment of the item bank, R_k, to satisfy the function:

$$i_k \equiv \max_j \left\{ I_{u_j}\left(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}\right) : j \in R_k \right\}.$$

Equation 3 chooses the next item with the maximum information at the provisional θ estimate. To accomplish that, we need *an item information function*. The concept of an item information function was introduced by Allan Birnbaum in 1968 and relates to the precision that each item provides with response to a given scoring function. For the 1PL, 2PL, and 3PL IRT models described earlier, the item information function can be written as:

$$I_i(\theta) = \left\{P_i(\theta)\left[1-P_i(\theta)\right]\right\}^{-1}\left[\partial P_i(\theta)/\partial\theta\right]^2,$$

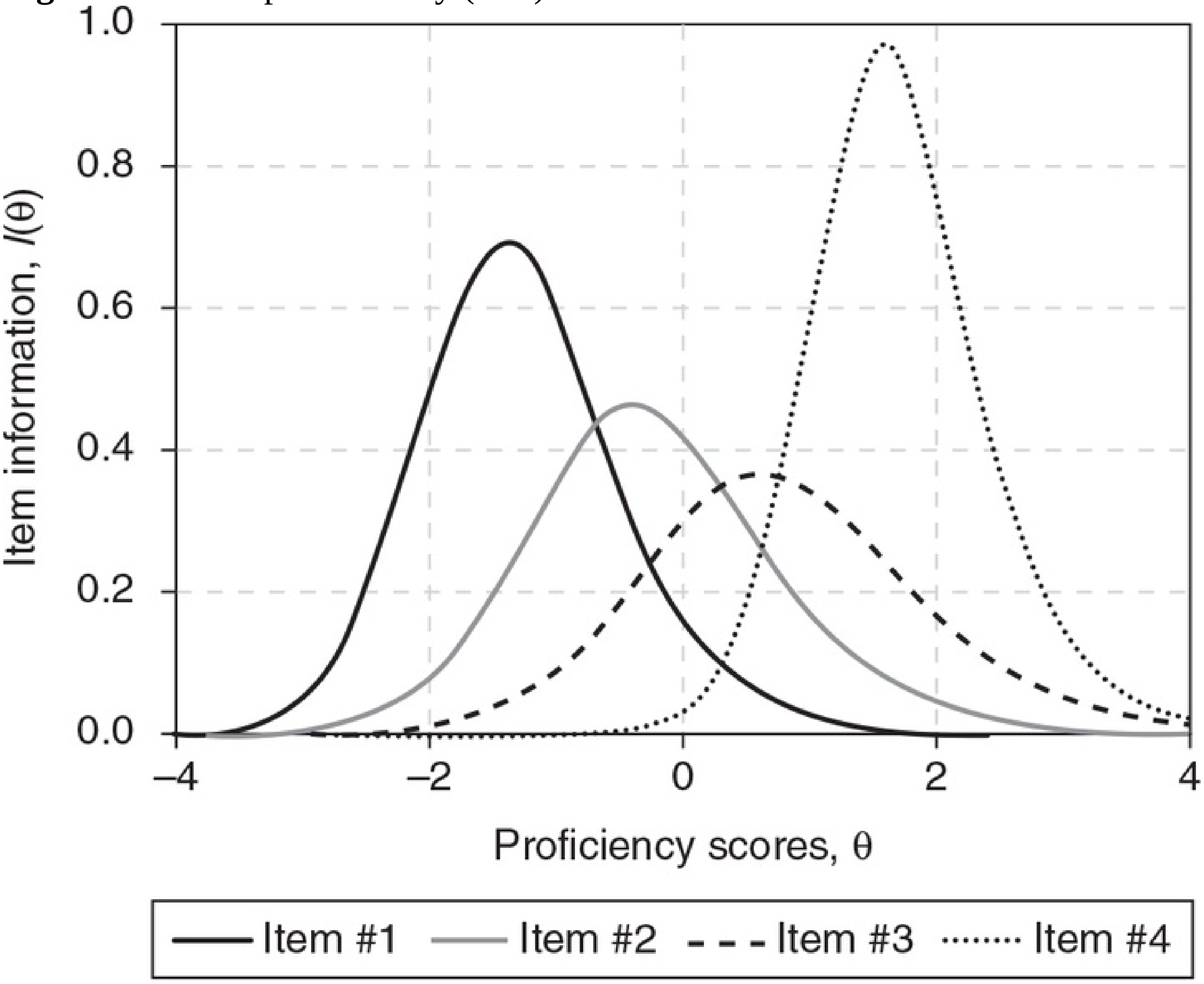
where ∂ denotes the first partial derivative of the IRT probability function, $P_i(\theta)$—see Equation 1. The probability function and associated first derivative terms change for each of the IRT models. For example, the 3PL item information function can be written as:

$$I_i(\theta) = \left\{a_i^2\left[1-P_i(\theta)\right]\left[P_i(\theta)-c\right]^2\right\}\left[P_i(\theta)(1-c)^2\right]^{-1},$$

(also see Equation 1), with obvious simplifications for the 2PL and 1PL models. Equation 3, which chooses the (unselected) item from the item bank with the maximum information at the provisional estimate of θ, can therefore be implemented by inserting into Equation 5 the requisite item parameter estimates and score estimate.

Consider the item information functions for four sample items shown in Figure 2. The associated 3PL item parameters are $a_1 = 1.1$, $b_1 = -1.5$, $c_1 = 0.12$ for Item 1; $a_2 = 0.9$, $b_2 = -0.5$, $c_2 = 0.12$ for Item 2; $a_3 = 0.8$, $b_3 = 0.5$, $c_3 = 0.12$ for Item 3; and $a_4 = 1.3$, $b_4 = 1.5$, $c_4 = 0.12$ for Item 4.

Figure 2 Item response theory (IRT) item information functions for 4 items



Based on the maximum information criterion for adaptive selection, we can visually confirm that Item 1 would most likely be chosen as most informative for examinees with proficiency scores below −0.67. Item 2 would be chosen for examinees with provisional proficiency scores falling within the interval $-0.67 \leq \theta < 0.33$, Item 3 would be selected for examinees with scores in the interval $0.33 \leq \theta < 0.8$, and Item 4 would be selected for examinees demonstrating proficiency at or above 0.8.

Another interesting implication of Figure 2 is the relatively small range of θ where information is maximal for Items 2 and 3. In both cases, the lower *a* parameters (0.9 and 0.8, respectively) reduce the effective range of adaptive utility for those 2 items. Conversely, Items 1 and 4 are more informative over a wide range of the proficiency scale. However, the increased utility of those 2 items also increases their likely exposure (overuse) within the examinee population. This point is briefly addressed in the following.

The test unit size being selected does not alter the basic adaptive algorithm shown earlier in Figure 1 because the item information (Equations 4 and 5) is additive across items or test units. That is, if we create a new test unit as a collection of 5 items, the information for those 5 items can be summed and used for selection.

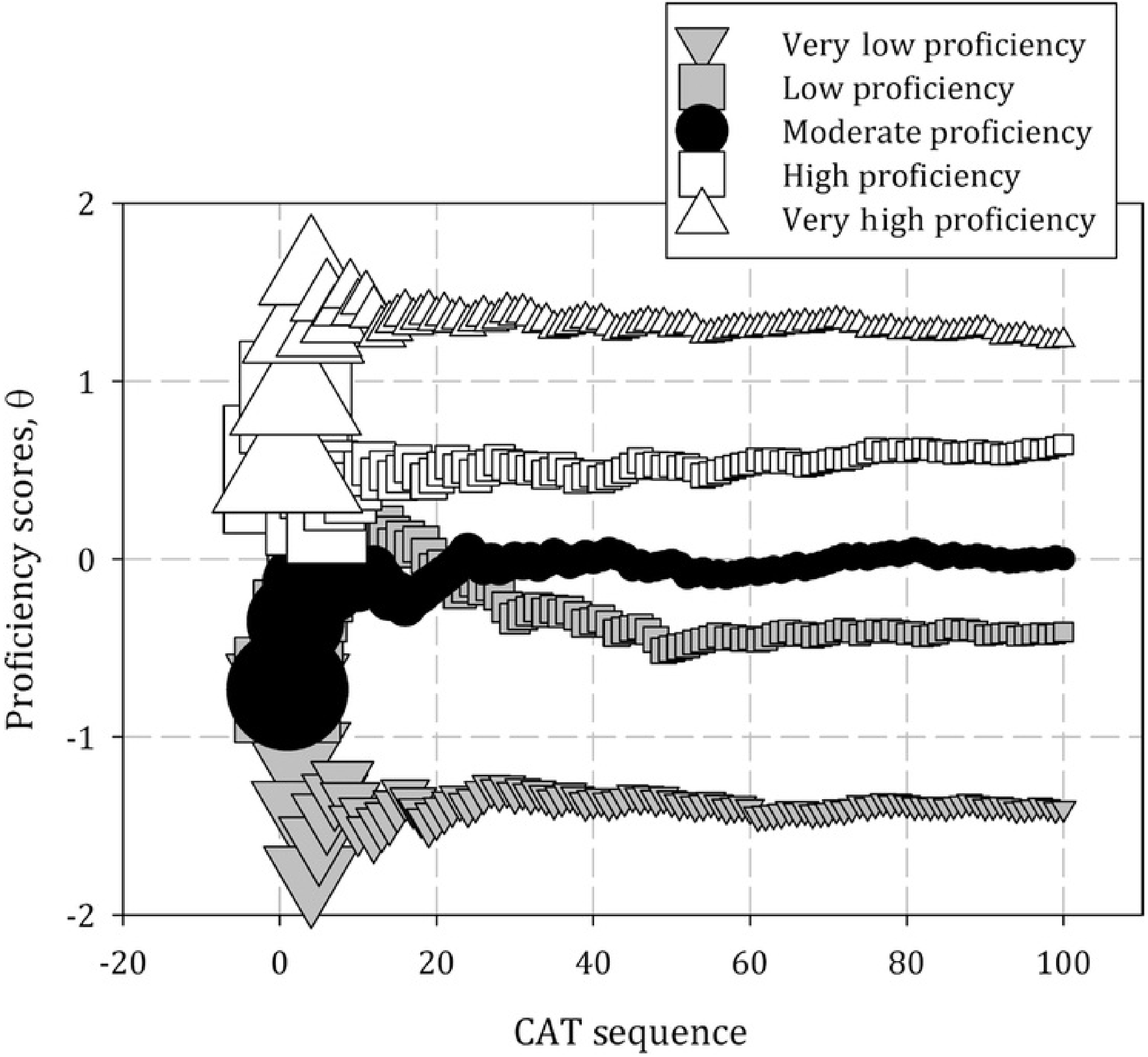
IRT item information functions directly contribute to the overall precision of the θ score estimates. That is, under IRT, the conditional measurement error variance of estimated proficiency scores is inversely proportional to the test information function (TIF) which is the sum of the item information functions. That is, the error variance of estimate of the proficiency scores is inversely proportional to the TIF:

$$\sigma^2(\hat{\theta} \mid \theta) = \left[\sum_{i=1}^{n} I_i(\theta)\right]^{-1},$$

where $I_i(\theta)$ is the item information function at some estimate of the proficiency score of interest. As noted earlier, the exact mathematical form of the information function varies by IRT model (e.g., see Equation 5 for the 3PL item information function). Each item adds some information to the TIF. The more item information we add to each examinee’s adaptive test, the smaller we force the error variance to be. This reduction in the error variance is the ultimate goal of most adaptive test algorithms.

Figure 3 shows what happens to the provisional proficiency scores and associated standard errors (the square root of the error variance from Equation 3) for five hypothetical examinees each taking a sequence of 100 adaptively administered items. These hypothetical examinees are each at a different level of proficiency (*very low* θ, *low* θ, *moderate* θ, *high* θ, and *very high* θ). Because of their proficiency differences, each examinee saw a different sequence of 100 items. For example, the very low proficiency examinee saw an easier set of items than the low proficiency examinees. The very high proficiency examinee saw the most difficult set of items. Note that the plotted estimated θ scores are IRT expected a posteriori scores mentioned earlier.

Figure 3 Expected a posteriori (EAP) θ score estimates for five examinees each taking a 100-item computerized adaptive testing (CAT)



The item pool used for this example comprises 600 items. The proficiency scale is shown as the vertical axis (−2.0 to +2.0). The sequence of 100 adaptively administered items is shown on the horizontal scale. Each plotted symbol is located at the current, provisional estimated θ. The size of each symbols is directly proportional to the standard error of estimate—that is, the square root of the TIF-based error variance from Equation 6. The standard errors are extremely large early in the CAT sequences, but eventually become quite small as more items are administered. All five examinees start with proficiency score estimates near zero, but then, the provisional estimates tend to fluctuate quite a bit. The trajectories of the estimated proficiency scores soon begin to separate for the five examinees after approximately 15 items are administered and tend to fully stabilize at 50–60 items. The standard errors continue to decrease in magnitude for the entire CAT sequence, as evidenced by the decreasing symbol sizes.

In practice, an adaptive test can achieve maximum test information (and minimum standard errors of estimate) in two ways. One way is to choose highly discriminating items that provide maximum item information within particular regions of the proficiency scale or at specific proficiency scores (e.g., see Figure 2). Or, we can merely continue adding items to increment the amount of information until a desired level of precision is achieved. Maximizing the test information at each examinee's score is tantamount to choosing a customized, optimally reliable test for each examinee.

However, maximizing the information may overexpose certain items within the examinee population. This is especially serious for testing programs that use the same item bank over extended periods of time. Overexposure of test items implies that some portion of items in the item bank are administered too often and can be easily be memorized and shared with examines testing at some later date. Almost any type of high-stakes use of test scores (e.g., granting entrance into graduate school, awarding scholarships, providing access to a highly coveted course placement, getting a high-paying or prestigious job, obtaining a professional license or certificate) must consider the possibility that there could be a group of cheaters intent on beating the odds (of random chance or luck) by employing well-thought-out strategies that provide them with any possible advantage of even slightly raising their scores. One of the most common security risks in high-stakes CAT involves groups of examinees collaborating to memorize and share items, especially when the same item database is active over a long period of time, and testing is nearly continuous during that time period.

There are methods of mitigating the overexposure risks. One approach is to increase the size of the active item database to reduce the likelihood of selecting a particular item. A second approach is to build different versions of the item bank that can be rotated in and out of active use over time. The third approach involves a modification to the CAT item selection algorithm. Extensive simulations are used to estimate item control parameters for all items in the bank. Those item control parameters are then used with a relatively simple probabilistic mechanism to buffer the likelihood of always choosing the most informative items. A more extensive discussion of item exposure controls is beyond the scope of this entry.

Concluding Comments

As presented in this entry, CAT is far more than a simple test delivery algorithm—it is a multifaceted collection of algorithms, test designs, and technologies for creating more efficiency tests. There is no single CAT delivery model or framework that universally works “best” for every application. But, CAT is continually evolving to incorporate new assessment applications or purposes and to take advantage of new CBT and psychometric technologies.

Richard M. Luecht

See also Cheating; Computer-Based Testing; Item Banking; Item Information Function; Testlet Response Theory

Further Readings

Birnbaum, A. (1968). Estimation of an ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 423–479). Reading, MA: Addison-Wesley.

Drasgow, F. (Ed.). (2016). Technology and testing: Improving educational and psychological measurement. New York, NY: Routledge.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 471–515). Washington, DC: American Council on Education/Praeger Publishers.

Hambleton, R. K., & Swaminathan, H. R. (1985). Item response theory: Principles and applications. Hingham, MA: Kluwer.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Luecht, R. M. (2014). Computerized adaptive multistage design considerations and operational issues. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 69–83). New York, NY: Taylor … Francis.

Luecht, R. M., & Sireci, S. G. (2011). A review of models for computer-based testing. New York, NY: The College Board (Research Report, 2011–2012).

Mislevy, R. J. (1986). Bayesian modal estimation in item response models. Psychometrika, 86, 177–195.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical considerations in computer-based testing. New York, NY: Springer.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). Computerized adaptive testing: From inquiry to operation. Washington, DC: American Psychological Association.

Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of adaptive testing (pp. 57–76). New York, NY: Springer.

van der Linden, W. J., & Glas, C. E. W. (2010). Elements of adaptive testing. New York, NY: Springer.

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). Computerized multistage testing: Theory and applications. London, UK: CRC Press.

Jeffrey D. Karpicke Jeffrey D. Karpicke Karpicke, Jeffrey D.

Concept Mapping

Concept mapping

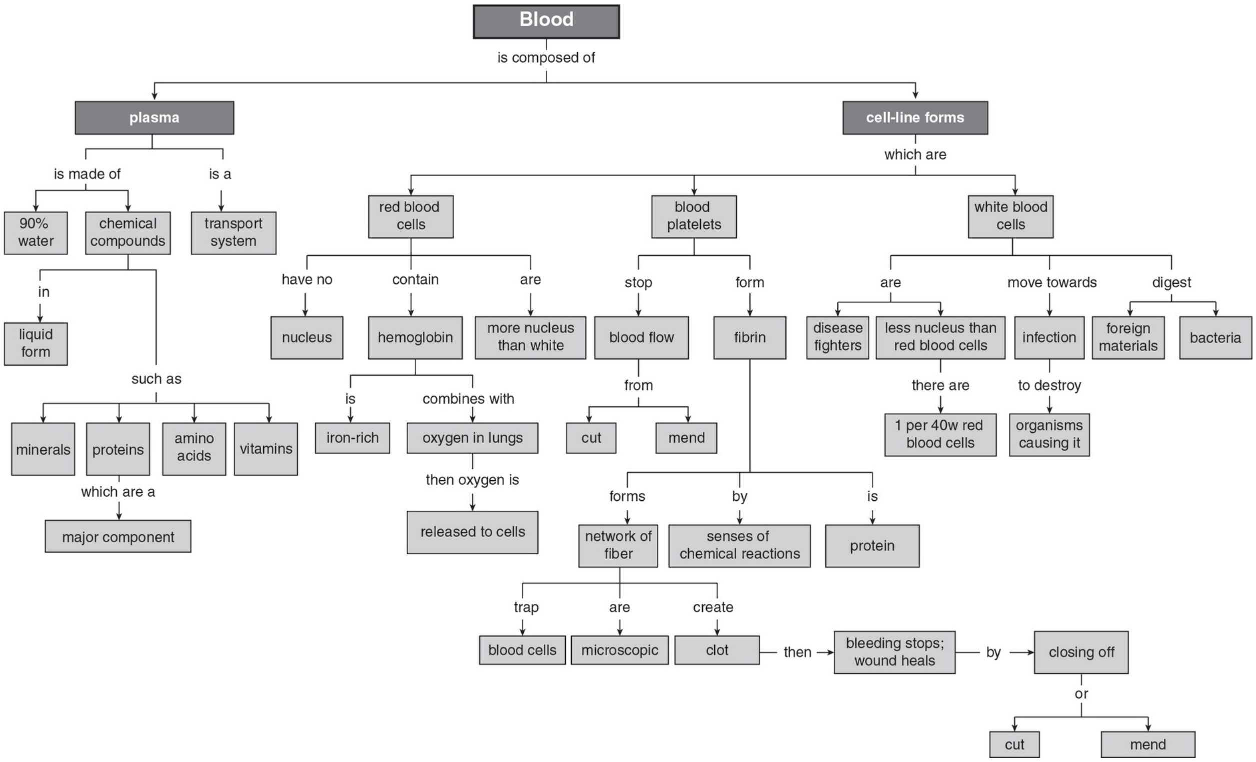
350

354

Concept Mapping

Concept maps are node-and-link diagrams that represent the key terms and relations among terms within a set of materials. *Concept mapping* refers to the activity of creating a concept map. There are a variety of ways to create concept maps, but all share common elements: People construct concept maps by identifying key terms or ideas, placing those key terms in nodes, drawing lines that link related terms, and writing a description of the nature of the relation along the link. Figure 1 shows an example of a concept map created by a college student while they read a text about the composition of blood. No sophisticated tools are needed to create concept maps—pencil and paper will suffice—but several computer programs have been developed to aid in the creation of concept maps. Concept mapping is done in educational settings in a variety of ways, from students creating concept maps as they study on their own (e.g., while they read a textbook) to teachers and students constructing maps as a collaborative classroom activity. Concept mapping may be used for a wide variety of purposes, including creative brainstorming, note-taking, outlining, and—the focus of this entry—as an activity intended to promote learning. Concept mapping enjoys widespread popularity in educational settings and among the general public.

Figure 1 Example of a concept map created by a student in an experiment by Karpicke and Blunt (2011)



Concept Mapping and Related Techniques

Concept maps bear a surface resemblance to semantic networks developed in cognitive psychology in the early 1970s. Such network models depict semantic knowledge as a set of interconnected nodes and assume that when one idea or concept is activated, the activation spreads throughout the network to other related notes. In the late 1970s, Joseph Novak developed concept mapping as a pedagogical tool. The original intent of concept mapping was to track students' conceptual change over time. For example, a student's knowledge about the composition of blood may change over the course of a semester-long anatomy class, and such changes would be reflected in the changing organization of concept maps produced by the student at different points in the semester. An assumption behind concept mapping is that when learners express their knowledge on a concept map, they express more, or express knowledge differently, relative to what they would express on a different assessment.

Concept mapping shares similarities with other mapping techniques, all of which can be considered types of *graphic organizers*. In a technique known as *knowledge mapping*, students create node-and-link diagrams, just as they do in concept mapping, but must use a predefined set of relations to do so (e.g., "part," “type,” “example”). There is no universal agreement about whether concept maps and knowledge maps are functionally similar activities, and no direct comparisons exist in the literature. *Mind mapping* is another technique that also involves representing knowledge in a node-and-link diagram, but mind maps typically center on a single concept (node) with several associated images and ideas radiating from this central node. Likewise, causal maps and flowcharts represent knowledge in node-and-link diagrams. While concept maps may represent cause-and-effect relations, concept mapping is generally considered to be different from mind maps, causal maps, and flowcharts.

Evaluating Concept Maps

A great deal of debate has focused on the most meaningful and informative ways to evaluate students’ concept maps. Perhaps the most straightforward way to assess a concept map is to tally the number of *idea units* represented on the map, whereby an idea unit is a proposition that expresses an idea or concept. For example, in the map in Figure 1, “blood is composed of plasma” was scored as one correct idea unit. Evaluations of concept maps can become considerably more sophisticated than this simple example, when one begins to consider the number of nodes, the number of links, and the overall organizational structure of links on a map. In a map like the one in Figure 1, nodes exist in different levels of a hierarchy, and students may identify *cross-links*, where a node in one section or level of a map is linked to a node in a different section or level. The presence of cross-links on a student’s map is thought to represent relatively deeper knowledge and insight about a domain.

Claims About Concept Mapping

The chief claim about concept mapping is that concept maps improve learning, but many additional claims about concept mapping have appeared in the literature and popular media. Concept mapping has been proposed to stimulate brainstorming and the generation of new ideas, aid in creativity, improve metacognitive monitoring (the self-assessment of one’s own knowledge), enhance critical thinking, and serve as an effective note-taking technique. Many of these claims have not been thoroughly examined in experimental or quasi-experimental research, for instance, by comparing a concept map condition to a plausible control condition and determining whether concept mapping improves the outcome of interest (e.g., idea generation or metacognitive accuracy). All of the claims mentioned here are plausible and perhaps true; but without more thorough research, no firm conclusions can be drawn.

One claim that has been examined in experimental research is that concept mapping improves students' affect, self-efficacy, and motivation. A 2006 meta-analysis of concept mapping research identified six papers that examined these outcomes, all of which reported positive effects of concept mapping. This represents promising support for the effectiveness of concept mapping in promoting students' affect, self-efficacy, and motivation, but given the relatively small number of studies in the literature, further exploration is warranted.

Mechanisms of Concept Mapping: Why Should Concept Mapping Promote Learning?

It is worth considering why concept mapping should be expected to promote learning. Although there is a fairly extensive research base on concept mapping, few studies have targeted the underlying cognitive processes that learners might engage in when they create concept maps. In the basic cognitive science literature, it is well established that a combination of *relational* and *item-specific processing* supports effective and durable encoding. Relational processing refers to tasks in which learners consider how items are similar to one another, whereas item-specific processing refers to tasks that emphasize how items are distinctive, unique, or different from one another. When trying to learn new information, engaging in both relational and item-specific encoding is a recipe for a robust mental model of the material.

Concept mapping would seem to emphasize relational processing by focusing on how terms are similar to one another and how ideas fit together within an organizational structure. The concept map shown in Figure 1 appears to provide a clear depiction of the overall relational structure of the text. It is possible that concept mapping also promotes distinctive or item-specific processing; perhaps this would be especially true when learners create cross-links or links that emphasize the distinctiveness of terms within categories. However, the literature is sparse when it comes to discussion of possible encoding mechanisms that concept mapping might afford.

One recent study, reported in 2015, examined the effects of concept mapping on relational and item-specific knowledge and suggested that some concept mapping activities may be detrimental to item-specific encoding. Standard mapping activities may be detrimental to item-specific encoding. Standard concept mapping instructions emphasize that learners should form many relations among items. As a consequence, learners may create overloaded categories in which too many terms become linked to higher level category nodes. Ultimately, the creation of overloaded categories hurts learning performance relative to other study strategies that also encouraged organizational or distinctive processing.

Does Concept Mapping Promote Learning?

The simple question of whether concept mapping promotes learning is not so simple after all because concept mapping is not a single prescribed activity. Concept mapping can be done in a variety of ways. For example, students might study a concept map as an advance organizer before a lesson, perhaps one created by a teacher or one that accompanies a text. Students might create maps while reading, or they might create them after they have read something (as a retrieval practice activity). Students might create maps on their own or in collaboration with other learners. And students might engage in concept mapping activities that offer varying degrees of support. For example, they might have access to a “node bank” that contains the key terms to be used on a map, they might be given a portion of a map and asked to fill out the remainder, or they may engage with an adaptive computer program that assists learners as they build concept maps (e.g., the Betty’s Brain intelligent tutoring system).

The most extensive analysis of the effectiveness of concept mapping was a 2006 meta-analysis that identified 55 experimental and quasi-experimental studies of concept mapping and knowledge mapping. In general, concept mapping produced positive effects on measures of student learning. The largest effects were observed in studies that compared concept mapping to relatively passive control conditions, like listening to material in lecture format. Studying concept maps produced small but positive effects on learning relative to studying by reading texts or outlines. In studies that compared concept mapping to other active control conditions (e.g., creating an outline rather than simply reading an outline), concept mapping showed even smaller but, nonetheless, positive effects on learning. In short, concept mapping tends to benefit learning, but the size of the effect depends on whether concept mapping is compared against passive or more active control conditions.

Future Directions

As noted earlier, concept mapping remains very popular in a range of educational and applied settings. However, many of the central claims about concept mapping require further research and investigation. Many studies have shown positive effects of concept mapping on learning, but there is a continuing need to identify the most effective ways to structure concept map activities to support effective encoding and promote learning.

Jeffrey D. Karpicke

Further Readings

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2015). From design to implementation to practice a learning by teaching system: Betty’s Brain. International Journal of Artificial Intelligence in Education.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. Journal of Educational Psychology, 106(3), 849–858. doi:10.1037/a0035934

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82(6), 407–428. doi:10.1037/0033-295x.82.6.407

Grimaldi, P. J., Poston, L., & Karpicke, J. D. (2015). How does creating a concept map affect item-specific encoding? Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(4), 1049–1061. doi:10.1037/xlm0000076

Hunt, R. R. (2012). Distinctive processing: The co-action of similarity and difference in memory. In B. H. Ross (Ed.), The psychology of learning and motivation (Vol. 56, pp. 1–46). San Diego, CA: Elsevier Academic Press.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. Science, 331(6018), 772–775. doi:10.1126/science.1199327

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. Review of Educational Research, 76(3), 413–448. doi:10.3102/00346543076003413

Novak, J. D., & Cañas, A. J. (2006). The theory underlying concept maps and how to construct and use them (Technical Report). Pensacola, FL: Institute for Human and Machine Cognition.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. Journal of Research in Science Teaching, 33(6), 569–600. doi:10.1002/(sici)1098-2736

Sharon F. Rallis Sharon F. Rallis Rallis, Sharon F.

Conceptual Framework

Conceptual framework

354

356

Conceptual Framework

A conceptual framework provides a map of the world a researcher intends to study. It captures what researchers see and how they make sense of what they are exploring. *Concept* means ideas, perceived facts, beliefs, mental pictures, perceptions, and theories. *Framework* indicates basic structure or underlying organizational elements. Drawing on these definitions, a conceptual framework is an organizing structure or scaffold that integrates related ideas, mental images, other research, and theories to provide focus and direction to the inquiry. It defines the “what”—the substantive focus—of the study and thus serves to guide and direct the on-going decision making required in any research endeavor. Ultimately, the conceptual framework puts forward an argument and establishes the significance of the study. After reviewing the elements that make up conceptual frameworks, this entry explains how to successfully build and then use conceptual frameworks.

Elements

Conceptual frameworks consist of three elements: the researcher’s personal experience and viewpoints, existing information and knowledge of the phenomenon under study, and relevant theoretical positions regarding the phenomenon. A clear, well-developed conceptual framework functions as an integrated system that discloses these perspectives, illustrates interrelationships, and establishes boundaries. Any argument or thesis that drives a study emerges from the conceptual framework, and tools for analysis and interpretation are embedded within the framework.

Building a Conceptual Framework

The first element in a conceptual framework for any study begins with the researcher—the researcher's knowledge, experiences, and interests related to the phenomenon. As thinking beings, researchers bring to the topic their interests, preferences, and interpretations. Systematic and rigorous inquiry requires that researchers make these central assumptions or claims explicit: Why have they chosen this topic? Why do they consider it important? What do they know about it already? What attitudes and opinions do the researchers hold regarding the topic? The answers help identify *sensitizing concepts* and *orienting perspectives* that suggest areas for focus, aspects and relationships to explore, possible ways to organize including boundaries to set, as well as *currents of thought* that can inform the inquiry.

The second element in the conceptual system is these currents of thought (i.e., relevant bodies of literature and the existing knowledge) about the phenomenon of interest. Recognizing these extant bodies or currents allows researchers to ground their work in scholarly and public discourse concerning what is already known about the phenomenon. Through the literature, researchers connect their particular interest to a larger, more general interest. Whatever the topic, someone has questioned, researched, or written about the general phenomena under consideration, so researchers critically read research studies, policy writings, reports about practice, evaluations, essays and opinion pieces, newspaper editorials and articles, and even popular communication on the topic. They ask: *What have scholars or "experts" said about this topic? What is the discourse in the public domain? What questions have already been raised or explored? What previous research can be built upon?* Relevant information and knowledge is woven into the framework to ground the study in what is already known, provide substantiation for points, clarify logic, define concepts, and suggest relevant theories.

Finally, the conceptual framework has a theoretical base that further connects the researcher's perspectives, the specific study focus, and the larger scholarly or public conversations about aspects of the phenomenon. A researcher asks: *What theoretical positions have informed my perspectives? What theories might usefully provide insights or direction for the inquiry project?* A theory is a set of propositions that describe, explain, and predict phenomena; it models some aspect of reality. Theory with a capital T consists of what Peter Burke calls an accepted and coherent set of statements, assumptions, or axioms that have been tested and accepted as explanations for particular phenomena. These Theories carry labels (e.g., self-efficacy) and are often attributed to an individual or group of individuals (e.g., Thomas Kuhn's *Structure of Scientific Revolutions*). In addition, many references to theory imply hunches or intuitive propositions believed to guide actions. As Carol Weiss has noted, a theory does not have to be universally accepted or correct; theory can be viewed as a set of working understandings or hypotheses that underlie action and guide analysis and interpretation. Researchers bring *theories* to their studies, and they find *Theories* among the scholarly discourse. Both enlighten and broaden researchers' perspectives, offer explanations, suggest patterns, and contribute to a generative foundational (i.e., based in theory) conceptual framework.

Using the Conceptual Framework

As previously stated, the conceptual framework defines "what" is to be studied. Specifically, the researcher uses the conceptual framework to

- describe and explain the phenomenon,
- embed the phenomenon in a context,
- construct an argument that articulates perspective,
- generate questions or hypotheses,
- sharpen the focus,
- propose strategies for action,
- provide categories for analysis, and
- link the questions to larger theoretical constructs and policy discussions.

The process of developing the framework is both inductive and deductive. It forces researchers to be explicit about their thinking and intended actions and to integrate their ideas with other research and theory. The framework becomes a selection tool that facilitates a coherent study, that is, a researcher uses it constantly to decide what is important, to explicate rationale and significance, to frame research questions, to choose the "how" (design and method), to choose which data to gather, to provide direction for analysis, and to interpret findings. The process also serves as a catalyst that raises the researchers' thinking from the particular and descriptive to contribute to some larger body of ideas contained in the research, writings, and experiences of others.

Sharon F. Rallis

See also Concept Mapping; Literature Review; Logic Models; Program Theory of Change

Further Readings

Blumer, H. (1954). What's wrong with social theory? American Sociological Review, 19(1), 3–10.

Bowen, G. A. (2006). Grounded theory and sensitizing concepts. International Journal of Qualitative Methods, 5(3), 2–9.

Burke, P. J. (2009). The elements of inquiry: A guide for consumers and producers of research (p. 62). Glendale, CA: Pyrczak.

Leshem, S., & Trafford, V. (2007). Overlooking the conceptual framework. Innovations in Education and Teaching International, 44(1), 93–105.

Rallis, S. F., & Rossman, G. B. (2012). The research journey: Introduction to inquiry (pp. 85–110). New York, NY: Guilford.

Schram, T. H. (2006). Conceptualizing and proposing qualitative research (2nd ed., p. 58). Upper Saddle River, NJ: Pearson Merrill.

Weiss, C. H. (1998). Evaluation: Methods for studying programs and policies (2nd ed., p. 55). Upper Saddle River, NJ: Prentice Hall.

Williams, B., & Hummelbrunner, R. (2011). Systems concepts in action: A practitioner's toolkit (pp. 18–23). Stanford, CA: Stanford University Press.

Colin P. West Colin P. West West, Colin P.

Thomas J. Beckman Thomas J. Beckman Beckman, Thomas J.

Concurrent Validity Concurrent validity

Concurrent Validity

Concurrent validity refers to the extent to which the results of a measure correlate with the results of an established measure of the same or a related underlying construct assessed within a similar time frame. This entry considers how concurrent validity fits within both the classical framework of validity and Samuel Messick’s unitary view of validity and provides examples of its importance and application within educational research.

Place in Validity Framework

In classical views of validity, concurrent validity is a type of criterion validity, which concerns the correlation between a measure and a standard regarded as a representative of the construct under consideration. If the measure is correlated with a future assessment, this is termed *predictive validity*. If the measure is correlated with an assessment in the same general time frame, this is termed *concurrent validity*. Conversely, poor correlation of the measures where correlation would be expected provides evidence against concurrent validity.

This validity concept aligns well with Messick’s commonly held unitary view of validity, in which concurrent validity is an example of validity evidence provided by relations to other variables. This type of validity is supported when two measures of the same construct correlate well with one another and called into question when such correlation is not seen.

Importance and Examples Within Educational

Research

Understanding correlations among measures of specific constructs is of great importance in educational research. Two examples will illustrate these concepts. First, a criterion standard test of medical knowledge might involve hundreds of examination items administered over many hours. A shorter medical knowledge assessment’s concurrent validity could be assessed by evaluating the correlation of results from the shorter examination with results from the criterion standard administered shortly before or after the abbreviated test. A strong correlation would provide evidence of concurrent validity which could then be supplemented by evaluations of other elements of validity. If little correlation was found, however, concurrent validity of the shorter measure would not be supported.

Second, an established instrument for depression diagnosis among medical students might be compared with results from a concurrent assessment of burnout. Strong observed correlation between these two measures would support concurrent validity of the burnout measure with the established depression measure. On the other hand, lack of correlation between the two measures would represent evidence against concurrent validity of the burnout measure in the evaluation of depression.

Colin P. West and Thomas J. Beckman

See also Correlation; Criterion-Based Validity Evidence; Internal Validity; Predictive Validity; Tests; Unitary View of Validity; Validity

Further Readings

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. American Journal of Medicine, 119, 166e7–166e16. doi:10.1016/j.amjmed.2005.10.036

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York, NY: American Council on Education and Macmillan.

Kun Zhang Kun Zhang Zhang, Kun

Conditional Independence Conditional independence

357

358

Conditional Independence

Statistical independence and conditional independence (CI) are important concepts in statistics, artificial intelligence, and related fields. Let *X*, *Y*, and *Z* denote three sets of random variables, and let *P* denote their probability distribution or density functions. *X* and *Y* are conditionally independent given *Z*, denoted by $X \perp Y \mid Z$, if and only if $P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$. It reflects the fact that given the values of *Z*, further knowing the values of *X* does not provide any additional information about *Y*. Generally speaking, such a CI relationship allows us to drop *X* when constructing a probabilistic model for *Y* with (*X*, Z), resulting in a parsimonious representation. Moreover, independence and CI play a central role in Bayesian network learning and causal discovery, which aims at recovering the underlying causal model from purely observational data.

A direct way to assess if $X \perp Y \mid Z$ is to estimate the involved probability density or distribution functions and then check on whether the definition is satisfied. However, density estimation in high dimensions is known to be difficult: In nonparametric joint or conditional density estimation, due to the curse of dimensionality, to achieve the same accuracy, the number of required data points grows exponentially in the data dimension.

Testing for CI is much more difficult than that for unconditional independence. For CI tests, traditional methods either focus on the discrete case, in which the chi-square test can be used, or impose simplifying assumptions to deal with the continuous case. In particular, the variables are often assumed to have linear relations with additive Gaussian errors. In that case, $X \perp Y \mid Z$ reduces to zero partial correlation or zero conditional correlation between *X* and *Y* given *Z*, which can be easily tested. However, nonlinearity and non-Gaussian noise are frequently encountered in practice and, accordingly, the partial correlation test may lead to incorrect conclusions.

CI is just one particular property associated with the distributions; to test for it, it is possible to avoid explicitly estimating the densities. There exist some ways to characterize the CI relation that do not explicitly involve the densities, and they inspired more efficient methods for CI testing. Note that when (X, Y, Z) is jointly Gaussian, $X \perp Y \mid Z$ is equivalent to the vanishing of the partial correlation coefficient between X and Y given Z. As its generalization, J. J. Daudin showed that in the general case, $X \perp Y \mid Z$ if and only if $f(X,Z)-E[f \mid Z]$ is always uncorrelated with $g(Y)-E[g \mid Z]$ for any square-integrable functions f and g. Here, $E[f \mid Z]$ denotes the conditional mean of $f(X, Z)$ given Z. In this way, CI is characterized by the uncorrelatedness of functions in suitable spaces. Kenji Fukumizu and others showed that one can use the reproducing kernel Hilbert spaces corresponding to the so-called characteristic kernels (e.g., the Gaussian kernel) instead of the square-integrable spaces and proposed a measure of conditional dependence. Kun Zhang and others further developed a kernel-based CI test. Such a nonparametric conditional dependence measure and CI test have received many applications in machine learning, statistics, and artificial intelligence.

Kun Zhang

See also Bayes’s Theorem; Bayesian Statistics; Partial Correlations

Further Readings

Daudin, J. J. (1980). Partial association measures and an application to qualitative regression. Biometrika, 67, 581–590.

Dawid, A. P. (1979). Conditional independence in statistical theory. Journal of the Royal Statistical Society. Series B, 41, 1–31.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), Advances in neural information processing systems 20 (pp. 489–496). Cambridge, MA: MIT Press.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence. In NIPS 20 (pp. 585–592). Cambridge, MA: MIT Press.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), Barcelona, Spain.

Robert L. Brennan Robert L. Brennan Brennan, Robert L.

Conditional Standard Error of Measurement Conditional standard error of measurement

358

Conditional Standard Error of Measurement

It is often assumed that classical test theory requires the standard errors of measurement to be constant for all examinees. This is not true. Rather standard errors of measurement can and do vary for examinees with different true scores. A conditional standard error of measurement (CSEM) is a measure of the variation of observed scores for an individual examinee with a particular true score. Measurement is more precise for examinees with small CSEMs.

In 1955, Frederic Lord developed the best known CSEM for number-correct scores. Its estimator is , where x_p is the number of correct dichotomously scored items for examinee *p,* and *k* is the total number of items in a test. Subsequently, in 1984, Leonard Feldt extended Lord’s method to tests in which items are nested within strata, such as fixed categories in a table of specifications. The Lord and Feldt formulas apply only to relatively simple tests with dichotomously scored items. In 1998, using the principles of generalizability (G) theory, Robert Brennan extended CSEMs to any type of raw scores obtained from many different test designs.

In most testing contexts, the scores reported to examinees are not raw scores; rather, the reported scores are transformed raw scores, called scale scores. For linear transformations, the previously mentioned methods can be used with simple adjustments. Usually scale–score transformations are nonlinear, however. If so, obtaining estimated CSEMs is almost always more complicated. Many methods for nonlinear transformations are developed in the 1990s. Item response theory can also be used to obtain estimated CSEMs for nonlinear transformations, although the theoretical basis for doing so is quite different from other methods.

Differentiating between CSEMs for raw and scale scores can have very important implications. For example, for raw scores, CSEMs are often considerably *larger* in the middle of the score distribution than in the ends. By contrast, for many nonlinear scale–score transformations, CSEMs are considerably *smaller* in the middle of the score distribution than in the ends. This is particularly likely for CSEMs obtained using IRT.

Robert L. Brennan

See also Item Response Theory; Generalizability Theory; True Score; Reliability; Standard Error of Measurement

Further Readings

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. Applied Psychological Measurement, 22, 307–331.

Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. Educational and Psychological Measurement, 59, 5–24.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. Educational and Psychological Measurement, 44, 883–891.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. Journal of Educational Measurement, 29, 285–307.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. Journal of Educational Measurement, 37, 1–20.

Yi-Fang Wu Yi-Fang Wu Wu, Yi-Fang

Confidence Interval

Confidence interval

358

362

Confidence Interval

The term *confidence interval* refers to an interval estimate that provides information about the uncertainty or the precision of estimation for some population parameter of interest. In statistical inference, confidence intervals are one method of interval estimation, and they are widely used in frequentist statistics. There are several ways to calculate confidence intervals. This entry first emphasizes the importance of confidence intervals by distinguishing interval estimation from point estimation. It then introduces a brief history of confidence intervals. The essentials of constructing confidence intervals are discussed, followed by a brief introduction to other types of intervals in the literature. Confidence intervals have been emphasized in the social and behavioral sciences, but they are often misinterpreted in statistical practice. Thus, the entry concludes with a discussion of common misunderstandings and misinterpretations of confidence intervals.

Interval Estimation Versus Point Estimation

The purpose of inferential statistics is to infer properties about an unknown population parameter using data collected from samples. This is usually done by point estimation, one of the most common forms of statistical inference. Using sample data, point estimation involves the calculation of a single value, which serves as a best guess or best estimate of the unknown population parameter that is of interest.

Instead of a single value, an interval estimate specifies a range within which the parameter is likely to lie. It provides a measure of accuracy of that single value.

In frequentist statistics, confidence intervals are the most widely used method for providing information on location and precision of the population parameter, and they can be directly used to infer significance levels. Confidence intervals can have a one-sided or two-sided confidence bound. They are numerical intervals constructed around the estimate of the unknown population parameter. Such an interval does not directly infer a property of the parameter; instead, it indicates a property of the procedure, as is typical for a frequentist statistical procedure.

The American Psychological Association’s *Publication Manual* strongly recommends the use of confidence intervals for reporting statistical analysis results. In fact, in the literature, it has been concluded that confidence intervals and null hypothesis significance testing are two approaches to answer the same research question. They give accessible and comprehensive point and interval information to support substantive understanding and interpretation. As George Casella and Roger L. Berger pointed out, in general, every confidence interval corresponds to a hypothesis testing and vice versa. Whenever possible, researchers should base discussion and interpretation of results on both point and interval estimates whenever possible.

Brief History of Confidence Intervals

In the early 19th century, Pierre-Simon Laplace and Carl Friedrich Gauss had already recognized the need for interval estimation to provide information about measures of accuracy. However, the term *confidence intervals* was not used until Jerzy Neyman’s presentation before the Royal Statistical Society in 1934. In the appendix of this paper entitled “On the Two Different Aspects of the Representative Method,” Neyman proposed a straightforward way to create an interval estimate and to determine how accurate the estimate is based on sample data. He called this new procedure *confidence intervals* and the ends of the confidence intervals *confidence bounds*. Also, the arbitrarily defined values termed *confidence coefficients* indicated how frequently the observed interval obtained from sample data contains the true population parameter if the experiment is repeated. Nowadays, a confidence coefficient is often referred to as a *confidence level* for its relation to null hypothesis significance testing. Neyman finally addressed the theory of confidence intervals extensively in 1937 in “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability.” In this paper, the mathematical assumptions, derivations, and proofs provided the philosophical and statistical foundation for confidence intervals.

Construction of Confidence Intervals

There are several approaches to the construction of confidence intervals. The approaches discussed in this entry are commonly used in psychological and educational testing, and they differ in the way of obtaining the standard error (SE) of the sampling distribution for the parameter of interest. The most common, standard procedure is to invert a test statistic as Casella and Berger demonstrated in their textbook *Statistical Inference*. What follows exemplifies the construction of confidence intervals for a population mean μ.

After data are observed, the sample mean may or may not be close to μ and the distance, , is the estimation error and is different for every sample. The margin of error (MOE) is defined as the largest likely estimation error and a confidence interval has this general form, . For the cases with a “likely” confidence coefficient of .95, we say about 95% of the values from the sampling distribution of the mean falls within MOE, which is 1.96 times the SE of the sampling distribution. By virtue of central limit theorem, the value 1.96, denoted $z_{.95}$, is the critical value of a standard normal distribution. The SE is computed in two ways, depending on whether the population standard deviation σ is known. For a known population standard deviation, the SE is , where *N* is the sample size. The 95% confidence interval of the population mean is . When the sample standard deviation *s* is used as an estimate of σ, the critical value comes from a *t* distribution associated with the proper degree of freedom. That is, the 95% confidence interval for μ is .

Compared to the procedure just shown, a more general approach to constructing a confidence interval is done under the likelihood theory. When estimates are constructed using the maximum likelihood principle, the technique of forming confidence intervals is called the asymptotic normal approximation, which works for a wide variety of statistical models. The SE is computed by , where is the second derivative of the log likelihood function with respect to θ, evaluated at , the maximum likelihood estimate of the parameter of interest.

For resampling methods, Thomas J. DiCiccio and Bradley Efron provided a heuristic overview of various bootstrapping confidence intervals that can be routinely constructed even for parameters of a complicated statistical model. A bootstrapping procedure yields a certain number of samples such that the bootstrapping SE is obtained for computing the MOE. This MOE value, together with the average of all bootstrapping sample estimates and a given confidence coefficient for obtaining a critical value from the *z* table, one can derive the bootstrapping confidence interval to depict the precision or stability of the estimator of interest. As DiCiccio and Efron showed, asymptotically, bootstrapping confidence intervals are not only good approximate confidence intervals, but more accurate than confidence intervals derived from standard procedures using sample variance and normality assumptions.

Confidence Intervals Versus Other Types of Interval Estimation

Several interval estimation approaches exist in addition to confidence intervals. This entry discusses three such intervals: prediction intervals, tolerance intervals, and Bayesian credible intervals.

Prediction Intervals

A prediction interval specifies the boundaries between which future observations fall. Prediction intervals are often used in regression analysis, where the predicted value for the parameter of interest is obtained, given what has already been observed. The interpretation of prediction intervals is similar to that of confidence intervals. Assuming a confidence coefficient of .95, we can say that the probability that a regression method produces an interval that contains the value of a future observation is 95%.

The difference between prediction intervals and confidence intervals is that the SE used in prediction intervals has to take into account the variability from the difference between the least-square solutions and the true regression as well as the variability of the future response variable. Thus, prediction intervals are always wider than confidence intervals.

Tolerance Intervals

A tolerance interval is an interval to cover a specified proportion of a population distribution with a given confidence for the purpose of predicting a range of likely outcomes. This statistical procedure is often used for quality control in manufacturing. To specify a tolerance interval, both the proportion of the population and a specified confidence level are required. This confidence level is the likelihood that the interval covers the specified, desired proportion of the population.

The width of tolerance intervals differs from that of confidence intervals. The width of a confidence interval, which approaches zero when the sample approaches the entire population, is solely due to sampling error. In contrast, the width of a tolerance interval is affected by not only sampling error but the variance in population.

Bayesian Credible Intervals

In Bayesian statistics, population parameters are random variables rather than fixed values as they are in frequentist statistics. The properties of a population parameter are inferred from the posterior distribution. For example, a 95% credible interval is the 2.5th and 97.5th percentiles of a unimodal posterior distribution.

In general, Bayesian credible intervals differ from frequentists' confidence intervals in some aspects. Credible intervals incorporate information from the prior distribution and the observed data, whereas confidence intervals are solely based on the data. Also, credible intervals and confidence intervals treat nuisance parameters in different ways. Simply put, Bayesian credible intervals treat the parameter being estimated as a random variable, and the resulting interval bounds as fixed values once the posterior distribution of the parameter is found. In contrast, in confidence intervals, the parameter is treated as a fixed value and the bounds are viewed as random variables that depend upon the observed data and can take different values. Prediction intervals mentioned earlier are also used in Bayesian statistics. Again, they are for predicting the distribution of individual future points, whereas Bayesian credible intervals are for predicting the distribution of estimates of the true population parameter that cannot be observed.

Common Misunderstandings and Misinterpretations of Confidence Intervals

There is much confusion about how to interpret a confidence interval. It is important to keep in mind that the interval, not the parameter, is the random quantity. Confidence intervals are probability statements of the procedure but not the probability of the parameter itself. For decades, people argued that confidence intervals were misleading; some even accused Neyman of being unclear about what that probability referred to. In fact, in his paper in 1935, Neyman explicitly displayed the whole concept of confidence intervals. In particular, in one of his formulas, a conclusive probability statement has conveyed all important concepts regarding confidence intervals; that is, . Let us assume α is 95%. The probability statement says if we were able to take repeated samples, based on the sample data observed, 95% of our intervals would contain the population parameter (i.e., $\theta_1 (n) \leq \theta \leq \theta_2 (n)$). It is hoped that for researchers this clarification of the definition is helpful to prevent from misunderstandings and misinterpretations of confidence intervals.

Conclusion

Confidence intervals are probability statements that combine a point estimate with the precision of that estimate and are commonly reported in tables or graphs in scientific writing. Confidence intervals do not allow for probability statements about the true population parameter as the parameters are fixed values in frequentist statistics. Instead, confidence intervals provide for probability statements about the performance of the procedure of constructing such intervals assuming we were able to do so repeatedly.

Researchers sometimes interpret confidence intervals as if they were Bayesian credible intervals, in which the probability statement is about the true parameter itself. The true parameter is unknown. Once data are observed, a confidence interval either contains the true parameter or not. Thus, for confidence intervals, it is false to say that there is a 95% probability that the true parameter lies in the calculated confidence bounds. Such statements and similar arguments should always be avoided.

Yi-Fang Wu

See also Bayesian Statistics; Central Limit Theorem; Inferential Statistics; Interval-Level Measurement; Standard Error of Measurement

Further Readings

Further Readings

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. Psychological Methods, 10, 389–396.

Casella, G., & Berger, R. L. (2002). Statistical inference (2nd ed.). Australia: Thomson Learning.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals (with discussion). Statistical Science, 11, 189–228.

Howe, W. G. (1969). Two-sided tolerance limits for normal populations—some improvements. Journal of the American Statistical Association, 64, 610–620.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97(4), 558–625.

Neyman, J. (1935). On the problem of confidence intervals. The Annals of Mathematical Statistics, 6, 111–116.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences, 236(767), 333–380.

Robert Boruch Robert Boruch Boruch, Robert

Confidentiality

Confidentiality

362

364

Confidentiality

It is customary to identify *confidentiality* as a property of information that is obtained on an identifiable individual or an entity. In education research, the information may take the form of a written hard document or digitized record, a video or audio recording, or an oral report on an observation from the researcher's memory. A *confidentiality assurance* to an individual or an entity entails a promise that, to the extent possible, the information on identifiable individuals or entities will not be disclosed outside the research context in which it is generated and for which it is used. Such an assurance is common in many social research sectors, including education. This entry further defines confidentiality and discusses how it relates to public information, private information, and professional codes of ethics. It then looks at approaches that reduce or obviate the need for confidentiality and the relationship of confidentiality to research that involves linking records over time or linking records from multiple sources.

Confidentiality is distinguished from and related to *privacy*, which is a property of the individual on whom information is obtained. One may assure an individual or a group that whatever information they provide will remain confidential, though the information's provision can be construed in a narrow sense as a reduction in privacy. Assurance then constitutes a limit on any further reduction of privacy.

Security is distinguishable from both confidentiality and privacy. It usually refers to the conditions under which information is maintained and used. Physical locks and electronic encryption schemes are examples of security measures.

Public Versus Private Information

At one extreme, much information on identifiable individuals is public. In the United States, for instance, administrative records on public school teachers are collected and maintained by state education agencies. That information—on teacher identity, teaching or administrative assignments, and other data—is, at times, publicly accessible on state data systems websites and is used in education research. In such records, and in other public systems on parole officers or nurses or others employed by government, salaries and other information on individuals are disclosed.

At the other extreme, law and regulation can restrict access to information and, in doing so, actualize an assurance of confidentiality. For example, records on students in education institutions, whether the institutions are public or private, are covered in the United States by the Family Educational Rights and Privacy Act. Such records cannot be disclosed to anyone who is not associated with the school unless certain conditions are met. One such condition involves asking the students to agree in advance to the disclosure of their records for research purposes, as is the case for some massive online open-access course data systems.

Between these two extremes, there is considerable variation in the factors that influence, or are influenced by, assurances of confidentiality. The World Wide Web, generally, and social media initiatives, in particular, engender far more complicated issues.

Professional Codes of Ethics and Confidentiality

Many professional associations have developed codes of ethics that attend to confidentiality in the context of educational and social research. Section 12 of the American Educational Research Association Code of Ethics published in 2011, for instance, considers the topic in detail. Among other things, it tells its members that "confidential information is protected," that "educational researchers take reasonable precautions to protect the confidentiality of information related to research participants… . (and) do not allow information gained in confidence to be used in ways would unfairly compromise research participants, students, employees, clients, or others" (p. 149). Other organizations that have addressed confidentiality issues, and to which education research is at times relevant, include the American Psychological Association, American Criminological Society, and American Sociological Association.

Obviating or Reducing the Need for Confidentiality Assurance

When education research relies solely on public records, as suggested earlier, it obviates the need to assure any person or entity of confidentiality. Although useful for many research purposes, however, public records are of limited use for many others. A controlled trial on approaches to reducing teacher turnover, for instance, might reasonably rely on public records on public school teachers to determine their positions over time. A similar study on teachers in the context of private schools would have to rely on surveys or other special arrangements to access institutional records because the records are not public.

Different approaches to assuring confidentiality, when confidentiality must be assured, can be classified as procedural, statistical, and legal. For instance, eliciting anonymous responses is a useful and common procedural approach in cross-sectional surveys. Anonymous responses are useless in longitudinal research unless one can rely on respondent-created aliases and consistent use of the aliases over time or on probabilistic matching algorithms. The anonymity is limited in that deductive disclosure of the identities of supposedly anonymous respondents can be accomplished at times though this might take considerable effort and skill. A survey of graduates of a university school of nursing in the United States at a point in time, for instance, may involve one male graduate whose responses to the survey can easily be coupled to auxiliary information on graduates to learn who the anonymous respondent is and to learn more about him.

Statistical approaches to assuring confidentiality in personal interviews and surveys involving identifiable people that deal with sensitive topics are underused in education and related research. Using these methods, the researcher can elicit sensitive information from identifiable individuals in a way that assures that the response cannot be tied directly to the status of the respondent. Some of the methods fall under the rubric of "randomized response."

Statutory approaches to assuring confidentiality of information on an identifiable

Statutory approaches to assuring confidentiality of information on an identifiable participant in research are at hand. Census bureaus in developed countries for instance, including the United States, Canada, United Kingdom, Sweden, Germany, and others, are governed by laws that prevent redisclosure of the information obtained by the census worker or the census agency. More to the point of education related research, several U.S. laws provide the researcher with protection from being compelled to disclose involuntarily the respondent’s confidential information to nonresearch entities such as a court or prosecuting office. These education-related sectors deal with adolescent use of controlled substances, criminal disorder, and mental and physical health, among others. The relevant legal protection, Certificates of Confidentiality, and conditions for them and limits on them can be found at the National Institutes of Health website.

Record Linkage and Integrated Data Systems in Education Research

Longitudinal research on children requires unique identifiers so as to link records over time. More generally, integrating or linking records from different sources has become important in the social and education sciences. One learns about the correlation between children’s early exposure to lead and their subsequent academic achievement only by linking their education records with, for instance, other records on their housing or health. The different record systems are in most countries governed by different government agencies. Each agency may have its own rules on the confidentiality of their records on identifiable individuals. Consequently, linkage agreements and agreement on principles of linkage can be complex.

The Organisation for Economic Co-operation and Development and the American Educational Research Association have initiated efforts to understand how productive record linkages across agencies within countries can be accomplished without compromising confidentiality assurances made to the people on whom records are kept. The specific context has been longitudinal information systems, but this work has larger implications. The Organisation for Economic Co-operation and Development–American Educational Research Association meetings in 2015 aimed to develop basic principles that would be acceptable to people in the countries involved (Ireland, the United Kingdom, the United States, Russia, the Slovak Republic, Norway, and others). One of the lessons of this and related meetings is that the benefits of linking records have to be documented well in any effort to balance the privacy values of the individuals on whom records are kept against the societal and scientific value of the research. The confidentiality of the records is a critical ingredient in this balance.

Robert Boruch

See also Family Educational Rights and Privacy Act; Health Insurance Portability and Accountability Act; Institutional Review Boards; Qualitative Research Methods

Further Readings

American Educational Research Association. (2011). Code of ethics. Educational Researcher, 40(3), 145–156. doi:10.3102/0013189X11410403

Boruch, R. F., & Cecil, J. S. (1979). Assuring the confidentiality of social research data. Philadelphia: University of Pennsylvania Press.

Bouza, C. N., Herra, C., & Mitra, P. (2010). A review of randomized response procedures: The qualitative variable case. Revista Investigacion Operacional, 31(3), 240–247.

Lensvelt-Mulders, G., Hox, J., van der Heijden, P., & Maas, C. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. Sociological Methods and Research, 33(3), 319–348.

National Institutes of Health. (2016, May 18). Certificates of Confidentiality (CoC). Retrieved from https://humansubjects.nih.gov/coc/index

Palys, T., & Lowman, J. (2014). Protecting research confidentiality. Toronto, Canada: James Larimore and Company.

Jennifer Randall Jennifer Randall Randall, Jennifer

Hyun Joo Jung Hyun Joo Jung Jung, Hyun Joo

Confirmatory Factor Analysis Confirmatory factor analysis

364

370

Confirmatory Factor Analysis

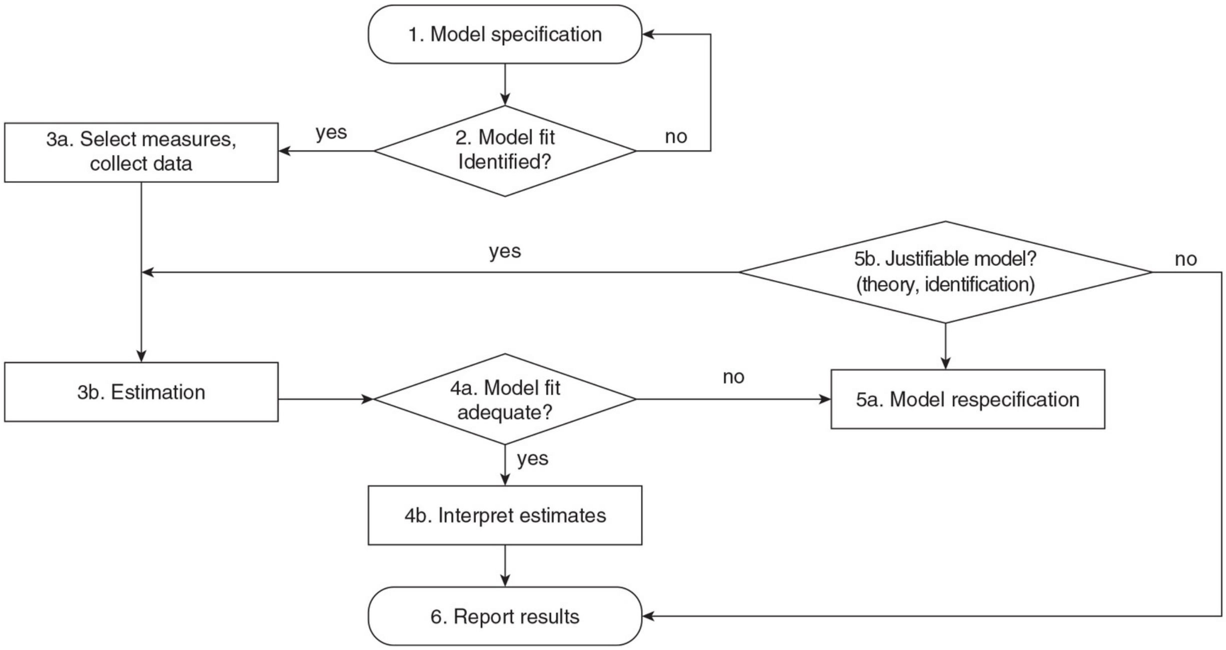
Confirmatory factor analysis (CFA) is a specific type of factor analysis that allows one to determine the extent of the hypothesized relationship between observed indicators and factors (underlying latent variables). CFA, unlike path analysis, allows the distinction between latent variables (referred to as factors) and the indicators (variables) used to measure these latent variables. With CFA models, the factors are assumed to cause the variation and covariation between the observed indicators which are fit to a correlation matrix. This assumption is the primary distinction between CFA and exploratory factor analysis models in which no hypothesis about the number of factors and the relationship between those factors and the indicators is proposed. Thus, CFA can be used for psychometric evaluation, construct validation, and testing measurement invariance. This entry begins with a discussion of the model specification followed by model identification, estimation, evaluation of model fit, and advanced applications of CFA.

CFA models have three primary characteristics:

1. Indicators are continuous variables with two components: (1) one underlying factor that is measured by the indicator and (2) and everything else which is referred to as error.
2. Measurement errors must be independent of each other and the factors.
3. The associations between the factors are not analyzed.

The basic steps in SEM are represented in Figure 1.

Figure 1 Steps in SEM implication framework



Within a structural equation modeling framework, CFA models serve two purposes: (1) to obtain parameter estimates for both the factor (i.e. the factor loadings, the variances, and covariances) and the indicator variables (i.e. residual error variances) and (2) to evaluate the extent to which the model fits the data.

Model Specification

Specification is defined as the representation of a hypothesis in the form of a structural equation model. Specification can take place either before or after data are collected. A CFA model can be defined using various models such as the linear structural relations model (LISREL), the covariance structure analysis model, Bentler-Weeks model, and/or reticular action model (RAM). Provided here are the best known models for continuous observable variables, LISREL and RAM. Figure 2 provides a visual representation of examples of a CFA model with two factors and six indicators.

In Figure 2, squares (or rectangles) and circles (or ellipses) represent observed variables and latent variables, respectively. Also, lines with a single arrowhead and a curved line with two arrowheads reflect hypothesized causal directions and covariances, respectively. This model has seven linear regression equations underlying it, a single structural equation, and six measurement equations. We write the LISREL with matrix notation as:

$$y = \Lambda_y \eta + \varepsilon$$

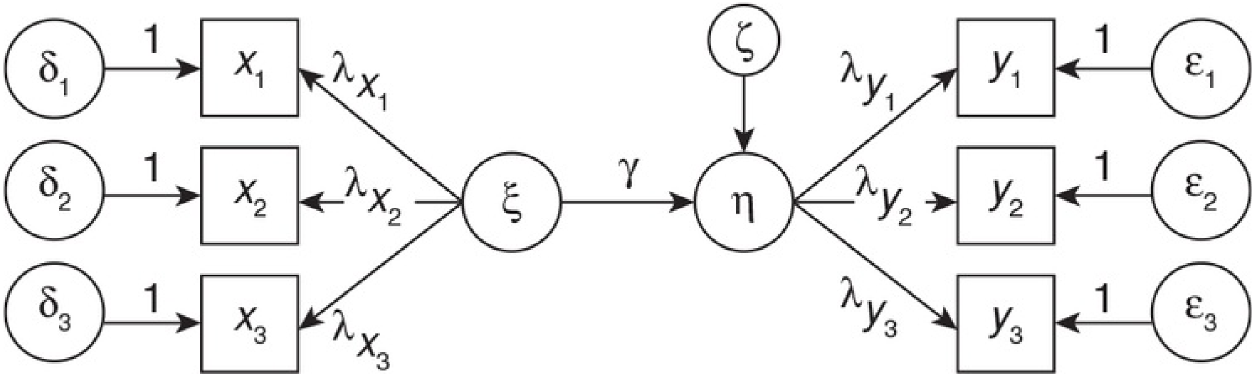
$$x = \Lambda_x \xi + \delta$$

$$\eta = B\eta + \Gamma\xi + \zeta,$$

where *x, y*: exogenous and endogenous variable vectors; Λ_x Λ_y: factor loading matrices; ε, δ: uniqueness vectors;

η, ξ: endogenous and exogenous latent variable vectors; *B*: regression coefficient matrix relating the latent endogenous variables to each other; Γ: regression coefficients matrix relating endogenous variables to exogenous variables; and ζ: structural disturbance vector.

Figure 2 Graphical presentation of an example of a CFA model using LISREL symbols. CFA = confirmatory factor analysis; LISREL = linear structural relations model.



A general covariance matrix for *y* and *x* can be written as:

$$\sum \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

$$= \begin{bmatrix} \Lambda_y (I - B)^{-1} (\Gamma \Phi \Gamma' + \Psi)(I - B)^{-1} \Lambda_y' + \Theta_s & \Lambda_y (I - B)^{-1} \Gamma \Phi \Lambda_x' \\ \Lambda_x \Phi \Gamma' (I - B)^{-1'} & \Lambda_x \Phi \Lambda_x' + \Theta_\varsigma \end{bmatrix},$$

where Φ: the variance–covariance matrix of the exogenous latent variables; Ψ: the variance–covariance matrix of the disturbance terms; and Θ_s, Θ_δ: the variance–covariance matrices of the measurement errors ε and ς.

In terms of the parameter vector Ω, we have $\Omega = (\Lambda_x, \Lambda_y, \Theta_s, \Theta_\delta, \Phi, B, \Gamma, \Psi)$.

Unlike with LISREL, the RAM, observable or unobservable, endogenous, exogenous variables are generally labelled as $v_1, v_2, \ldots, v_\mathrm{n}$. Thus, only one residual variance–covariance matrix of all variables is specified. We write the RAM with matrix notation as:

$$v = Av + u,$$

where *A*: coefficient matrix; *v*: latent and observed variable vector; and *u*: residual vector.

To express the covariance matrix of LISREL as a RAM, we write:

$$\Sigma\begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 & 0 & 0 \\ (I-B)^{-1}\Gamma & (1-B)^{-1} & 0 & 0 \\ \Lambda_x & 0 & \mathrm{I} & 0 \\ \Lambda_y(I-B)^{-1}\Gamma & \Lambda_y(I-B)^{-1} & 0 & \mathrm{I} \end{bmatrix}.$$

If response variables are ordered categorical or dichotomous data, latent continuous response variables, z^*, can be used for the CFA model. For the ordered categorical case, observed variables can be defined as:

$$z = c\,,\ \text{if}\ \ \tau_c < z^* < \tau_{c+1}\,,$$

where z^*: underlying continuous variable; τ_c: thresholds as parameters for the categories $c = 0, 1, 2, \ldots, C-1$, for a variable with C categories, and where $\tau_0 = -\infty$, $\tau_c = -\infty$.

Also, the linear “inner” model describes the relationship for a set of p latent response variables y^* and a set of q latent response variables x^*,

$$y^* = \upsilon_y + \Lambda_y\eta + \varepsilon$$

$$x^* = \upsilon_x + \Lambda_x\xi + \delta,$$

where v_x, v_y: parameter vector of intercepts. Modeling follows the same general structure as when testing a CFA with continuous indicators:

$$\Sigma^{*} = \Lambda^{*}\Phi^{*}\Lambda^{*'} + \Theta^{*},$$

where Σ*: the variance–covariance matrix for modeling relationship between latent response variables; Λ*: the factor loading matrix;

Φ*: the latent variance–covariance matrix among one or more latent variables; and Θ_s: the variance–covariance matrices of the measurement errors.

Identification

Identification is a key concern in model specification. Parameters to be estimated in a model are identified if a unique set for them can be obtained given the model. When every parameter in a model is identified, the model as a whole is identified. The minimum condition of identifiability is that there should be at least as much known information as unknown information (i.e., number of free parameters). The degrees of freedom for a model typically equal the difference between the known and unknown information. Ideally, prior to estimation, the identification of all parameters would be verified. However, in practice, identification is linked to (re)specification.

Estimation

Estimation is the process of finding values for the unknown parameters that minimize the discrepancy between the observed covariance matrix and the estimated (or implied) covariance matrix given the model and the data. For example, in Figure 2, CFA obtains simultaneous estimates of the population coefficients for the seven equations as well as parameters including factor variances, covariances, and error variances using estimation methods. The most commonly used method of estimation is maximum likelihood (ML), which analyzes the covariance matrix of data. The estimates are the ones that maximize the likelihood that data were drawn from the population of interest.

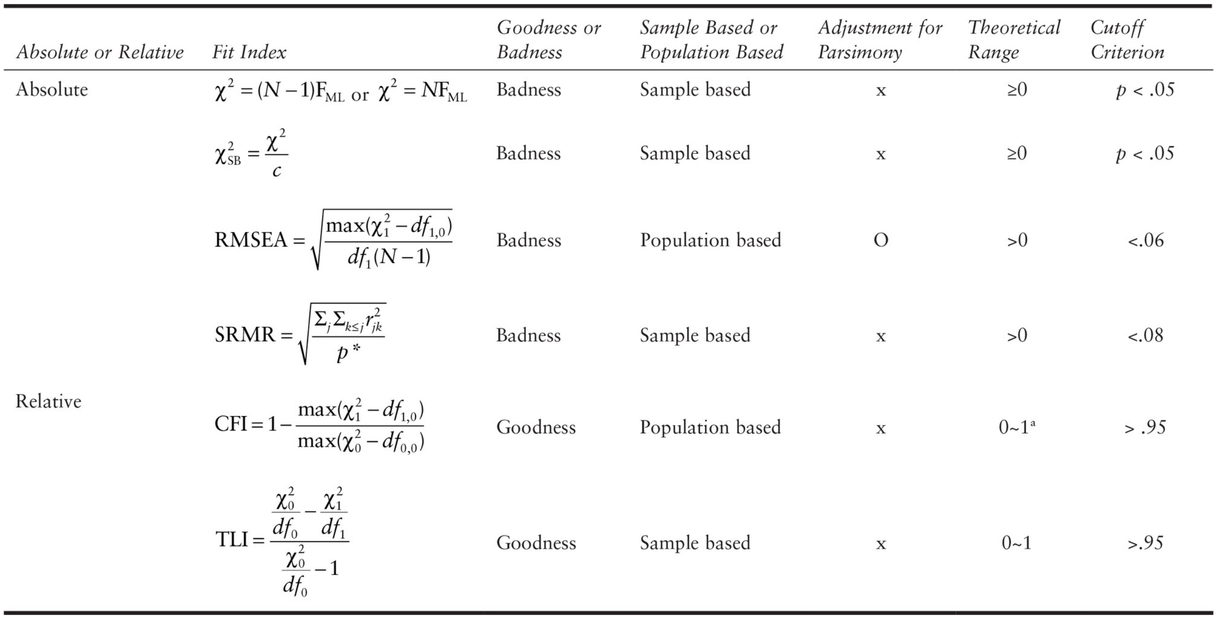
The statistical assumptions of ML estimation include multivariate normality for the joint population distribution of the endogenous variables (which implies that they are continuous), independence of observations, exogenous variables and disturbances, unstandardized observations, and large samples (e.g. N = 200, but $N \geq 200$ needed when analyzing a complex model or outcomes with nonnormal distributions and using an estimation method other than ML). When analyzing categorical variables with asymmetrical distributions, the ML method is probably not appropriate. For categorical indicators, a full information version of ML method can be used. Full information version of ML also directly analyzes the raw data using methods for numerical integration to estimate response probabilities in joint multivariate distributions of the latent response variables, which is the underlying continuous and normally distributed continuum. One alternative method is the fully weighted least square (WLS) method, which does not assume particular distributional form and thus can be applied to continuous or categorical variables.

To reduce the computation complexity of the fully WLS, robust WLS estimation can be used. Robust WLS methods use only the diagonal elements in the weight matrix from full WLS estimation. In the Mplus program, mean-adjusted least squares and mean-and variance-adjusted weighted least squares options are available. Values of the same fit indices can differ over these two methods due to different chi-square and degrees of freedom values. Simulation studies generally favor the mean-and variance-adjusted WLS method over the mean-adjusted WLS. If these estimation methods are applied to data that violate the underlying statistical assumptions of each method, misleading results may be obtained.

Model Fit

Evaluation of model fit concerns whether the specified model explains the data or should be rejected or respecified. Also, if we wish to choose among multiple competing models, fit indices can be used to compare these models. To compare nested models, most of the practical fit indices include the chi-square statistic, which tests the exact-fit hypothesis that there is no difference between the predicted covariance matrix and the population covariance matrix. The chi-square statistic is regarded as undesirable because the chi-square test is quite sensitive to sample size. To overcome this problem, other approximate fit indices have been developed. Indeed, there are more than 50 fit indices introduced in published entries. Here, we summarize the most popular indices.

Alternative fit indices can be classified under absolute and relative (comparative, incremental) fit indices. Absolute fit indices are functions of the discrepancies and include “goodness-of-fit index” (GFI), adjusted GFI, root mean square error approximation, and standardized root mean square residual. Relative indices reflect increments in fit of the researcher’s model over a null model and include incremental fit index, normed fit index, comparative fit index, relative noncentrality index, and Tucker–Lewis index (also called nonnormed fit index). It is customary to report the model chi-square with its degrees of freedom (*df*) and *p* value, root mean square error approximation with its 90% confidence interval, comparative fit index, and standardized root mean square residual as a minimum set of fit indices. Commonly reported alternative fit indices and related information are represented in Table 1.



Absolute or Relative	*Fit Index*	*Goodness or Badness*	*Sample Based or Population Based*	*Adjustment for Parsimony*	*Theoretical Range*	*Cutoff Criterion*
Absolute	$\chi^2 = (N-1)\mathrm{F}_{\mathrm{ML}}$ or $\chi^2 = N\mathrm{F}_{\mathrm{ML}}$	Badness	Sample based	x	≥0	$p < .05$
	$\chi^2_{\mathrm{SB}} = \frac{\chi^2}{c}$	Badness	Sample based	x	≥0	$p < .05$
	$\mathrm{RMSEA} = \sqrt{\frac{\max(\chi^2_1 - df_{1,0})}{df_1(N-1)}}$	Badness	Population based	O	>0	<.06
	$\mathrm{SRMR} = \sqrt{\frac{\Sigma_j \Sigma_{k \le j} r^2_{jk}}{p\,*}}$	Badness	Sample based	x	>0	<.08
Relative	$\mathrm{CFI} = 1 - \frac{\max(\chi^2_1 - df_{1,0})}{\max(\chi^2_0 - df_{0,0})}$	Goodness	Population based	x	0~1[a]	> .95
	$\mathrm{TLI} = \frac{\frac{\chi^2_0}{df_0} - \frac{\chi^2_1}{df_1}}{\frac{\chi^2_0}{df_0} - 1}$	Goodness	Sample based	x	0~1	>.95

a Tucker–Lewis index can have a negative value, which indicates an extremely misspecified model.

F_{ML} = discrepancy function for the ML estimation procedure; = Satorra-Bentler scaled chi-square; *c* = scaling correction factor; 0 = baseline model; 1 = hypothesized model; = standardized residual from a covariance matrix with *j* rows and *k* columns; *p** = the number of nonduplicated elements in the covariance matrix.

Also it can exceed 1, which indicates an extremely well-fitting model.

In Table 1, indices are classified into GFIs where larger values indicate better fit; or conversely, “badness-of-fit” indices in which smaller values indicate improving fit. All relative fit indices are GFIs; absolute fit indices can be either of two types of indices. Also, while sample-bases indices reflect the discrepancy between the model-implied covariance matrix and the sample covariance matrix, population-based fit indices estimate the discrepancy between the reproduced covariance matrix by the model and the population covariance matrix. For some fit indices, they are adjusted depending on their model complexity. Suggested cutoff criteria are not universal for every situation, and so residuals and modification indices should also be investigated for model evaluation. For CFA with categorical data, can be used instead of chi-square. For evaluating nonnested models, Akaike information criterion, consistent information criterion, Bayesian information criterion, cross-validation index, and expected cross-validation index have been proposed.

Although these fit indices provide helpful information for evaluating models, it is most important that the choice between alternative models should be decided based on theoretical rather than statistical considerations as fit measures are not based on meaningful theories but only average or overall model–data correspondence.

Cautions

Researchers and/or practitioners interested in applying factor analytic models with their own data should be mindful of three cautions. First, incorrect model specifications (based on either bad theory or hypotheses) may lead to false conclusions. Moreover, as the CFA model is meant as a test of the measuring instrument as a whole, limited feedback can be provided to the researcher with respect to individual items. Readers should refer to models based in item response theory for drawing such conclusions. Finally, the impact/role of sample size in factor analytic models is considerable. For example, relatively large sample sizes are required to obtain reliable estimates particularly when there are a large number of variables.

Jennifer Randall and Hyun Joo Jung

See also Exploratory Factor Analysis; Path Analysis; Structural Equation Modeling

Further Readings

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), Second International Symposium on Information Theory (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238–246.

Bentler, P. M. (1995). EQS: Structural equations program manual. Encino, CA:

Multivariate Software.

Bollen, K. A. (1989). Structural equations with latent variables. New York, NY: Wiley.

Hoyle, R. H. (2012). Introduction and overview. In R. H. Hoyle (Ed.), Handbook of structural equation modeling (pp. 3–16). New York, NY: Guilford Press.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

Jöreskog. K. G. (1970). A general method for the analysis of covariance structures. Biometrika, 57, 239–251.

Kline, R. (2016). Principles and practice of structural equation modeling (4th ed.). New York, NY: The Guilford Press.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 43–65.

Robert Boruch Robert Boruch Boruch, Robert

Conflict of Interest Conflict of interest

Conflict of Interest

In the area of research, conflict of interest usually involves researchers being in a position to personally benefit from particular information generated in the research or personally benefiting directly or indirectly as a consequence of engaging in the research. Typically, though not exclusively, conflict of interest involves financial gain for oneself, one’s family or friends, or one’s business associates. The information of typical concern is what one learns about the opportunity for personal gain that is incidental to the declared purpose of the research and that may undermine the confidentiality assurances or understandings that the research engenders.

The Singapore Statement on Research Integrity, developed in 2002 at the Second World Congress on Research Integrity, gives a succinct definition and directive on conflict of interest as part of its list of 14 responsibilities for researchers: “9. Conflict of Interest: Researchers should disclose financial and other conflicts that could compromise the trustworthiness of their work in research proposals, publications, and public communications as well as in all review activities.” This entry discusses how the topic of conflict of interest is handled in professional codes of ethics, how it comes into play in research and advisory work, and the consequences of failure to disclose actual or potential conflicts of interest.

In developed countries, the codes of ethics of professional organizations attend in detail to the topic of conflict of interest in research. The codes issued and revised periodically by the American Educational Research Association (AERA), American Sociological Association, and American Psychological Association are among the most explicit in dealing with the subject. The AERA Code, for example, provides guidance on potential conflicts in direct research, in the workplace, and outside the workplace.

Section 10 of the AERA’s Code of Ethics, adopted in 2011, declares that researchers must disclose:

> Relevant sources of financial support and relevant personal or professional relationships that may have the appearance of or potential for a conflict of interest to an employer or client, to the sponsors of their professional work, and to the public in written and verbal reports. (p. 148)

In research on the use of beer or wine by adolescents, for example, the receipt of a research from a brewery must be disclosed partly because the research results (finding that wine drinkers are more polite and achieve higher test scores than beer drinkers) can be influenced by the sponsorship.

Further, the AERA code declares that the researcher must not seek to gain from confidential or proprietary information that they obtain as part of their employment or relationship with a client unless they have permission from the employer or client. Research on students who invent robotic devices, for example, can generate information on the device or on the student’s need for capital investment in production, information that could be exploited unfairly by the researcher.

Some publishers of education research require that authors disclose potential conflicts when they have a journal article accepted for publication. These conflicts can include forms of financial support or financial interests related to the research.

In the research work environment, potential or actual conflict of interest is sometimes less clear, but nonetheless important. A university researcher, for instance, may choose to be first author on a report for which junior colleagues developed new ideas and did most of the research work. Sexual exploitation of vulnerable colleagues (student researchers or research staff) is also a type of actual conflict of interest though it may not be labeled as such.

Research intensive universities formally explicate conflict of interest policies that relevant faculty and research personnel must abide by. For example, Stanford University’s detailed policy emphasizes financial gain issues and informs readers of ways to mitigate conflicts. The list of Further Readings at the end of this entry includes Stanford’s policy and other examples of institutional and organizational policy statements that address conflict of interest.

Education researchers, at times, serve on boards of trustees for foundations or on the boards of directors of nonprofit organizations or boards of for-profit corporations. These researchers are not directly involved in research in these contexts, but nonetheless may have potential conflicts of interest. The National Council of Nonprofits in the United States has issued policy guidance on conflicts of interest so as to inform the ways that board members think about the topic. The guidance includes recommendations that board members disclose potential conflicts of interest and that members abstain from votes and decisions that may involve a conflict of interest. Some research-focused organizations require board members to submit detailed statements listing potential conflicts of interest involving their employment, business ownership and involvement, investments and financial interests, and gifts and gratuities.

Peer review of proposals to support research may engender conflicts of interest. The U.S. Institute of Education Sciences, for instance, requires that reviewers recuse themselves from participating in a review panel in several different circumstances, including when they “have professional differences that could reasonably be viewed as affecting the objectivity of their review” (2006, p. 6) or have a financial interest in a for-profit organization that has an application before the panel.

People can go to prison for failure to disclose conflicts of interest when it is clear that they make personal gain despite their conflict. It is easy to find recurrent episodes of the so-called insider trading in the commercial sector, some of which have resulted in legal sanctions. There is no evidence of any such sanctions in education or other social research sectors.

Even when legal sanctions do not apply or are not pursued, however, social sanctions may apply when academic researchers are careless about conflicts of interest. For instance, in 2013, two economists who published evaluation research on the effectiveness of private prison programs initially failed to disclose that they had been financially supported by private prison services corporations for the research. A complaint alleged the failure to disclose the information violated policies of the economists’ university. The complaint and the university’s response received wide coverage in the news media.

Robert Boruch

See also American Educational Research Association; American Psychological Association; Ethical Issues in Educational Research; Institutional Review Boards; Interviewer Bias

Further Readings

American Economic Association. (n.d.). Disclosure policy. Retrieved from https://www.aeaweb.org/journals/policies/disclosure-policy

American Educational Research Association. (n.d.). Professional ethics. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/CodeofEthics/tabid/102

American Institutes for Research. (n.d.). Code of conduct. Retrieved from http://www.air.org/page/code-conduct

Institute of Education Sciences. (2006, January 24). Procedures for peer review of grant applications. Retrieved from http://www2.ed.gov/policy/rschstat/guid/ies/peerreviewgrants.pdf

National Council of Nonprofits. (n.d.). Tools … resources: Conflict. Retrieved from https://www.councilofnonprofits.org/tools-resources/conflict

Singapore Statement on Research Integrity. (2010, September 22). Retrieved from http://www.singaporestatement.org/statement.html

Stanford University. (n.d.). Research … scholarship: Conflicts of interest. Retrieved from https://doresearch.stanford.edu/research-scholarship/conflicts-interest

University of Oxford. (8 April, 2011). Illustrative examples of conflict of interest. Retrieved from https://www.admin.ox.ac.uk/researchsupport/integrity/conflict/policy/illustrativ

Jennifer A. Brussow Jennifer A. Brussow Brussow, Jennifer A.

Consequential Validity Evidence Consequential validity evidence

371

374

Consequential Validity Evidence

Consequential validity evidence provides information about the social consequences that result from using a test for a particular purpose. Various types of evidence can be presented to provide information about a test's consequential validity; these types of evidence include subgroup scores, results of test-based classification decisions (e.g., instructional or curricular differences, negative social consequences within a peer group, differences in opportunity), and errors in test use. Evidence supporting consequential validity is typically used to demonstrate how intended outcomes have been achieved, a lack of differential impact across subgroups, and the presence of positive and absence of negative systemic effects resulting from the testing program.

In 1989, Samuel Messick introduced the idea of a consequential basis for validity; since that time, there has been a great deal of scholarship produced regarding whether consequential validity evidence is needed to support the interpretation of a test's results. Some scholars argue that since consequential evidence deals with ethical rather than measurement considerations, it should not be considered as part of the validity argument. Others argue that ethical issues should be included in the scope of validity, and consideration of the social consequences resulting from a test's interpretation is an important ethical consideration when constructing a validity argument. The following sections of this entry outline some possible types of consequential validity evidence, then explore the main arguments forwarded by scholars in favor of including consequential validity evidence in the validity argument and by scholars in favor of considering consequences outside of the framework of validity.

Although scholarship continues to be published on both sides of this debate, in 2008, Gregory Cizek and colleagues found that just 2.5% of the 283 tests they reviewed provided this type of evidence in their validity arguments. Although consequential validity evidence has been indicated as a source of validity evidence in the *Standards for Educational and Psychological Testing*, these results clearly show that it has failed to gain traction as an important source of validity evidence.

Types of Consequential Validity Evidence

One of the most commonly reported types of consequential validity is differences in subgroup scores. Psychometric analyses that detect differential item functioning or differential test functioning can provide information on whether certain subgroups—especially protected subgroups—are consistently performing lower than reference groups. This type of bias both suggests that the test may suffer from excessive construct-irrelevant variance and indicates that consequences resulting from test classification decisions may differentially impact different subgroups.

A related source of consequential validity evidence is a description of the results of test-based classification decisions. For example, students classified into different proficiency bands as a result of their scores on an academic achievement test may receive different instruction and/or curriculum as a result of their performance. In some instances, decisions about students' placement in special education programs, more intensive supports, or alternative settings can be influenced by test scores. Lower-performing students may also experience fewer opportunities to succeed as a result of their low test scores. Additionally, low test scores can result in negative social consequences within students' peer groups; lower-performing students are sometimes ostracized as a result of their lower academic achievement.

Evidence regarding the interpretation of test results should ideally include information about consequences resulting from the recommended interpretation of test results as well as any interpretations that may extend beyond the recommendations of the test developer. Although assembling evidence for the consequences of the test's interpretations outside of the recommended interpretations may be challenging, there are several avenues that test developers can pursue. Consequential validity evidence for a new test can attempt to anticipate and describe these possibilities. Established tests' consequential validity evidence can collect evidence from the populations who are impacted by test classification decisions to provide concrete statistics regarding these consequences.

Some of the test’s uses and/or interpretations extending beyond the test developer’s intentions may constitute errors in test use. Scholars disagree on whether consequential validity evidence should be provided to describe the consequences of erroneous test uses, although most agree that erroneous test use does not need to be considered within the validity argument.

The Case for Consequential Validity Evidence

Messick’s 1989 validity framework presented validity as a unitary yet multifaceted concept that required many types of evidence to be adequately supported. Within this framework, construct validity was presented as the unifying force under which the rest of the validity framework resides. To argue for construct validity, many types of validity evidence must be brought to bear as part of the validity argument. To illustrate this idea, this work included a diagram that Messick termed the “progressive matrix,” which was intended to illustrate the way that types of validity evidence stood in relation to each other. This matrix included social consequences in its fourth and final cell, which represented the intersection between test use and a consequential basis. The other three cells included construct validity, construct validity plus relevance/utility, and value implications. In later works, Messick specified six types of evidence to support a validity argument: content, substantive, structural, generalizability, external, and consequential. All of these types of evidence were presented as aspects of construct validity.

Throughout his scholarship, Messick has argued that the social consequences of test use are an ethical matter that must be taken into consideration to argue for an interpretation of a test as valid. Messick argues that anticipated consequences of legitimate test use constitute part of the nomological network that provides the framework for construct theory and are therefore part of the unifying construct validity argument. Because these consequences are part of the nomological network, the consequences of a test’s use and interpretation must necessarily be sources of evidence for construct validity and, therefore, for the test’s value and worth.

Additionally, the social consequences of specific score interpretations contribute to score meaning and are therefore also part of the overarching validity argument, especially because their value implications may not line up with the construct’s implications for relative levels of the trait in question. Messick, Michael Kane, and other scholars have argued that negative consequences from a test’s uses can render a score use unacceptable and therefore invalid. Consequential validity evidence is especially important, given that test scores are typically used to make decision inferences about individuals at different score levels, and these decision inferences are at least partially based on the score users’ assumptions about the consequences resulting from different decisions. If decisions result in adverse conditions or negative systemic effects for all or most of the populations, the decision rules and therefore the test’s interpretation should be rejected as invalid. Kane proposes three major categories of outcomes that should be considered when assembling consequential validity evidence:

1. the extent to which the intended outcomes are achieved,
2. differential impact on groups (particularly adverse effect on legally protected groups), and
3. positive and negative systemic effects (particularly in education).

Both positive and negative as well as intended and unintended consequences resulting from the testing program should be considered when assembling consequential validity evidence as part of a validity argument.

The Case Against Consequential Validity Evidence

Although some scholars support the inclusion of consequential validity evidence in a validity argument, other scholars argue that consequential evidence, while important for consideration, should not be considered under the umbrella of validity evidence. Scholars arguing against the use of consequential validity evidence in the validity argument—such as James Popham, Lorrie Shepard, and others—assert that inserting social consequences into the concept of validity leads to further confusion surrounding the issue. The concept of validity has already undergone several transformations throughout its history, and many practitioners are currently unclear about its definition. Common misconceptions include reference to different types of validity rather than different types of validity evidence and statements about the validity of tests rather than the validity of score interpretations. Including social consequences as a required part of a validity argument may sacrifice the clarity of the definition of validity set forth in the 1985 *Standards for Educational and Psychological Testing*, which defined validity as “the degree to which th[e] evidence supports the inferences that are made from [test] scores" (p. 9).

A similar argument is that while consequences of test use and interpretation should be addressed by test developers and policy makers, this evidence should not be considered as part of the validity argument. By including consequential validity evidence, the validity argument is incorrectly expanded to include considerations of possible consequences that are outside of the scope of test development. While the test development process can assemble validity evidence showing that the test adequately and accurately measures the construct, thereby yielding accurate inferences about examinees' proficiency levels, policy decisions that are made regarding the consequences of those inferences cannot be rightly considered as part of the validity evidence for the test's interpretation and use. If consequences resulting from test scores are unreasonable, that does not necessarily impact the validity of test-based inferences about examinees' proficiency.

In fact, inclusion of consequential validity evidence can result in test misuse in terms of errors in procedure and/or policy and can detract from the validity of correct, legitimate use of the test.

An argument taking a different view of adverse consequences of test use posits that observed adverse consequences of test use only indicate a lack of validity if they are traceable to a gap in some other part of the validity argument such as construct underrepresentation or construct-irrelevant variance. These are both aspects of construct validity evidence that must be considered in a validity argument. This position argues that while consequences should be considered, consequential validity evidence should not be included in the validity argument; instead, sources of invalidity that lead to adverse consequences should be identified and accounted for using the other types of validity evidence outlined in Messick's framework or the 2014 *Standards for Educational and Psychological Testing*. By viewing challenges to the validity argument in terms of the other types of validity evidence, the issue of validity remains a question of scientific measurement rather than of ethics.

Jennifer A. Brussow

See also Construct-Related Validity Evidence; Content-Related Validity Evidence; Criterion-Based Validity Evidence; Unitary View of Validity; Validity; Validity, History of

Further Readings

American Educational Research Association, American Psychological Association, … National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Washington, DC: American Council on Education … National Council on Measurement in Education.

Newton, P., & Shaw, S. (2014). Validity in educational and psychological assessment. Thousand Oaks, CA: Sage.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. Educational Measurement: Issues and Practice, 16(2), 9–13.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16(2), 5–24.

W. Jake Thompson W. Jake Thompson Thompson, W. Jake

Construct Irrelevance Construct irrelevance

374

376

Construct Irrelevance

Construct irrelevance, as the name might suggest, refers to measuring phenomena that are not included in the definition of the construct. This is generally considered to be one of the two biggest threats to the validity of an assessment, along with construct underrepresentation. Although construct underrepresentation involves an assessment not adequately measuring all aspects of the construct of interest, construct irrelevance (sometimes referred to as construct irrelevant variance) occurs when an assessment is measuring more than just the construct of interest. This entry examines how construct irrelevance occurs and its implications and then provides illustrative examples of construct irrelevance.

Construct Irrelevance in Practice

The underlying assumption of all assessments is that the score that is produced reflects a test taker’s ability on the construct of interest. In an educational setting, this construct is often an educational content area. For example, if a student takes a math assessment, we assume that the score that student receives represents, and is an accurate measure of, his math ability.

As noted previously, construct irrelevance occurs when the score produced by the assessment is dependent on more than just the construct of interest. Continuing with the example of the math assessment, suppose that some of the questions were word problems. These word problems will require some level of reading and comprehension ability in order to understand the question and respond appropriately. This is construct irrelevance. We are not interested in measuring a student’s reading ability with this assessment. That is, reading ability is irrelevant to math ability, as we have defined the construct.

The implication of including these items on the assessment is that a student's score may not accurately reflect that student's ability on the construct of interest. Suppose a student with poor reading ability but high math ability took the assessment previously described. This student might be fully capable of performing the math necessary to correctly answer the word problems, but because of her poor reading ability, she doesn't understand the question and gets the item wrong.

As this type of construct irrelevant error variance accumulates over many word problems, the student's math ability score is going to be biased down. The student will get many questions wrong that someone with a high level of math ability should get correct, due simply to her reading ability. Thus, the score she receives is not an accurate representation of her true math ability. This can in turn influence which performance levels students are placed in, which then has implications for many of the decisions in educational settings that are based on student test scores.

This example clearly demonstrates the importance of construct irrelevance to the overall validity argument of an assessment. If an assessment contains variance that is irrelevant to the construct, the scores will be biased, and thus it becomes extremely difficult to justify using those scores to make decisions. Therefore, it is important to ensure that sources of construct irrelevance are minimized as much as possible to safeguard the validity of the scores and their intended uses. Common sources of construct irrelevance and methods used to detect some of these sources are discussed in the following section.

Examples of Construct Irrelevance

In operational psychometrics, construct irrelevance is most commonly associated with differential item functioning (DIF) and differential test functioning. This is because at their core, both DIF and differential test functioning represent construct irrelevance. Thus, possible sources of DIF are also possible sources of construct irrelevance. If performance on an item can be predicted by group membership after accounting for ability level on the construct on interest, then construct irrelevance is present. Most commonly, the groups of interest in these analyses are gender, ethnic, and socioeconomic groups.

When a reading assessment is administered, there is no intention for that assessment to also measure a student's gender or race. Thus, if one group (e.g., females) is favored over another (e.g., males) after accounting for ability level, construct irrelevant variance is introduced into the scores. Similarly, if an item or the assessment overall favors a particular racial or socioeconomic group after accounting for ability, then the assessment is not only measuring the construct of interest, but also the racial or socioeconomic group membership.

Construct irrelevance is not limited to group membership, however. For example, individuals' performance on an assessment may be influenced by their motivation or their preconceived notions about how they will perform (e.g., stereotype threat). Thus, the assessment would be measuring not only the construct of interest but also the individual's motivation.

This can also be true of tests that are overly long. If individuals get tired at the end of a long assessment, then the assessment now measures exhaustion in addition to the construct of interest. When providing scores, only the construct of interest should be included. Therefore, it is of the utmost importance construct irrelevance is considered throughout the assessment development from item writing (to lower the chance of construct irrelevance through DIF) to the blueprint of assessment (to avoid test designs that may invite additional sources of error).

W. Jake Thompson

See also Construct-Related Validity Evidence; Construct Underrepresentation; Differential Item Functioning; Threats to Research Validity; Validity

Further Readings

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Advances in Health Sciences Education, 7, 235–241.

Haladyna, T. M., & Downing, S. M. (2005). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23(1), 17–27.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13–23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.

Yi-Hsin Chen Yi-Hsin Chen Chen, Yi-Hsin

Isaac Y. Li Isaac Y. Li Li, Isaac Y.

Walter Chason Walter Chason Chason, Walter

Construct Underrepresentation Construct underrepresentation

376

380

Construct Underrepresentation

Construct underrepresentation occurs when a test does not adequately measure all aspects of the construct of interest. There are various sources of construct underrepresentation. This entry first discusses the sources of construct underrepresentation, the effects of construct underrepresentation on test use and score interpretation, and how to minimize construct underrepresentation. It then looks at specific ways of dealing with potential sources of construct underrepresentation in different types of assessment.

For a construct having multiple facets, when one of the facets is not tapped in the measurement, the construct is underrepresented. For example, if a particular test intended as a comprehensive measure of anxiety measures only psychological reactions and not emotional, cognitive, or situational components, it might underrepresent the intended construct. In addition, covering only trivial content in the curriculum will create construct underrepresentation.

Other times, construct underrepresentation is due to inadequate use of test questions. For example, a test of reading comprehension intended to measure children’s ability to read and interpret stories might not contain a sufficient variety of reading passages or might ignore a common type of reading material. An examination of sufficient length will be a fairer, more accurate, and reliable sample of important knowledge. Maldistribution of examination items leads to oversampling of some content areas and undersampling of others; too few test questions results in failure to adequately sample the learning content in the achievement domain desired. Consequently, the reliability of the examination suffers.

From the perspective of test development, multiple-choice tests are designed to measure skills ranging from lower order (e.g., recognition, recall) to higher order (e.g., reason, synthesis, application, evaluation) skills. However, it takes time, energy, and expertise to create multiple-choice items that tap higher order thinking skills; therefore, many multiple-choice tests overrepresent lower order skills and underrepresent higher order skills. If the test purporting to measure a broad range of cognitive skills employs few items assessing evaluation and application, its scores will be invalid because they underrepresent higher order thinking in the domain.

Construct underrepresentation also occurs when assessment objectives are deficiently considered. Items at a low level of cognitive function require only rote memorization to recall isolated facts that may not reflect the integrated knowledge to support critical thinking or problem-solving ability for real-world situations. A test full of this type of questions underrepresents the construct. Also, teaching to the test leads to scores that are an inaccurate reflection of the knowledge domain and leaving out items that require higher order cognition and problem-solving skills. When teaching to the tests, teachers focus on some subjects at the expense of others, some aspects of a subject at the expense of others, and some students at the expense of others. Consequently, learning becomes narrow, shallow, and transient as a result of construct underrepresentation.

Effects on Test Use and Score Interpretation

In education, test scores are interpreted, acted upon, and used as the basis for inferences and decision making. The extent to which these consequences are aligned with intended purposes and are appropriate and meaningful, given the available sample and data, is the scope of validity. Construct validity is about the meaning of the scores. Evidence is gathered to support the argument for solid validity in the measured construct that renders meaning for the scores. Construct underrepresentation occurs when important elements of the construct are missing from the measurement instrument and it cannot be eliminated but only minimized.

As a source of invalidity, construct underrepresentation negatively affects the soundness of score meaning, relevancy, and value implications. When construct representation is found to contribute to social consequences, the construct and/or the measure needs to be adapted in keeping with such a finding. For example, in cross-cultural comparisons, it is crucial to ask whether a less studied cultural group conceives of or values the construct in the same manner as the most often studied group upon which the construct and the measure were developed. The answer to this question reveals how much the obtained scores reflect construct underrepresentation and/or construct irrelevancy. The presence of construct underrepresentation cautions against direct test use for that group because the intended meanings of the scores may have been contaminated by the factors within the context.

How to Minimize Construct Underrepresentation

The breadth of content specifications for a test should reflect the breadth of the construct invoked in score interpretation. Equally important is for measurement research in general and construct validation in particular to entail multiple measures of each construct under investigation. Critically, test developers and users must be conscious of the relationship between score meaning and social consequences. Samuel Messick argued that if social consequences occur that are traceable to construct underrepresentation and/or construct-irrelevant variance, the construct and/or measure need to be modified to incorporate these findings; if they are not, then they are not part of validity.

During the process of validation, empirical evidence and a compelling argument are presented to support the intended inference and to show that alternative or competing inferences are not more viable. In particular, the degree to which construct underrepresentation and construct-irrelevant variance are problems needs to be analyzed. Validation puts both the test and the theory under scrutiny; and if research findings obtained from test scores and the theory disagree, then this discrepancy must be resolved by drafting a new test or postulating a new theory or both.

Because invalidity threats like construct underrepresentation can happen to both the test and the explanatory theory, validation as an ongoing process needs to account for both dimensions. Threats to the interpretability of obtained scores from construct underrepresentation can be minimized by clearly defining how particular psychological or educational tests are to be used. When constrained with a single test score, a strategy can be employed to triangulate on the referent construct by incorporating multiple formats of items or tasks in a composite total construct by incorporating multiple formats of items or tasks in a composite total score. Messick pointed out the potential role of social consequences in expounding the values and meanings of test score and test use. Therefore, considering values and social consequences in the validation process and minding potential impacts from legitimate/illegitimate use and interpretation of test scores is a way of minimizing construct underrepresentation.

When developing a criterion-based assessment, construct representation must be well reflected in test specifications, often created with the help of an in-depth needs analysis of the requirements from the test users and the skills and ability certain levels of obtained scores should possess. For higher level constructs, use of multiple-choice items can be limited and constructed-response question formats used more often. Test and item specifications can be further developed following field tests and feedback from stakeholders. Follow-up studies, as part of the validation process, are necessary for maintaining construct representation.

Rasch analysis is a powerful tool for evaluating construct validity. The Rasch model assumes a hypothetical unidimensional line along which persons and items are located according to their ability and difficulty magnitude. The items that fall close enough to the hypothetical line contribute to the measurement of the single dimension defined in the construct theory. Long distances between the items on the line indicate that there are big differences between item difficulties, so people who fall in ability close to this part of the line are not as precisely measured by means of the test. Gaps along the unidimensional continuum are indications of construct underrepresentation.

Dealing With Sources of Construct Underrepresentation

This section gives a few examples how the issue of construct underrepresentation is handled to illustrate the issues discussed previously.

Test Accommodation

Construct underrepresentation endangers validity of test accommodation. For example, if speed is part of the intended construct, it is inappropriate to allow for extra time, a common accommodation, in the test administration. Because speed will not be part of the construct measured by the extended-time test, scores obtained on the test with extended administration time may underrepresent the obtained on the test with extended administration time may underrepresent the construct measured with the strictly timed test. Similarly, it would be inappropriate to translate a reading comprehension test used for selection into an organization’s training program if reading comprehension in English is important to successful participation in the program.

Valid test accommodations avoid creating construct underrepresentation; those that reduce construct representation are invalid. Valid accommodations include offering an example in a reading comprehension test, enlarging the text print, or allowing the use of eyeglasses; invalid accommodations may include reading the text to a person who is visually impaired because doing so reduces construct representation by removing the element of text decoding.

Computer-Based Testing (CBT)

How test questions are developed, selected, and calibrated within a CBT can impact representation of the construct measured. Typical procedures adopted by CBT such as algorithmic item writing, computerized-adaptive testing (CAT), and unidimensional item calibration can potentially cause construct underrepresentation. First, the item selection algorithms used in adaptive testing can possibly lead to construct underrepresentation. In many CATs, an algorithm is used to select items or testlets to be administered to an examinee. The typical algorithm attempts to align item difficulty with estimated proficiency and limits how frequently a specific question can be administered. These activities can reduce content coverage (and thus construct representation) at the test score level for a particular examinee.

Although CAT algorithms include content constraints to ensure content coverage of items, it is important to evaluate the content quality of CATs administered to specific individuals and acknowledge that content quality and construct representation are continuous qualities, which are unlikely to be guaranteed within an item selection algorithm that is inherently binary. Therefore, we need evidence that construct representation does not differ across levels of proficiency and the CATs given are roughly parallel forms.

Second, with automatic item generation, which CBT uses to develop a test, it is possible that parallel items have differing statistical and substantive properties. Comprehensive field testing of generated items may circumvent this problem, but methods to determine or predict item properties need to be further developed in order to ensure that algorithm-produced items have desirable and expected in order to ensure that algorithm-produced items have desirable and expected psychometric properties. The validity of the resulting measurement will be undermined if the predicted item parameters do not adequately represent the true attributes of the generated item.

Third, construct underrepresentation in CBT can also stem from the use of an item calibration model that puts statistical criteria ahead of qualitative criteria, such as construct representation. For example, with a unidimensional item response theory model, construct-relevant items that do not fit the model may be eliminated and might result in a particular substantive area insufficiently represented and a poorly represented construct.

Language Use in Science Assessments

The absence of linguistic complexity from content area tests in text structures, genres, or styles of rhetorical organization common in the scientific discipline can potentially cause construct underrepresentation. In science learning, students are often expected to produce and comprehend explicit procedure recounts and/or research articles, arguments with claim and evidence, explanations, and comparisons. Therefore, if students are not expected to make meaning from argumentation and explanation texts on science tests, this may suggest that these assessments suffer from construct underrepresentation; that is, the inclusion or exclusion of specific text structures may pose a threat to the validity of these tests.

For test developers, because specific linguistic knowledge is a component of content area mastery and linguistic features are part of the target construct on content area tests, definitions of the science achievement construct used in assessment design should explicitly include a description of the linguistic features that are necessary for participation in grade-level written and oral discourse.

Both lexical and grammatical elements of language and text-level organizing structures (rhetorical organization and cohesion) should be analyzed. Work is needed to investigate the styles of rhetorical organization used in these assessments and to match this organization to that common in the measured domain. If text-level styles of organization such as argumentation are central to the domain but absent from content area tests, this may suggest language-related construct underrepresentation.

Game-Based Assessments and Simulation-Based Assessments

In game-based assessments use cases, construct-representation evidence for validity is found from domain analysis research. The evidence-centered assessment design framework includes elements and processes that embody this research. Failure to evoke aspects of the targeted capabilities constitutes construct underrepresentation. Construct representation may be improved by including interaction, an array of actions and representations, and open-ended spaces for assembling and carrying out strategies.

Simulation-based assessments are often used in medical fields, where validity research helps determine whether limitations in the simulation model led to construct underrepresentation. Only part of the real-world tasks can be simulated seamlessly with high fidelity. For example, the use of standardized patients—laypeople trained to portray real patients—provides a potentially valuable means for assessing skills such as the ability to collect a patient history and the ability to communicate with the patient. With standardized patients, it is difficult, however, to simulate abnormal physical findings. This limitation restricts the range of problems that can be presented, which in turn may reduce the likelihood that the examinee will check for abnormal findings—even though the examinee would have in the real world—and it also may lead examinees to record that those findings were absent despite the fact that they did not check for them.

The complexity of the problem of developing tasks and creating variables makes it clear that a lengthy program of test development and refinement is likely to be necessary before optimal solutions can be found for the problem of variable identification. Currently, answers remain context specific.

Yi-Hsin Chen, Isaac Y. Li, and Walter Chason

See also Construct Irrelevance; Construct-Related Validity Evidence; Cross-Cultural Research

Further Readings

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational … Psychological Testing (US). (2014).

Standards for educational and psychological testing. American Educational Research Association.

Drasgow, F. (Ed.). (2016). Technology and testing: Improving educational and psychological measurement. New York, NY: Routledge.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. Educational Measurement: Issues and Practice, 20(3), 16–25.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York, NY: Macmillan.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics, Vol. 26: Psychometrics (pp. 45–79). Amsterdam, the Netherlands: Elsevier Science B.V.

Ji Seung Yang Ji Seung Yang Yang, Ji Seung

Monica Morell Monica Morell Morell, Monica

Yang Liu Yang Liu Liu, Yang

Constructed-Response Items Constructed-response items

380

383

Constructed-Response Items

Constructed-response items refer to a wide range of test items that require examinees to produce answers in various formats; they are often contrasted or compared to multiple-choice (or selected-response) items in which examinees are required to select one or multiple appropriate options out of a given list. In practice, any items that do not take the selected response item format (e.g., multiple-choice or true/false items) can be referred to as constructed-response items. The term itself does not refer to a single format of items but implies flexibility in item formats. Because of this flexibility, “items” can be extended to “tasks” or “exercises” that are included not only in a written test but also in a performance test. The *Standards for Educational and Psychological Testing* defines constructed-response items, tasks, or exercises as follows:

> Items, tasks, or exercises for which test takers must create their own responses or products rather than choose a response from a specified set. Short-answer items require a few words or a number as an answer; extended-response items require at least a few sentences and may include diagram, mathematical proofs, essay, or problem solutions such as network repairs or other work products. (pp. 217–218)

In large-scale assessments and formative assessments, the constructed-response item format is primarily used to measure a complex set of skills or composition of knowledge that cannot be easily summarized in a short list of response options. Due to the complexity in the skill sets to be measured, writing items as well as grading and analyzing item responses are inevitably accompanied by a certain level of complexity (e.g., nature of multidimensional latent traits or skills). This entry reviews various formats for constructed-response items within different contexts and addresses the issues in developing, grading, and analyzing constructed-response items for educational assessments.

Item Formats for Constructed-Responses

The constructed-response item format exhibits great diversity, reflecting distinct characteristics of the content domain (e.g., language arts, mathematics, science, social studies, or computer science) and cognitive demand (e.g., knowledge, skill, or ability) to be measured. The taxonomy of constructed-response item format has been developed with contributions made by many researchers. With some variabilities, all taxonomies include a common dichotomy: whether an item requires open-ended or closed-ended response. Here, the distinction between open-ended or closed-ended lies on the existence of a well-defined (or constrained) scope for the set of skills or ability to be measured. In the meantime, the taxonomy developed by Steven Osterlind and William Merz includes reasoning competency (predictive, analytical, and interpretive reasoning, and factual recall) and cognitive continuum (convergent vs. divergent thinking), while Thomas Haladyna’s taxonomy includes scoring (objective vs. subjective) and outcome dimensions (product vs. performance). Extant studies identified the following numerous constructed-response item formats: anecdotal, cloze (embedded answers), demonstration, discussion, essay, exhibition, experiment, fill in the blank, grid in response, interview, observation, oral report, performance, portfolio, project, research paper, review, self/peer test, short answer, writing sample, and video-based task.

While the constructed-response item can take various formats, the most commonly used item formats in large-scale assessments (e.g., National Assessment of Educational Progress or Programme for International Student Assessment) are arguably limited to cloze, fill in the blank, grid in response, and short answer for all content domains. In addition, essay writing and oral exams (e.g., the speaking section in the TOEFL, a test of English as a foreign language) are used to measure language competency. Finally, it should be noted that the item formats and taxonomy for constructed-response items need to be and will be even more varied and extended as modern technology (e.g., computers, tablets, or motion-detection devices) plays a more significant role in the educational learning and assessment environment.

Developing Constructed-Response Items

The general standard procedure of item writing can be routinely applied to develop constructed-response items; in addition to that, specific recommendations and guidelines for developing constructed-response items have been offered in multiple resources as well. Drew Gitomer emphasized the coherence among task, rubric, and scoring apparatus that are built on well-defined constructs. Haladyna listed four categories of concerns that need to be addressed in writing constructed-response items—content, formatting/style, writing directions/stimulus, and context—and provided detailed guidance within each category. As can be seen from most of the literature on constructed-response items, a clear definition of the knowledge domain and skills to be measured, and an appropriateness of the item format is the key in writing constructed-response items. Both ensuring a precise mapping of items on the test blueprint with clear domain definitions and avoiding construct-irrelevant features are closely related to content or construct validity of test scores. In addition, the most distinctive and important aspect of writing constructed-response items is that the item writing cannot be separated from the concerns in scoring. Thomas Horgan and Gavin Murphy found more than 75% of textbooks recommend anonymous scoring, scoring one item at a time, and using a rubric or an ideal answer. Haladyna also suggests that item writers should “provide information about scoring criteria.” The following section addresses some methodological issues and practices in scoring constructed responses.

Issues in Scoring Constructed Responses

To analyze item response data for assessing quality of items (e.g., to obtain item difficulty and discrimination information), it is necessary to transform various forms of constructed responses (either product or performance) into numeric data. The assignment of numeric values to constructed responses can be achieved by using a rubric (or scoring guide) that can be holistic, analytical, or developmental. While a holistic rubric can be efficient in that it describes the overall quality of a product or performance in a single rating number (e.g., SAT writing), a drawback of the holistic rubric is that it does not provide precise and useful information. Accordingly, holistic rubrics are more appropriate for summative assessment rather than formative assessment in which explicit feedback is crucial to diagnose. On the other hand, an analytical rubric details characteristics of a product or performance with respect to each analytical category, so that multiple subrating numbers are included in the final rating. Assuming that the analytical rubric is well constructed with clearly distinctive categories and characteristics, using analytical rubrics can provide more explicit feedback. A developmental rubric describes the developmental characteristics of a product or performance either holistically or analytically.

Regardless of the type of rubric being used, multiple raters are typically involved to secure fairness and objectiveness in scoring; therefore, ensuring an appropriate level of interrater reliability through developing rubrics and training raters is critical. Agreement among raters can be quantified by using various coefficients. For dichotomously scored data, κ coefficients, intraclass correlation, or tetrachoric correlation are often used. For polytomous data, weighted κ, intraclass correlation, or polychoric correlation can be calculated. These approaches are, however, limited in that the analysis is somewhat post hoc and definitive. Furthermore, other reliability coefficients such as internal consistency or test–retest reliability are often calculated separately after a satisfactory level of interrater reliability is established.

Researchers can also apply the generalizability theory to investigate the reliability of constructed-response items in a more comprehensive and systematic fashion. In particular, when scoring constructed responses is involved in multiple facets of the measurement structure (e.g., multiple raters, multiple tasks, and multiple occasions), generalizability theory provides a theoretical and analytical framework that enables researchers to specify sources of measurement error and to gauge how much variability in scores can be attributed to different sources of error. A typical generalizability theory application is composed of two studies: a G study in which a variance component for each facet is estimated and a D study in which an optimal measurement structure to ensure a certain level of reliability can be determined given the variance components estimated in the G study. Through the D study, researchers can make decisions on how many raters, occasions, or items are needed to ensure a desired level of reliability considering cost efficiency.

With advances in technology, automated scoring has become an alternative approach to score constructed responses, particularly for large-scale assessment to save time and reduce the cost of resources while ensuring scoring consistency. A few large testing companies (e.g., Educational Testing Service, Pearson

A few large testing companies (e.g., Educational Testing Service, Pearson, Pacific Metrics) have developed computer programs for scoring essays, contents, and math. To ensure the quality of scoring, an automated scoring system is often adopted through a carefully composed protocol with excessive caution. Usually, it serves as a cross validation for human scoring at the beginning and is gradually extended to a substitution of human scoring after eliminating systematic errors in an iterative process. However, caution should be taken if human scoring is to be completely substituted by automated scoring because many large-scale assessments or standardized tests are used to make high-stakes decisions.

Analysis of Constructed-Response Item Data

Once scoring constructed responses is completed by raters, the next step is to evaluate the psychometric properties of the items using the scored item responses. Measurement theories such as classical test theory or item response theory provide analytical tools for assessing items.

Within the classical test theory framework, the item discrimination can be calculated as the correlation between item scores and total scores for both dichotomously and polytomously scored item responses. The item difficulty (also sometimes referred as item easiness) for a dichotomously scored item is defined as the percentage of examinees who answer to the item correctly. However, the possible rubric scores for constructed-response items often exhibit more than two categories; therefore, the ratio of the weighted sum scores to theoretically possible maximum sum scores across examinees can be alternatively used. For example, let there be 1, 2, and 3 rubric scores for an item, and observed numbers of examinees for a corresponding rubric category are 25, 5, and 20 for Item A and 15, 15, and 20 for Item B. The difficulty parameters for these two items can be calculated as follows:

$$1\times 25+2\times 5+3\times 203\times (25+5+20)=0.63,$$

$$1\times 15+2\times 15+3\times 203\times 15+15+20=0.70.$$

While these numbers reveal that Item A is more difficult than Item B, 0.63 and 0.70 do not provide further information for each specific score category.

In contrast, item response theory defines item discrimination and difficulty as parameters to model the relation between item responses and a latent trait or ability. Among various item response models, Fumiko Samejima’s graded response model, Erling Andersen’s rating scale model, and Geoff Masters’ partial credit model are often used for polytomously scored item responses. In case the order of rubric categories is not predetermined, R. Darrell Bock’s nominal categories model can be utilized to determine which category requires a higher level of latent trait or ability to order the categories. These aforementioned models extract different information from item responses based on different assumptions. For example, both rating scale and partial credit models are extended from Georg Rasch’s model that assumes the same discrimination parameters across items. On the other hand, nominal categories and graded response models allow varying discrimination across items. If there is a theoretical or empirical rationale for a categorized latent trait rather than a continuous latent trait, various cognitive diagnosis models can be also used for item analysis.

Another issue in analyzing constructed response item data is multidimensionality of latent traits because an item often taps on multiple skills or cognitive domains. With advances in multidimensional item response theory or item factor analysis, the measurement models become more flexible so that researchers can model more complex relations among latent traits.

Constructed-Response Items Versus Multiple-Choice Items

Because constructed-response items require more resources in scoring relative to cost-efficient multiple-choice items, there have been multiple studies to compare utility of the two types of items in large-scale assessment settings. However, consensus has not been made among researchers in educational testing because previous studies vary in terms of methods that include samples, subjects, and contexts. For example, while the scores on constructed-response items and multiple-choice items are highly correlated among the same group of examinees, gender is a factor that might be related to responses to different item formats according to Mark Pomplun and Nita Sundbye. Some also argue that carefully developed multiple-choices items can serve as a substitute for some short-answer items, particularly in large-scale assessments. Therefore, using constructed-response items in large-scale assessments is incorporated not because the constructed-response item formats are exclusively appropriate to measure the skill or domains but because using only multiple-choice items in assessments could introduce a less desirable phenomenon such as excessive focus on testing skills for multiple-choice items.

Ji Seung Yang, Monica Morell, and Yang Liu

See also Automated Essay Evaluation; Classical Test Theory; Cognitive Diagnosis; Difficulty Index; Discrimination Index; Generalizability Theory; InterRater Reliability; Item Response Theory; Multiple-Choice Items; Performance-Based Assessment; Portfolio Assessment; Reliability; Rubrics; Scales; Tests; True Score; True-False Items; Validity

Further Readings

American Educational Research Association, American Psychological Association, … National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Gitomer, D. H. (2007). Teacher quality in a changing policy landscape: Improvements in the teacher pool. Princeton, NJ: Educational Testing Service.

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. New York, NY: Taylor … Francis.

Hogan, T. P., & Murphy G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. Applied Measurement in Education, 20(4), 427–441.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. Educational Measurement: Issues and Practice, 33(2), 19–28.

Osterlind, S. J., & Merz, W. R. (1994). Building a taxonomy for constructed-response test items. Educational Assessment, 2(2), 133–147.

Pomplun, M., & Sundbye, N. (1999). Gender differences in constructed response reading items. Applied Measurement in Education, 12(1), 95–109.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. Handbook of Modern Item Response Theory, 29–164.

Jill Hendrickson Lohmeier Jill Hendrickson Lohmeier Lohmeier, Jill Hendrickson

Constructivist Approach Constructivist approach

383

386

Constructivist Approach

Constructivism is the epistemological idea that we construct our knowledge by linking new information to what we already know, rather than simply being passive recipients of knowledge. Thus, a *constructivist approach to education* is one in which educators encourage students to solve problems by actively engaging in tasks that require the learners to create an interpretation of the outside world in order to construct their own new knowledge rather than relying on instructor driven didactic methods only. In addition to being at odds with a more didactic approach to education, constructivism is a cognitive learning theory, in which active cognitive processing is a fundamental element of learning.

The emphasis on the learner’s active cognitive processes in constructivism is generally considered to be at odds with a behaviorist approach to learning and instruction, which in its simplest form suggests that learning occurs as a result of conditioning and can be measured by changes in behaviors. While behavior changes may occur as a result of learning with a constructivist approach, the focus for both the instructor and the learner is more on the cognitive processes and newly constructed knowledge used to solve problems and less on the behaviors that occur as a result of that learning. This entry describes the history of the constructivist philosophy, beginning with Lev Vygotsky and John Dewey. An explanation of constructivist learning philosophy, constructivist approaches to education, examples of constructivist activities, and criticism of constructivism are also included.

History of Constructivism

The constructivist approach has received much attention in the 21st century; however, constructivism was first formally discussed in early 20th century in Russia by Vygotsky, in the United States by Dewey, and later by Jean Piaget of Switzerland. Although Vygotsky and Dewey wrote about educational practice during the same era, it is unlikely that the two truly discussed their ideas due the political climate of the time. It is also unclear as to how familiar they were with each other’s work because Vygotsky’s work was not widely published until the 1970s, almost 40 years after his early death due to tuberculosis. Even so, both Vygotsky and Dewey encouraged educational practices in which the learners are engaged in thinking about practical, everyday, real problems rather than simply memorizing rote facts. Both Vygotsky and Dewey were concerned with creating good citizens through education, thus they believed that helping students become thinking adults who could solve novel problems was more important than simply telling students exactly what they should know.

Following Dewey’s work in the United States, in mid-20th century in Switzerland, Jean Piaget emerged as an advocate for constructivist approaches to education based on his studies of human development. Although Piaget’s theories primarily focused on developmental stages, his theories about learning included the ideas of assimilation and accommodation, which are constructivist in nature. Assimilation occurs when a child encounters something new and in order to understand it, the children incorporates it into the knowledge the child already has. Accommodation is an opposite approach to the construction of knowledge in which a learner modifies the learner’s interpretation of the world when encountering new information that does not fit into the knowledge base that the child already has. A simple example of these concepts is when a child encounters a dog that the child has never seen, the child may be told that it is a dog and add that to the child’s category of dogs. If the child encounters a coyote, thinks it is a dog, and then is told that it is actually a coyote, the child must add the category of coyotes to the child’s knowledge base. In both cases, the child is constructing the child’s own knowledge by linking the new information to the child’s own interpretation and understanding of the world.

Years before Vygotsky or Dewey began touting the virtues of active learning through practical problem solving, several other famous educators also integrated elements of what might be defined as a constructivist approach into their educational practice. Both the German Friedrich Froebel (the father of kindergarten) and the Italian Maria Montessori began suggesting the importance of active learning and problem solving for children in the 19th century. Their ideas, which included allowing children to work together to solve practical ideas, which included allowing children to work together to solve practical problems both indoors and outdoors through play and “work,” are consistent with the basic tenets of constructivism. Montessori’s philosophy of teaching was also child centered, such that children were encouraged to be active learners, presented with appropriate learning materials, so that they might fully explore and understand concepts as well as learn important life lessons.

Twenty-first-Century Constructivism

In the beginning of the 21st century, constructivism started becoming a catch phrase in educational theory, practice, and research. Educators and policy makers began to emphasize the importance of learning the so-called 21st-century skills, and there was increasing concern about the abilities of students to solve real-world problems. Although an emphasis on educational testing also became prevalent in the United States during this time, it was often at odds with educational researchers’ and philosophers’ consideration of using constructivist methods in order to increase students’ engagement and abilities to solve novel real-world problems. Although many administrators and thus teachers began emphasizing more instructor centered approaches focused on the content on standardized tests, constructivist instruction remained student centered, with the instructor more often acting as a guide or facilitator, rather than the authoritative driver of the lessons.

In constructivist-based lessons, students are given real-world problems to solve and are expected to begin by creating connections from the knowledge they have to new information, so that they might construct the knowledge they need to fully answer the problems. The connections formed between what they already know and the new information include their interpretations of the new ideas and experiences. The learners often need to find additional sources of information through this process. This is where the guide (or teacher) can be most useful. By guiding the discovery, the teacher can scaffold or start where the student’s knowledge level is and build from that level of knowledge. Using guided instruction, the teacher can also be sure that the students learn (or construct) the knowledge they need for a lesson.

Constructivist-based lessons are most often used in science classrooms, where students may be allowed to explore scientific concepts through the scientific method as well as through shared learning experiences with other students. Perhaps constructivism has become most popular in science classrooms because the materials lend themselves well to exploration and active learning.

Constructivism is sometimes considered “hands-on” learning and is often confused with active learning. Although active learning is an essential element in constructivist learning, the activity must occur foremost within the learners in their thoughts. Although it is easy to see engagement when learners are physically actively involved in the learning process, physical activity is not always essential for knowledge construction. A constructivist-based classroom is often filled with both physically and mentally active students. The learning process is expected to be interactive and dynamic, with students coming together to discuss and solve problems. New challenges and problems often emerge through the process. The constructivist approach relies on the process of evolving problem solving to create critical thinking skills as part of the newly constructed knowledge set.

Constructivist Lesson Plan Example

A teacher beginning a lesson in a constructivist classroom generally begins with a big picture concept. A high school science teacher or college professor might begin by talking about a concept such as sea level rise. After explaining the general idea of sea level rise, the instructor might ask the students to “pair and share” what they know about and wonder about sea level rise. The instructor might then ask them to work in small groups to come up with three questions they would like to know the answers about sea level rise. At that point, the instructor might ask the students to create a hypothesis about one of those questions and what might contribute to sea level rise. Then they might be asked to conduct an experiment in which they test that hypothesis in a laboratory setting.

Following the experiment, the students might be asked to find information that supports their findings. The students might then spend time discussing the question in the classroom, do Internet research on the topic, and write a response to the question. Following this work, the instructor might bring all of the groups together to discuss all of their findings and how they fit together. They then could have a classroom discussion about the scientific causes of sea level rise, the global causes of current sea level rise, and the potential impacts of sea level rise on different countries.

The constructivist approach suggests that the learners in this lesson would take knowledge that they had before the lesson about water, the sea, and other related issues and use it to interpret their findings and construct new knowledge about sea level rise and all of the connected issues that emerged during the lessons, the experiments, the presentations, and the discussions. This constructivist approach is in sharp contrast to a didactic lesson in which the teacher presents a lecture about sea level rise and its causes and impacts and then gives the students an in-class exam to assess learning.

Criticism of Constructivism

Constructivist approaches to education have been criticized for forcing learners to "reinvent the wheel." Critics have suggested that it is often more efficient to simply tell learners how things are rather than force them to discover what has already been discovered. For example, critics of constructivism might suggest that it is a waste of a student's time to expect them to rediscover Newton's laws, when an instructor can more simply explain the laws in lectures.

It is likely that constructivists would argue that they do not expect students to discover every fact on their own but expect that their need for that knowledge will arise from their attempts to solve real-world problems. When the students arrive at a point where they need to understand Newton's laws, then an instructor can serve as a guide and help them find the material that presents these ideas. The instructor can then allow the students to discuss and work together to construct their own interpretation and understanding of these laws.

Jill Hendrickson Lohmeier

See also Active Learning; Critical Thinking; Epistemologies, Teacher and Student; Learning Theories; Montessori Schools; Social Cognitive Theory

Further Readings

Dewey, J. (1966). Democracy and education. New York, NY: Free Press.

Fosnot, C. T. (2013). Constructivism: Theory, perspectives, and practice. Teachers College Press.

Glasersfeld, E. v. (1995). Radical constructivism: A way of knowing and learning. Washington, DC: Falmer.

Gunstone, P. J. F. R. F. (2013). The content of science: A constructivist approach to its teaching and learning. Routledge.

Steffe, L. P., & Gale, J. E. (Eds.). (1995). Constructivism in education. Hillsdale, NJ: Erlbaum.

Ultanir, E. (2012). An epistemological glance at the constructivist approach: Constructivist learning in Dewey, Piaget, and Montessori. International Journal of Instruction, 5(2), 195–212.

Jennifer A. Brussow Jennifer A. Brussow Brussow, Jennifer A.

Construct-Related Validity Evidence Construct-related validity evidence

386

389

Construct-Related Validity Evidence

Construct-related validity evidence demonstrates whether a test measures its intended construct, where a construct can be defined as a conceptual abstraction used to understand the unobservable latent variable that is responsible for scores on a given measure. Constructs are said to be situated within the nomological network, which was originally proposed by Lee Cronbach and Paul Meehl in 1955. *Nomologic* refers to rules of nature, and the nomological network situates a construct in terms of its relationship to other, known constructs and behaviors in order to provide a theoretical context for the construct. This theoretical context in turn suggests avenues through which construct-related validity evidence can be provided, for example, in terms of its relationship to other constructs or traits.

According to our current understanding of validity, construct validity is the only type of validity, and thus, construct-related validity evidence encompasses all possible types of validity evidence. Samuel Messick’s 1989 framework redefined validity as a unified concept by defining all validity as construct validity; this definition effectively subsumes all possible types of validity evidence into the larger category of construct validity evidence. However, the term *construct validity evidence* is also sometimes used to refer to specific sources of validity evidence; this sense of the term recalls the historical definition of construct validity as a specific type of validity. In earlier decades when validity was conceptualized as having multiple types, construct validity frequently appeared alongside content validity and criterion-related validity as one of the main types of validity, and construct validity had its own sources of validity evidence. Although this conceptualization of types of validity is not the modern view, construct-related validity evidence is still discussed in the literature. The following sections outline the sources of construct-related validity evidence, provide an overview of the historical definition of construct validity as a type of validity, and provide an overview of our current understanding of construct validity as the sole type of validity in the unified theory.

Sources of Construct-Related Validity Evidence

Construct-related validity evidence supporting an item's nomological validity attempts to provide quantitative evidence to position the construct within the nomological network. In order to assemble nomological validity evidence, it is useful to consider both convergent and discriminant validity evidence. Convergent validity evidence rests on the assumption that constructs that are closely related in the theoretical framework of the nomological network should also be correlated when measured in reality. Convergent validity evidence can be provided in terms of a measure's correlation with other measures with strong validity arguments that assess theoretically related constructs. For example, if the construct of intelligence is thought to be closely related to working memory, then examinees' results on a test thought to measure intelligence should be highly correlated with their results on a measure of working memory.

Discriminant validity evidence is the counterpart of convergent validity evidence. This type of validity evidence is used to demonstrate that constructs that have no relationship or an inverse relationship in the nomological net are also not correlated in reality. Discriminant validity evidence can consist of a measure's low or negative correlation with other measures assessing theoretically opposed concepts. For example, if extraversion and introversion are assumed to lie at opposite ends of a spectrum, then examinees' results on a measure of extraversion should negatively correlate with their results on a measure of introversion.

In order to provide an approach to assessing the construct validity of a measure or set of measures, Campbell and Fiske developed the multitrait–multimethod matrix (commonly referred to as MTMM) in 1959. The multitrait–multimethod matrix provides a way to track correlations across multiple measures, measuring the same construct via different methods and different constructs by the same method. Through this process, both convergent and discriminant validity evidence is collected.

Another method of collecting construct-related validity evidence is to observe the effect of experimental variables on test scores. For example, if a test is designed to measure participants’ skill in two-digit addition problems, one would expect that practice in solving these types of problems would improve test scores. By collecting data from the measure before and after participants take a practice session, the researcher can assess whether the practice impacts test scores. If practice on the skill in question fails to improve test scores, the measure may not actually be assessing the construct in question. Of course, this example would not hold true for measures of constructs that would not be expected to change with practice, such as most personality traits. In this case, evidence that practice fails to affect participants’ scores would constitute validity evidence supporting the construct.

Other commonly used sources of construct-related validity evidence include statistical analyses such as factor analysis and structural equation modeling. By conducting a factor analysis to determine how much of the variance a factor accounts for, the researchers can provide evidence to support the presence of their construct and the adequacy of their measure. Such empirical data have been frequently used in validity arguments for decades and continue to be a popular source of validity evidence.

If content-related validity evidence turns out to yield negative findings that disconfirm the hypothesis, the researcher must consider the possible implications of this finding. Either the construct is improperly defined within the nomological network, the construct is well defined but the measure either assesses a different construct or is overly subject to construct-irrelevant variance, or the construct is poorly defined and the measure assesses a different construct. Negative construct-related validity evidence should always prompt researchers to more closely examine the construct and measure for which they are seeking to construct a validity argument.

Historical Definition

The idea of construct validity was formulated by Cronbach and Meehl for the first edition of the *Standards for Educational and Psychological Testing,* which was published in 1954. The idea of construct validity was more abstract than previous ideas about validity types and was motivated at least partially in order to deal with the perplexing issue of validating personality tests, which criterion and content validity were ill-equipped to handle. This initial definition called for both logical and empirical evidence to justify the inference of an underlying construct given test performance. However, construct validity itself was poorly defined, and it was presented as a fallback option of sorts. According to this initial edition of the *Standards*, logical approaches to validation were suitable for achievement or proficiency tests, empirical approaches were suitable for tests of aptitude or disorder, and both approaches were to be employed for tests of more nebulous concepts like personality—hence, construct validity. By using both logical and empirical approaches, researchers could attempt to simultaneously validate the test and its underlying theory.

Cronbach and Meehl’s 1955 paper attempted to more clearly define the idea of construct validity and the processes for collecting construct-related validity evidence. This landmark publication introduced the idea of the nomological network and suggested several different sources of construct validity evidence. These sources included group differences in test scores, relationships between test scores as expressed by correlations and factor analyses, item correlations within tests, performance stability over time, and analysis of cognitive processes underlying test performance. These types of validity evidence are still important to include when constructing validity arguments. Indeed, Cronbach and Meehl also put forth guidelines for what to include in a validity argument. In order to construct a validity argument, researchers were to explain the proposed interpretation of test scores in terms of the construct and its theoretical surroundings in the nomological net, express how adequate they believed the validity argument substantiated these claims, and detail the evidence and logical reasoning that supported the validity claims.

Current Definition

Messick’s 1989 chapter on validity marked the field’s move away from an understanding of types of validity. Instead, Messick’s unifying theory of validity promoted construct validity to the encompassing idea that all types of validity evidence support. With this understanding of validity as a unitary concept, all validity evidence can be said to be construct validity evidence. In the years since Messick’s publication, many scholars have debated the nature of construct validity, and several positions regarding construct validity theory have emerged. Brief summaries of several recent papers typifying some of these positions will follow.

Robert Lissitz and Karen Samuelson argue against the need for nomological networks, claiming instead that construct definition can take place without the need for the theory building originally envisioned by Cronbach and Meehl. Especially in the case of academic achievement tests, it is possible for the construct test measures to be logically extrapolated from its content rather than positioned in an external theoretical environment. A similar yet diverging viewpoint has been expressed by scholars such as Susan Embretson, James Pellegrino, and Joanna Gorin, who believe that while theory continues to be critical to construct validity evidence, the theories underlying constructs could best be understood in terms of internal underlying processes such as cognitive processes, skills, and knowledge. This understanding of constructs in terms of their internal processes also negates the need for the nomological network.

Another challenge to the nomological network— and to the definition of validity more generally—came from Denny Boorsboom, Gideon Mellenbergh, and Jaap van Heerden in 2004. These researchers contended that while our current understanding considers validity to be an epistemological matter, it should more accurately be considered an ontological concern. That is, instead of being concerned with how we know things, validity should be concerned with how things actually are. If an ontological definition of validity is accepted, then only two conditions must be met in order for a test to be considered valid. The construct in question must exist, and variation in test scores must be caused by that construct. This definition greatly reduces the scope of validity as a concept, sidestepping many of the concerns about inclusion of social issues in validity arguments.

This brief summary of several positions on validity from the 2000s and 2010s illustrates how construct validity continues to be a contentious issue within validity scholarship. From the foundational idea of the nomological net through the more recent subsummation of all other types of validity into the overarching idea of construct validity, all of the assumptions underlying construct validity are being challenged in the literature. Indeed, because all validity is construct validity under the current unified understanding of validity theory, construct validity and its sources of evidence will likely continue to evolve through time.

Jennifer A. Brussow

See also Consequential Validity Evidence; Content-Related Validity Evidence; Criterion-Based Validity Evidence; Multitrait–Multimethod Matrix; Unitary View of Validity; Validity; Validity, History of

Further Readings

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological Review, 111(4), 1061–1071.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher, 36(8), 437–448.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Washington, DC: American Council on Education … National Council on Measurement in Education.

Newton, P., & Shaw, S. (2014). Validity in educational and psychological assessment. Thousand Oaks, CA: Sage.

Chi Yan Lam Chi Yan Lam Lam, Chi Yan

Consumer-Oriented Evaluation Approach Consumer-oriented evaluation approach

389

392

Consumer-Oriented Evaluation Approach

The consumer-oriented approach to evaluation is the evaluation orientation advocated by evaluation expert and philosopher Michael Scriven. The approach stems from the belief that evaluation ought to serve the consumer, that is, the ultimate end user of the particular object under evaluation, the evaluand—be it a program, a curriculum, a policy, a product, or a service. This entry first discusses the history and the key aspects of the consumer-oriented evaluation approach, including the centrality of the consumer, the goal of the evaluation, and the role of the evaluation and the evaluator. It then looks at the techniques used in consumer-oriented evaluation, the checklist developed by Scriven for this evaluation approach, and the advantages and challenges of the approach.

The consumer-oriented evaluation approach arose in the 1960s in reaction to the then-prevailing stances that saw evaluation as an exercise in value-free measurement of whether program goals were achieved. The consumer-oriented evaluation approach reminds evaluators, and those who commission and use evaluation, that an evaluation ought to produce a determination about the merit, worth, and/or significance of the evaluand and that the basis of evaluation ought to be referenced to the needs of consumers.

Centrality of the Consumer

At the core of the consumer-oriented evaluation approach is the stance that evaluation should be oriented toward the needs of the consumer. Scriven argues that an evaluation’s task and goal should be directed toward the consumer (end user) primarily and, to a much lesser extent, the program developers and other stakeholders. Scriven recognizes that consumers’ values may not always align with the values of developers, funders, or even the delivery partners. The author also observes that the consumer is not necessarily concerned with goals that program developers set out to achieve with an evaluand nor should they have to contend with what developers’ intentions are. Rather, what truly matters to consumers is that an object has value, that is, merit, worth, and/or significance.

Goal of Evaluation

The significance of the consumer-oriented evaluation approach is best understood in context of the historical confluences that gave rise to it. One major source has been the limitations and flaws associated with objective-oriented approaches to evaluation or what has sometimes been referred to as Tylerian approach to evaluation. Proponents of Ralph Tyler’s approach to evaluation see it as the determination of whether objectives have been achieved or not.

Scriven’s critique of Tyler’s approach is that conceptualizing evaluation in a goal-oriented way is narrow, for goals, as prescribed by program developers, can be flawed, incomplete, unrealistic, or inadequate in addressing the social ills that prompted the creation of the intervention in the first place. Evaluating in this fashion ignores the true needs of the consumer. This stance is echoed in contemporary discourse that emphasizes the importance of placing the learner first.

The methodological implication is that evaluation is not merely a technical exercise in measurement between what was set out and what was the case in reality, as is the case with evaluation conducted in the Tylerian tradition, but in bringing evidence to bear in reaching an informed judgment about an object’s value independent of what developers set out to do. Scriven implores evaluators to understand all effects of an intervention, unconstrained by what developers had sought to achieve, and assess the needs of the users. On the bases of these two assessments, the evaluator advances a judgment concerning the value of the object. The determination of merit, worth, and significance of an evaluand is the singular goal of evaluation.

Role of Evaluation

Scriven further distinguishes the goal of evaluation from the role of evaluation. The author identifies two legitimate roles of evaluation, that of summative and formative evaluation. Summative evaluation advances a summary judgment concerning the overall value of the evaluand. For practical reasons, a summative evaluation is performed when the object is ready to be evaluated summatively, that is, when the evaluand has developed fully and when the evaluand is operating with sufficient regularities in its operation and is producing stable effects. To help program developers ready an evaluand for summative evaluation, an evaluation may be conducted formatively to identify shortcomings and deficiencies.

In either formative or summative evaluation, the evaluation activities are not materially different; the two only differ in purpose. Put simply, when the customer tastes the soup, it is summative evaluation. When the chef tastes the soup just before serving it, it is formative evaluation. In this way, summative evaluation constitutes the core of the role of evaluation in the consumer-oriented approach to evaluation.

Role of the Evaluator

Scriven sees that evaluation—and by extension, the evaluator—carries the ethical and moral imperative to determine whether an object contributes to the welfare of consumers. To that end, Scriven sees the proper role of the evaluator to be that of “enlightened surrogate consumer”; the evaluator discharges such responsibility by making informed judgments on consumer’s behalf. In his writing, Scriven often cites the magazine *Consumer Reports* as illustrative of the consumer-oriented approach to evaluation.

Techniques

To aid in putting the consumer-oriented approach to evaluation in practice, Scriven advances several techniques.

Goal-Free Evaluation

Scriven advances the notion of a goal-free evaluation in an attempt to offer an alternative to goal-based approaches to evaluation. In a goal-free evaluation, the evaluator ignores the stated program goals on purpose. Instead, the evaluator investigates all possible outcomes—both anticipated and unanticipated—of a program. According to Scriven, advantages of goal-free evaluation “are that it is less intrusive than goal-based evaluation; more adaptable to mainstream goal shifts; better at finding side effects; less prone to social, perceptual, and cognitive bias; more professionally challenging; and more equitable in taking a wide range of values into account” (cited in Stufflebeam … Shrinkfield, 2007, p. 374).

Needs and Needs Assessment

One of the advantages to adopting program developers’ goals in an evaluation is that assumptions about what constitutes valuable or meaningful outcomes have been made ahead of time. In the absence of adopting developers’ goals in a goal-free evaluation, the quandary arises of whose values ought to be represented in an evaluation and by what means they could be established. Scriven resolves this issue by placing the onus upon the evaluator to explicate the needs of the consumer through a needs assessment.

Scriven is specific in how he defines a need. Consistent with the author’s stance on orienting the evaluation toward the consumer, the author defines a need as “anything essential for a satisfactory mode of existence, anything without which that mode of existence of level of performance would fall below a satisfactory level” (cited in Stufflebeam … Shrinkfield, 2007, p. 375). A need defined in this way carries the notion of what is essential or necessary to the consumer. An example of such a need would be functional literacy. The findings of a needs assessment provide the basis to compare against observed outcomes in an evaluation.

Key Evaluation Checklist

Another major contribution Scriven has made to advance the approach is the creation of the Key Evaluation Checklist. The Key Evaluation Checklist draws together a comprehensive list of considerations and action items that the author views to be essential to conducting evaluation in ways consistent with the consumer-oriented approach. The checklist is organized into four sections and comprises 18 checkpoints. The remainder of this section summarizes each of the major sections.

Section A: Preliminaries

Section A: Preliminaries

The first section of the checklist invites the evaluator to consider those issues that would have bearing on the design, execution, and reporting of the evaluation itself. Three checkpoints are identified: creating an executive summary of the most pertinent information concerning the evaluation itself; clarifying the intended audience of the evaluation, the role of the evaluator, stakeholders of the program, and the questions the evaluation is to answer; and, finally, the design and methods that would be employed to answer those questions.

Part B: Foundations

The second section invites the evaluator to establish a detailed description of the evaluand. Five checkpoints comprise this section: establishing the background and context surrounding the evaluand; defining and describing the evaluand and its composition; identifying the consumers or what Scriven sometimes refers to as “impactees”; uncovering what resources are made available to enable operation; and, finally, what values (needs) ought to be used in the evaluation of the evaluand.

Part C: Subevaluations

The third section concerns the processes of constructing evaluative claims. Five checkpoints comprise the section: establishing program processes, specifically around the means by which the evaluand achieves intended goals; establishing outcomes; establishing the costs associated with operating the evaluand (which can manifest in different forms, from monetary, nonmonetary, and nonmonetizable costs to direct, indirect, maintenance, and residual costs); comparing observations made of the evaluand to the needs and expectations put forth by consumers; and finally, establishing the extent to which claims can be generalized.

Part D: Synthesis

The last section concerns the construction of evaluative conclusions and implications stemming from the evaluation inquiry. Five checkpoints comprise the section: advancing a synthesis claim into the overall value of the evaluand; advancing recommendations, explanations, predictions, and redesigns, if appropriate; concerning the evaluand; reporting on the evaluation; and finally, subjecting the evaluation itself to scrutiny by engaging in a meta-evaluation process.

In sum, the four sections of the checklist advance a methodology for conducting a consumer-oriented program. The document that discusses the Key Evaluation Checklist is comprehensive and is freely distributed via the Evaluation Checklist website hosted by Western Michigan University.

Advantages and Challenges

The primary advantage of the consumer-oriented approach to evaluation is that it produces a comprehensive account and assessment concerning the value of an evaluand. The findings from such evaluations serve an important function in protecting consumer interest, a laudable goal. The benefit of the consumer-oriented evaluation comes from the systematic and comprehensive nature of the approach, which itself is grounded in philosophical arguments concerning the fundamental goal and role of evaluation. The comprehensive nature of the consumer-oriented approach to evaluation also imposes challenges in its execution. Satisfying the approach fully requires a highly competent evaluator and sufficient resources.

Chi Yan Lam

See also Formative Evaluation; Goal-Free Evaluation; Program Evaluation; Summative Evaluation

Further Readings

Scriven, M. (2003). Evaluation theory and metatheory. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), International handbook of educational evaluation (pp. 15–30). New York, NY: Springer.

Scriven, M. (2013, March 22). Key evaluation checklist [Draft; latest version]. Retrieved from http://www.michaelscriven.info/images/KEC_3.22.2013.pdf

Shadish, W. R., Jr., Cook, T. D., & Leviton, L. C. (1991). Foundations of program evaluation: Theories of practice. Thousand Oaks, CA: Sage.

Stufflebeam, D. L., & Shrinkfield, A. (2007). Evaluation theory, models, and application. San Francisco, CA: Jossey-Bass.

Hsiu-Fang Hsieh Hsiu-Fang Hsieh, Hsiu-Fang

Sarah Shannon Sarah Shannon Shannon, Sarah

Content Analysis

Content analysis

392

394

Content Analysis

Content analysis is an analytic method used in either quantitative or qualitative research for the systematic reduction and interpretation of text or video data. Data can be generated from a variety of sources including (a) individual or focus group interviews; (b) responses to open-ended survey items; (c) text from social media; (d) printed materials such as research articles, newspapers, or books; (e) video-taped simulations; or (f) naturally occurring conversational events. It is also used in case study research. The aim of content analysis is to describe data as an abstract interpretation.

Use of content analysis as a research technique dates to the 1900s when it was used in communication research primarily to describe the quantity (frequency) rather than quality (meaning) of content contained in textual data. Since this early use, qualitative content analysis has gained popularity as a means to interpret data by identifying codes and common themes (manifest content) and then constructing underlying meanings (latent content). Content analysis is estimated to have been used as a qualitative analytic method in more than 3,000 research studies between 2005 and 2015 in such diverse fields as education, business, economics, social work, social science, and health sciences, including nursing, psychology, medicine, rehabilitation, gerontology, and public, environmental, and occupational health.

At least three distinct approaches to content analysis have emerged. These approaches differ in terms of study design, sampling decisions, and analytic strategies used, particularly how coding schemes are developed. The selection of approaches to content analysis largely depends on the research purpose and the availability of existing knowledge in the area of interest, particularly related models or theories. When existing knowledge around a phenomenon of interest is largely absent and the purpose of a study is to create knowledge, an *inductive approach* or *conventional qualitative content analysis* is appropriate where codes and themes are generated directly from the data.

When prior research or theory exists and the purpose of the research is to confirm, expand, or refine this existing understanding of a phenomenon, a more *deductive approach* or *directed qualitative content analysis* is appropriate using existing knowledge or theory to build the initial coding structure. When quantification of a specific content is desired, a summative content analysis approach is appropriate to identify and tally keywords or concepts.

As with any research method, sampling decisions are critical to meet study goals when using content analysis. Generally, sampling in a qualitative design seeks to maximize diversity of data around the phenomena of interest. Sample sizes may vary considerably when using content analysis depending on the research question. To understand a complex emotional event, researchers might conduct in-depth interviews with a small number of participants, while to understand what terms are used to describe a physical symptom, researchers might analyze written responses to an open-ended survey item from hundreds of participants. Using a directed content analysis approach, a researcher might purposively sample a particular group to refine or extend existing knowledge or theory about a particular phenomenon to a new population.

The development of the initial coding scheme and overall approach to coding differs depending on the specific content analysis approach chosen. With a directed content analysis approach, the researcher develops an initial coding scheme from existing theory or knowledge, using the data to modify or expand these codes. In a conventional content analysis approach, the initial coding scheme emerges from the data. With either approach, generally, it is helpful to first immerse oneself in the data to obtain a sense of the whole. Then data are coded through an iterative process. It is important to identify a consistent unit of coding, which might range from a single word to short paragraphs. Coding serves to reduce and condense the data based on its content and meaning. Finally, the relationships between codes are constructed by arranging them within categories and themes.

The process of abstraction from the raw data to meaningful themes requires establishing trustworthiness through strategies to ensure credibility, transferability, dependability, confirmability, and authenticity. A checklist for each phase of data analysis is helpful. In the data preparation phase, data collection methods, sampling strategies, and selection of units of analysis should be reviewed for trustworthiness. In the data organization phase, trustworthiness issues relate to categorization and abstraction, interpretation, and representativeness of results. Assessment of intercoder reliability is important in content analysis to establish credibility of the analytic process and findings.

Results of content analyses are presented through descriptive writing but should be complemented with figures and tables as appropriate. Examples include conceptual diagrams showing the relationships between codes and themes or tables showing codes in rank order of use, potentially for different groups of study participants. Although direct data quotes are used to illustrate findings, interpretation and presentation of the findings is essential. Commercial qualitative research software options are increasing that assist with managing data coding. These programs are helpful for handling large amounts of data and recognizing subtle patterns, but they are not substitutes for actual data analysis.

Qualitative content analysis has been criticized for lacking depth in abstraction. It is also of limited use for developing theory, in contrast to grounded theory methodology. Content analysis has several distinct advantages, however: (a) the analytic approach to data is unobtrusive and nonreactive, (b) novice researchers can learn basic techniques quickly, in contrast to other qualitative methodologies such as phenomenology where deep understanding is sought, and (c) it is more time efficient than methods such as ethnography where sustained immersion in the field is required. When choosing content analysis as an analytic approach, researchers should clarify which approach matches their research question, goal, and overall purpose. Content analysis remains one of the most widely used research strategies because it is fast and effective for finding patterns within multiple types of qualitative data.

Hsiu-Fang Hsieh and Sarah Shannon

See also Case Study Method; Qualitative Data Analysis; Qualitative Research Methods; Quantitative Research Methods; Trustworthiness

Further Readings

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. Nursing Research, 57(2), 113–117. doi:10.1097/01 .NNR.0000313482.33917.7d

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. Nurse Education Today, 24(2), 105–112. doi:http://dx.doi.org/10.1016/j.nedt.2003.10.001

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277–1288. doi:10.1177/1049732305276687

Kohlbacher, F. (2006). The use of qualitative content analysis in case study research. Forum: Qualitative Social Research, 7(1). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/75/153

Mayring, P. (2000). Qualitative content analysis. Forum: Qualitative Social Research, 1(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1089/2385

Morgan, D. L. (1993). Qualitative content analysis: A guide to paths not taken. Qualitative Health Research, 3(1), 112–121.

Sara E. Witmer Sara E. Witmer Witmer, Sara E.

Heather Schmitt Heather Schmitt Schmitt, Heather

Content Standard

Content standard

394

395

Content Standard

A content standard is a general statement that describes what students should know and be able to do following their participation in educational programming. Content standards are developed to offer consistency and coherence to educational programming, to eliminate redundancy in content covered over time, and to provide a foundation for the development of effective instructional and assessment programs. In contrast to achievement standards and/or performance standards, which describe the specific level at which students are expected to perform, content standards describe more generally what students are expected to learn. Within K–12 educational settings, content standards are typically developed to be grade level and discipline specific and are organized in a way that reflects a logical progression of essential knowledge and skills within a given content area. An example of a first-grade English language arts content standard from the Common Core State Standards is: “Ask and answer questions about key details in a text.” An example of an eighth-grade English language arts content standard is: “Cite the textual evidence that most strongly supports an analysis of what the text says explicitly as well as inferences drawn from the text.” The current entry describes recent trends in the development of content standards that correspond to the standards-based reform movement, considerations for diverse learners, and ongoing tensions in the development and use of content standards in K–12 school settings. Although content standards may be developed for a variety of educational programs (e.g., early education, adult education, graduate education), across a variety of countries, and by a variety of organizations, the focus of the current entry is on content standards as they relate to K–12 education in the United States.

content standards as they relate to K–12 education in the United States.

Standards-Based Reform and the Development of Content Standards

During the 1980s and 1990s, widespread concern with the status of U.S. public education relative to other countries ushered in educational reform efforts, with the intent to improve teaching and learning. At the core of these efforts was the standards-based reform movement. According to standards-based reform, student achievement will rise when (a) high expectations for student learning are clearly articulated, (b) assessment programs are designed to measure student progress toward those expectations, and (c) consequences are attached to student achievement, as measured by the assessment programs. The 2001 reauthorization of the Elementary and Secondary Education Act, namely, the No Child Left Behind Act of 2001, correspondingly required states to develop challenging academic content standards in English/language arts, math, and science, along with assessment programs that measured student progress toward those standards. Schools and teachers were offered flexibility in how they taught, but all students were expected to learn, at a minimum, the content articulated in the standards. Various consequences were applied to schools that did not demonstrate appropriate adequate yearly progress. In addition to the federally required content standards, some states and districts developed content standards in disciplines outside of those mentioned in the No Child Left Behind Act, such as health, fine arts, social studies, and citizenship.

Given a growing concern that states had disparate content standards and achievement standards at each grade level, the Council of Chief State School Officers and National Governors Association Center for Best Practices collaborated in 2009 to develop and validate a common set of standards. The Common Core State Standards were designed by a team of teachers, school administrators, and other experts in education from 48 different states. In the 2015reauthorization of Elementary and Secondary Education Act, namely, the Every Student Succeeds Act of 2015, an emphasis on state approval of content standards was maintained, along with the requirement for assessment programs to be developed and implemented to monitor student progress toward those standards.

Considerations for Diverse Learners

In the United States, federal law has required that all students, including those with disabilities, have access to learning and assessment according to the same grade-level content standards. This requirement was intended to ensure that all students experience the intended benefits of standards-based reform (i.e., improved teaching and learning). Through the provision of appropriate accommodation supports and specially designed instruction, students with disabilities are expected to achieve the same educational outcomes as their peers without disabilities. To facilitate access to the content standards for students with particularly significant cognitive disabilities, some states developed extended standards linked to the original content standards. The following is an example of an Ohio Grade 8 content standard in reading, along with the associated extended standard for students with significant cognitive disabilities: “Cite the textual evidence that most strongly supports an analysis of what the text says explicitly as well as inferences drawn from the text” and “Identify details that support answers to literal questions.” Goals that are written as a part of Individualized Education Programs that are developed for students receiving special education services are expected to be linked to the content standards.

Ongoing Tensions and Concerns

In the United States, content standards are typically developed and validated using an iterative process that involves input from a variety of individuals, including scholars, teachers, and the general public. Some scholars have expressed concern about the development of a particularly broad set of standards that fails to foster depth of student knowledge within each academic discipline. Finally, it is important to note that the impact of content standards on student learning may depend not only on the quality of the standards but on the extent to which teacher professional development, instructional materials, and assessment are aligned to those standards.

Sara E. Witmer and Heather Schmitt

See also Alignment; Common Core State Standards; Curriculum; No Child Left Behind Act; Standards-Based Assessment; State Standards

Further Readings

Browder, D. M., Wakeman, S. Y., Flowers, C., Rickelman, R. J., Pugalee, D., & Karvonen, M. (2007). Creating access to the general curriculum with links to grade-level content for students with significant cognitive disabilities: An explanation of the concept. The Journal of Special Education, 41, 2–16. doi:10.1177/00224669070410010101

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new US intended curriculum. Educational Researcher, 40(3), 103–116.

Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. Educational Evaluation and Policy Analysis, 31, 238–268. doi:10.3102/0162373709335465

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of U.S. mathematical and science content standards from an international perspective. Journal of Curriculum Studies, 37, 525–559. doi:10.1080/0022027042000 294682

Tonya Rutherford-Hemming Tonya Rutherford-Hemming Rutherford-Hemming, Tonya

Content Validity Ratio Content validity ratio

396

398

Content Validity Ratio

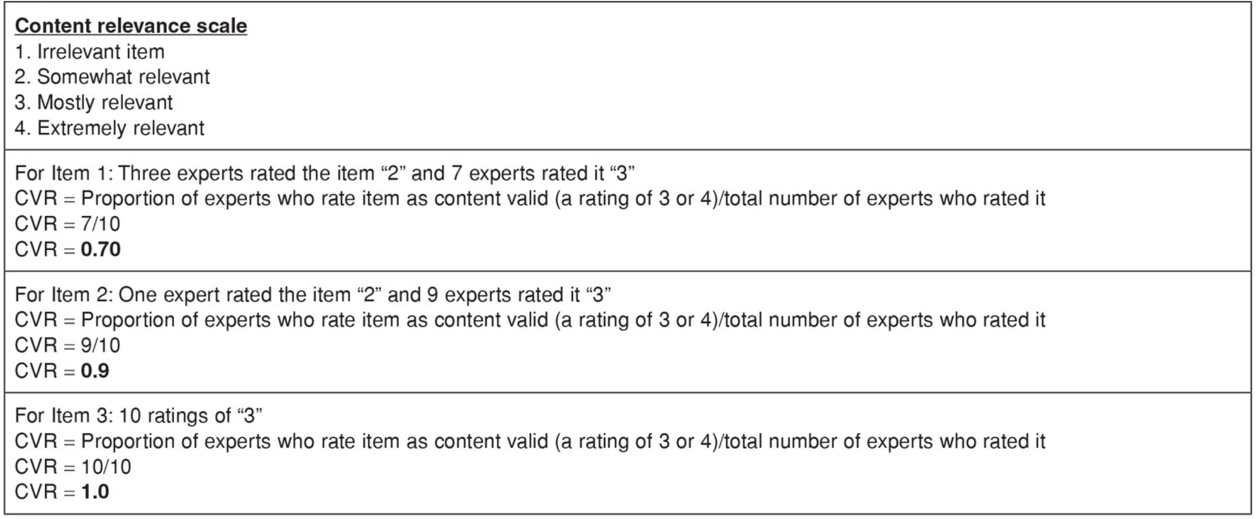
Validity is the degree to which an instrument measures what it is supposed to measure. Content validity (CV) determines the degree to which the items on the measurement instrument represent the entire content domain. Experts familiar with the content domain of the instrument evaluate and determine if the items are valid. A CV ratio (CVR) is a numeric value indicating the instrument’s degree of validity determined from expert’s ratings of CV. One rule of thumb suggests that a CVR of at least 0.78 is necessary to deem an item or scale as valid.

In order for a research study to provide accurate and meaningful results, the instrument used to test the hypothesis must be valid. Use of a measurement instrument that is not valid may produce meaningless results.

Methods to Calculate CVR

A CVR can be calculated for each item and overall for an instrument. There are two ways to calculate item and scale (overall) CVR. The first method was developed by Mary R. Lynn in 1986. Experts rate each item using a four-point ordinal scale (1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, and 4 = *highly relevant*). The item CVR is calculated as the number of experts giving a rating of 3 or 4 divided by the total number of experts who evaluated the item. The scale CVR is a proportion of items that met validity (i.e., at least 0.78) out of the total number of items. Figure 1 provides an example of how to calculate item CVR using this method. Figure 2 depicts how to calculate a scale CVR using this method.

Figure 1 Example of calculating item content validity ratio (CVR)



Content relevance scale
1. Irrelevant item
2. Somewhat relevant
3. Mostly relevant
4. Extremely relevant

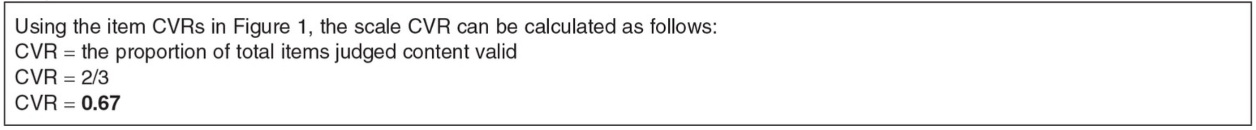
For Item 1: Three experts rated the item “2” and 7 experts rated it “3”
CVR = Proportion of experts who rate item as content valid (a rating of 3 or 4)/total number of experts who rated it
CVR = 7/10
CVR = **0.70**

For Item 2: One expert rated the item “2” and 9 experts rated it “3”
CVR = Proportion of experts who rate item as content valid (a rating of 3 or 4)/total number of experts who rated it
CVR = 9/10
CVR = **0.9**

For Item 3: 10 ratings of “3”
CVR = Proportion of experts who rate item as content valid (a rating of 3 or 4)/total number of experts who rated it
CVR = 10/10
CVR = **1.0**

Source: Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382–385.

Figure 2 Example of calculating scale content validity ratio (CVR)

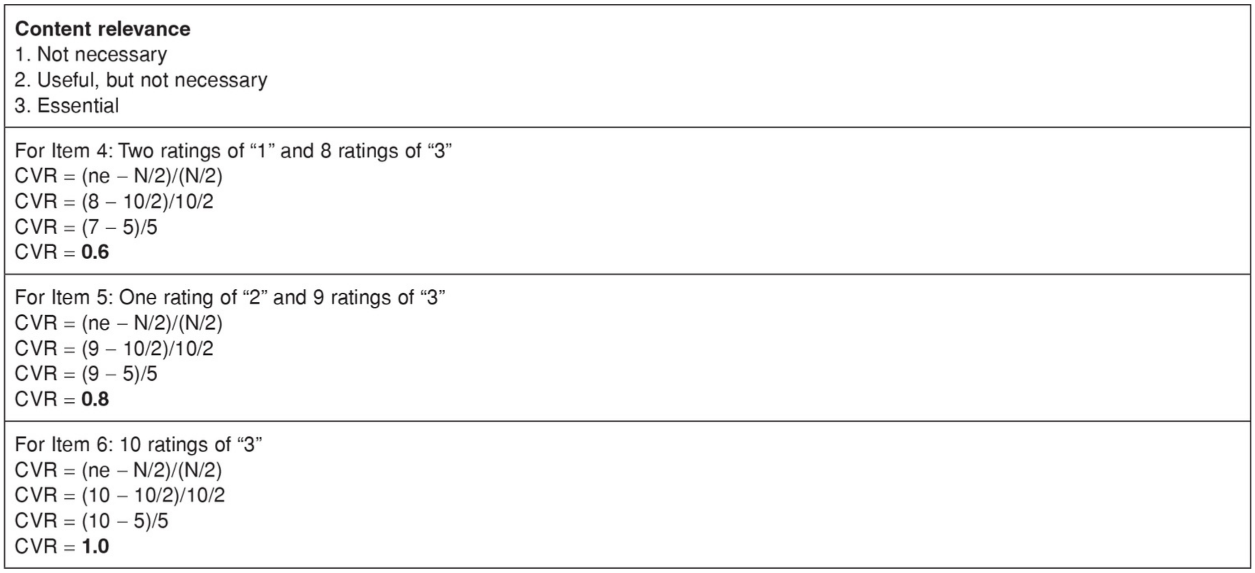


Using the item CVRs in Figure 1, the scale CVR can be calculated as follows:
CVR = the proportion of total items judged content valid
CVR = 2/3
CVR = **0.67**

Source: Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382–385.

A second method to calculate a CVR was developed by C. H. Lawshe in 1975. Experts rate each item using a four-point ordinal scale: 3 = *essential*; 2 = *useful, but not essential*; and 1 = *not necessary*. To calculate an item CVR, the following formula is used: CVR = $(n_e - N/2)/(N/2)$. In this ratio, n_e is the number of content experts who indicated that the item was essential (i.e., a rating of “3”). N is the total number of content experts. The mean CVR of all items computes an overall scale CVR. Figure 3 provides an example of how to calculate item CVR using this method, while Figure 4 demonstrates how to calculate a scale CVR.

Figure 3 Example of calculating item content validity ratio (CVR)



Content relevance
1. Not necessary
2. Useful, but not necessary
3. Essential

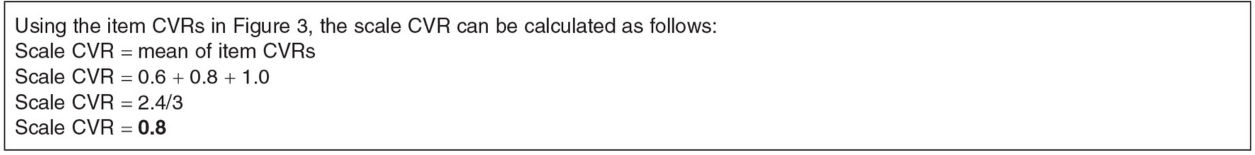
For Item 4: Two ratings of "1" and 8 ratings of "3"
CVR = (ne – N/2)/(N/2)
CVR = (8 – 10/2)/10/2
CVR = (7 – 5)/5
CVR = **0.6**

For Item 5: One rating of "2" and 9 ratings of "3"
CVR = (ne – N/2)/(N/2)
CVR = (9 – 10/2)/10/2
CVR = (9 – 5)/5
CVR = **0.8**

For Item 6: 10 ratings of "3"
CVR = (ne – N/2)/(N/2)
CVR = (10 – 10/2)/10/2
CVR = (10 – 5)/5
CVR = **1.0**

Source: Lawshe, C.H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x

Figure 4 Example of calculating scale content validity ratio (CVR)



Using the item CVRs in Figure 3, the scale CVR can be calculated as follows:
Scale CVR = mean of item CVRs
Scale CVR = 0.6 + 0.8 + 1.0
Scale CVR = 2.4/3
Scale CVR = **0.8**

Source: Lawshe, C.H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x

Additional CVR Procedures

According to Lynn, measurement instrument should be evaluated by at least six experts. These experts should be individuals who have published, presented, and/or are known nationally and internationally for their expertise in the content area. This ensures that the assessment of the validity tool is based on global practices and not standard local practices.

CV should be obtained from experts anonymously to avoid bias. Individuals who are familiar with the person requesting a review of the measurement tool are less likely to provide honest and valuable feedback.

If the CVR is less than 0.78 on an individual item, that item should be revised or deleted. Any feedback provided by experts should be considered in the revisions. If the overall CVR does not meet validity, revisions should be made and the instrument sent to at least six different experts for second review. This process is repeated until the scale CVR meets validity standards. Sending the instrument to different experts increases the rigor of the validating process; it also decreases bias from reviewers who have previously seen the instrument.

Tonya Rutherford-Hemming

See also Validity

Further Readings

Boulet, J. R., Jeffries, P. R., Hatala, R. A., Korndorffer, J. R., Feinstein, D. M., & Roche, J. P. (2011). Research regarding methods of assessing learning outcomes. Simulation in Healthcare, 6, S48–S51.

Davis, L. L. (1992). Instrument review: Getting the most from your panel of experts. Applied Nursing Research, 5, 194–197.

Lawshe, C.H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x

Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382–385.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what’s being reported? Critique and recommendations. Research in Nursing … Health, 29, 489–497.

Polit, D. F., & Beck, C.T. (2014). Essentials of nursing research: Appraising evidence for nursing practice (8th ed.). Philadelphia, PA: Wolters Kluwer.

Danielle N. Dupuis Danielle N. Dupuis Dupuis, Danielle N.

Content-Related Validity Evidence Content-related validity evidence

398

400

Content-Related Validity Evidence

Validation of test scores involves collecting evidence and developing an argument that supports a particular use (i.e., an inference or decision) of the test scores. For a validity argument to be correct, it must be supported by evidence and be logical and coherent. There are various types of evidence that can be used to support a validity argument, including content-related validity evidence, criterion-related validity evidence, and evidence related to reliability and dimensional structure. The type of evidence needed to support the use of the test scores depends on the type of inference or decision being made. As such, test scores can only be said to be valid for a particular use. If multiple inferences or decisions are to be made based on a set of test scores, multiple types of evidence may be required. Even if a single inference or decision is made, multiple types of evidence may still be required to support the test score use. After briefly reviewing the three types of validity evidence, this entry focuses on the basics of content-related validity evidence, including providing an example of its use.

Types of Validity Evidence

Validity evidence can be classified into three basic categories: content-related evidence, criterion-related evidence, and evidence related to reliability and dimensional structure. Most test score uses require some evidence from all three categories. Content-related validity evidence is evidence about the extent to which the test accurately represents the target domain. For achievement tests, the target domain is most often a particular subject matter domain (e.g., seventh-grade mathematics), and for ability tests, the target domain is most often a particular mental ability (e.g., quantitative reasoning). Criterion-related validity evidence is evidence that relates the test scores to one or more external criterion (often observable behaviors). Evidence related to reliability and dimensional structure are types of evidence about the internal structure of a test (i.e., the composition of the items and subtests). In addition, reliability evidence is evidence about the consistency or reproducibility of the test scores across various test conditions (e.g., raters and time).

Content-Related Validity Evidence

Content-related validity evidence is most important when making an inference about a target domain based on a sample of observations taken from that target domain. Evidence related to both the definition of the target domain and the representativeness or relevance of the sample of observations (items and tasks) taken from the target domain are important aspects of content-related validity evidence. Both types of evidence rely on the judgment of experts and are therefore subjective. Content-related validity evidence is often confused with, or is thought to be the same as, face validity evidence. This confusion is understandable because on the surface the two types of validity evidence have many commonalities. However, the two types of validity evidence differ in who is making the judgment about validity. When seeking evidence about face validity, the test takers and test users are asked if the test appears to measure what the test developers say the test measures. In contrast, when seeking evidence about content validity, individuals with expert knowledge in the target domain are asked if the test content (items and tasks) represents or is relevant to the target domain.

In the context of ability testing, content-related evidence is evidence about the relevance of the tasks or items to the latent trait or mental ability of interest. In the context of achievement testing, content-related evidence is evidence about the representativeness of the tasks or items to the subject matter domain being measured. Because content-related validity evidence relies on expert judgment and is therefore subjective, it is more appropriate for tests of specific knowledge and skills (i.e., achievement tests) than for tests of mental abilities or latent traits. Defining the target domain and sampling from that domain are easier when specific knowledge and skills are being measured. Defining the target domain and sampling from it are also easier when the knowledge and skills that comprise the domain are stable over time. Content-related validity evidence alone cannot support a particular use of test scores but is one part of the evidence that can be used to support a validity argument around a particular use of test scores.

Content-Related Validity Evidence Study

The goal of content validation work is to establish that the sample of observations (items and tasks) that comprise the test is representative of (or relevant to, if developing an ability test) the target domain. Collecting evidence related to content validity involves four basic steps: (1) identifying and selecting subject matter experts, (2) defining the target domain, (3) developing a procedure to sample observations (items and tasks) from the target domain, and (4) evaluating the effectiveness of the validation work.

When identifying and selecting subject matter experts, test developers should consider a variety of perspectives on the target domain to ensure it is thoroughly defined and described. In a content-related validity evidence study, the subject matter experts play a critical role and the success of the validation work depends, in part, on their ability to fully and accurately define and describe the target domain. However, subject matter experts should not be the sole source of information used to define and describe a target domain but should be viewed as one of many sources of information that can be used to define and describe a target domain.

Defining the target domain involves fully describing all of the knowledge and skills that comprise the target domain. When describing the knowledge and skills that comprise the target domain, it is important for test developers to be as specific as possible to enable item writers to design and create test items and/or tasks with some ease. Test developers should rely on the subject matter experts, in addition to other sources of information, to describe the knowledge and skills comprising the target domain. Additional sources of information might include previously developed tests and the research literature.

Sampling from the content domain should have a rationale, which itself should be documented. Random sampling is probably not feasible to ensure content coverage. Instead, systematic sampling that ensures the target domain is accurately represented should be considered.

Evaluating the effectiveness of the validation work involves assessing the extent to which the sampling procedure produced items and tasks that are representative of the target domain, such that test scores and the inferences and/or decisions made from those test scores are valid. This work should rely, in part, on the judgments of the subject matter experts. Procedures for quantifying the amount or representative of the content coverage have been developed and should be considered.

An Example: Measuring Seventh-Grade Mathematics Achievement

Test developers have been tasked with creating a measure of students' seventh-grade mathematics achievement. The test will be used to assess students' end-of-year knowledge and skills in seventh-grade mathematics and will measure teacher and/or school effectiveness.

The most obvious subject matter experts when assessing students' end-of-year knowledge and skills in seventh-grade mathematics are teachers of seventh-grade mathematics. Because all students (not requiring accommodations) will be taking this test, it is important to include teachers who teach students at all ability levels, including special education teachers and honors and gifted/talented teachers. Other subject matter experts might include school curriculum directors and mathematics education researchers.

Again, defining the target domain involves fully describing (and documenting) the target domain. In this example, the target domain is seventh-grade mathematics achievement, specifically, the knowledge and skills that comprise the seventh-grade mathematics standards and curriculum. It is often easier to start by identifying the key topics that comprise the target domain. For seventh-grade mathematics, this includes the concepts of ratio, proportion, and slope. Further, subtopics could also be identified, such as the different types of ratios—part–part ratios versus part–whole ratios. In addition to identifying the key topics or concepts (knowledge) that comprise the seventh-grade mathematics standards and curriculum, it may be important to also identify the key uses of those concepts (skills) such as the ability to identify or provide a definition of the concept or to solve problems using the concept.

Once the key concepts (topics) and the key uses of those concepts have been identified, a test blueprint can be developed, and items and tasks can be created that are representative of those key concepts and uses. Item and task development includes creating items and tasks that represent the intersection between each of the three key topics (i.e., ratio, proportion, and slope) and the two key uses (i.e., provide a definition and solve a problem) in the test blueprint, such sample items might include having students provide a definition of a proportion and solve a missing-value proportion problem.

After items and tasks have been developed and assembled into a measure of seventh-grade mathematics knowledge and skills, the extent to which the items and tasks are representative of seventh-grade mathematics knowledge and skills (i.e., the target domain) can be assessed. The subject matter experts should be asked to judge the adequacy of the content coverage. In addition, the adequacy of content coverage can be quantified using percentages. If a particular key topic or concept is found to not have adequate coverage, test developers can request items writers create additional items or tasks until all key topics of concepts have adequate coverage such that the target domain is represented.

Danielle N. Dupuis

See also Construct-Related Validity Evidence; Criterion-Based Validity Evidence; Reliability; Tests; Validity

Further Readings

American Educational Research Association, American Psychological Association, … National Council for Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: Author.

Downing, S. M., & Haladyna, T. M. (2006). Handbook of test development. Mahwah, NJ: Erlbaum.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. Journal of Business Research, 57, 98–107.

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment, 7, 238–247.

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563–575.

Contingency Table Analysis

Contingency table analysis

400

400

Contingency Table Analysis

See Two-Way Chi-Square

Michael T. Brannick Michael T. Brannick Brannick, Michael T.

Control Variables

Control variables

400

402

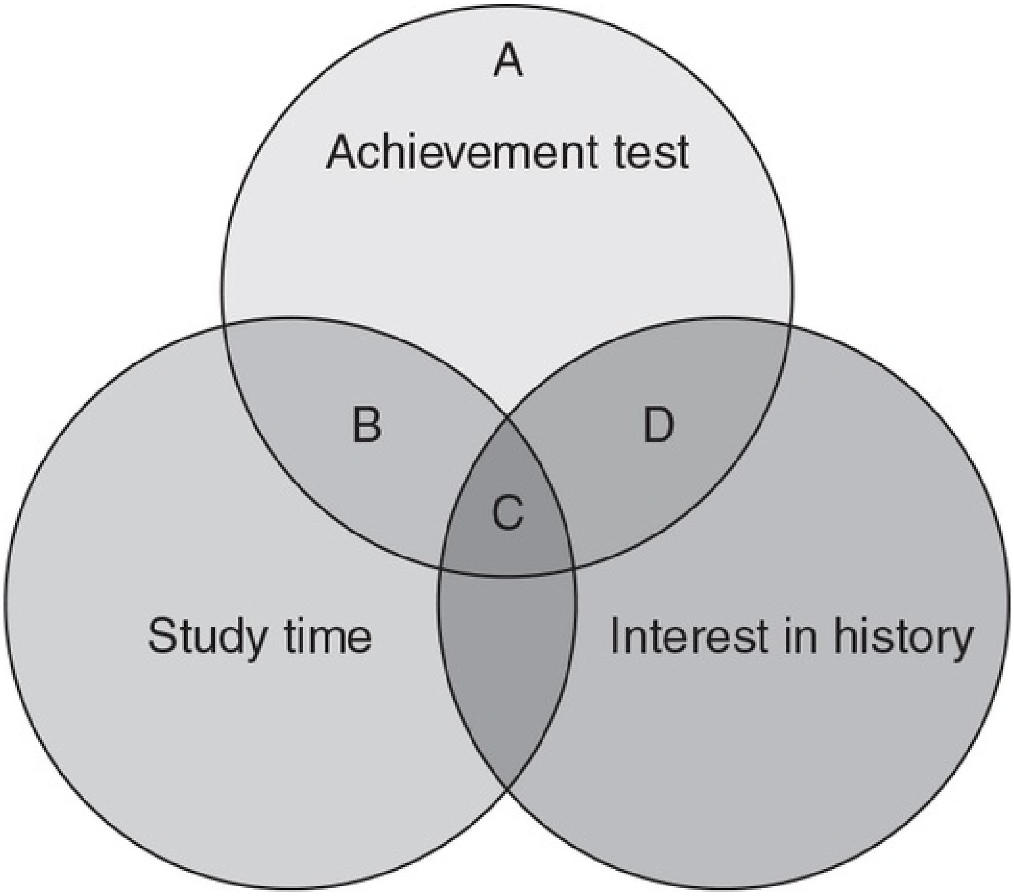
Control Variables

In correlational research, a control variable might be labeled a confounding variable or nuisance variable that is “held constant” by statistical means. Suppose we want to know the relations between length of study time and scores on a test of American history, but we are worried that interest in history might be an alternate explanation of the association. If we allowed students to choose and report their own study times for the test, and we also measured the participants’ interest in history, we could adjust the relations between study time and test score by statistically holding constant scores on interest in history. In such a study, interest in history would be described as a control variable.

Statistical Control

The mathematics of statistical control is based on correlation and regression, which can be illustrated graphically. In Figure 1, the variance of the distribution of American history test scores is partitioned into 4 areas labeled A, B, C, and D. Partition A is that part of the variance in test scores that is accounted for by neither study time nor interest in history—this is what cannot be predicted by either variable. Partition B is accounted for by study time alone. Partition C is shared by both study time and by interest in history—those more interested in history might spend more time studying, and thus either or both can account for this part of achievement. Finally, Partition D is the variance in achievement accounted for by interest in history alone.

Figure 1 Statistical control through removal of shared variance



The magnitude of association is indicated by the degree of overlap, that is, by the size of the shared portions. If study time and interest in history do an excellent job of predicting achievement, we would see Areas B, C, and D expand and Area A would shrink. However, if study time and interest in history were highly correlated, they would largely overlap one another, and the area marked C would increase, leading to smaller areas for B and D.

What statistical control does is remove the shared variance. In statistical terms, the *partial correlation* removes the control variable from both other variables of interest. In our example, we could compute a partial correlation between study time and achievement controlling for interest in history. The partial correlation would represent the ratio of B to (A + B), that is, the association of what is left of achievement with what is left of study time once interest in history is removed from both. The *semipartial correlation* removes the control variable from only one of the variables of interest. For example, we could compute the semipartial correlation between study time and achievement, holding constant interest in history for study time only. In this semipartial correlation, the association between study time and achievement would represent the ratio B to (A + B + C

+ D) because we would remove interest in history from study time, but not from achievement. The semipartial correlation is closely related to the regression coefficient. In essence, the multiple regression equation holds constant or controls each independent variable for all other independent variables.

Pros and Cons

The strength of control variables is that they allow the user to conclude that a focal variable accounts for variance in the dependent variable *above and beyond* the control variables in a regression equation. In our running example, if we computed a multiple regression and the association between time spent and achievement was significant, we could conclude that study time was related to achievement, even after holding interest constant. In other words, interest in history could be ruled out as an alternate explanation of our results. This would help to make a strong case for the value of study time.

On the other hand, suppose that time spent studying is highly correlated with interest in history because those most interested in history spent the most time studying. In such a case, the Area C in Figure 1 would become large, Area B would become small, and when we applied the regression model to the data, we would most likely find that time spent studying was not associated with achievement once interest in history was controlled. Should we infer that time spent studying is wasted? Clearly not; it could be that interest leads to studying, which leads to good test scores.

Applications

Control variables may not be advisable when theoretical understanding is the purpose of the research, and several articles have been written about the appropriate use of control variables. In organizational research, authors since 2000 suggest that researchers avoid including control variables in a regression equation simply because the controls are available for the analysis. Control variables should have a clear theoretical role in the analysis that is explained in the article’s introduction.

Beyond theoretical justification, there are measurement considerations for the inclusion of control variables. Control variables chosen for the analysis should be measured well and subject to the same standards of reliability and validity as the focal variables. One should avoid using variables that are proxies for the actual variables of interest (e.g., participant sex should be avoided as a proxy for interest in typically masculine or feminine interests).

When control variables are included in the analysis, they should be included in the summary table of descriptive statistics along with the focal variables. Results should be reported both including and excluding the control variables. When control variables are used, care is needed in making inferences because statistical control holds constant things that may be connected in ways not considered by the analysis, such as our example of interest leading to study time leading to achievement.

Michael T. Brannick

See also Causal Inference; Correlation; Descriptive Statistics; Experimental Designs; Multiple Linear Regression; Part Correlations; Partial Correlations

Further Readings

Becker, T. E., Atinc, G., Breaugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2015). Statistical control in correlational studies: 10 essential recommenations for organizational researchers. Journal of Organizational Behavior. doi:10.1002/job.2053

Howell, D. C. (2013). Statistical methods for psychology (8th ed.). Belmont, CA: Wadsworth.

Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), Minnesota studies in the philosophy of science (*Vol. 4*, pp. 372–402). Minneapolis: University of Minnesota Press.

Pedhazur, E. (1997). Multiple regression in behavioral research (3rd ed.). Fort Worth, TX: Harcourt Brace.

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. Organizational Research Methods, 14, 287–305. doi:10.1177/1094428110369842

Jackie Waterfield Jackie Waterfield Waterfield, Jackie

Convenience Sampling Convenience sampling

Convenience Sampling

Convenience sampling (also known as availability sampling) is a method where the selection of participants (or other units of analysis) is based on their ready availability. This availability is usually in terms of geographical proximity (e.g., students in the researcher’s own college or in neighboring colleges) but may involve other types of accessibility, such as known contacts.

As sample selection is based on the researcher’s choice, convenience sampling is a form of nonprobability sampling distinct from forms of probability sampling such as (stratified) random sampling or cluster sampling. Convenience sampling differs from quota sampling—another form of nonprobability sampling, in which selection is based on certain identified characteristics—in not specifically seeking representativeness.

Like other nonprobability sampling methods, convenience sampling has certain practical advantages. It does not require an exhaustive list of the study population, which is needed for random sampling, and has clear logistical and resource benefits in terms of travel, cost, and time expenditure. However, these advantages are at the price of certain biases, such as sampling error and undercoverage. Sampling error means that the sampling method provides a sample whose characteristics (e.g., participants’ age, educational level, or socioeconomic status) differ systematically from those of the population of interest. Undercoverage means that certain individuals in the population of interest are excluded by the sampling method (e.g., the researcher’s interest is in staff in community colleges, liberal arts colleges, and universities, but a convenience sample only accesses staff in community or liberal arts colleges).

If quantitative data are collected, a convenience sample’s lack of assured representativeness causes difficulties at the data analysis stage. As the sample is representativeness causes difficulties at the data analysis stage. As the sample is not representative in the way that a probability sample is, using a sample statistic (e.g., a sample proportion) to estimate a population parameter (e.g., a population proportion) is inadvisable, as such an estimate is likely to be biased. Furthermore, using statistical hypothesis tests is questionable, as these assume random sampling. Inferential statistics applied to convenience samples therefore make an assumption that the sample is comparable to a random sample from the same population (an assumption that is normally untestable). In qualitative research, however, this strict empirical representativeness is not normally at issue. What matters here is that members of the sample are relevant to the aims of the study—this is more a notion of theoretical than of statistical generalization and does not require the same concern for empirical representativeness.

> Although convenience sampling has methodological shortcomings, these can be mitigated by:
> describing the demographic and other characteristics of the sample in detail, and if possible, comparing these with those of the relevant population, so that readers of the study can evaluate its representativeness;
> making efforts to gain the participation of all intended participants, so that response bias or
> self-selection does not compound a lack of representativeness; and
> ensuring that the participants recruited are theoretically relevant to the study, so that selection is not based *solely* on convenience.

Jackie Waterfield

See also External Validity; Quota Sampling; Selection Bias; Simple Random Sampling

Further Readings

Fink, A. (2013). How to conduct surveys: A step-by-step guide (5th ed.). Thousand Oaks, CA: Sage.

Lohr, S. L. (2008). Coverage and sampling. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), International handbook of survey methodology (pp. 97–112). New York, NY: Erlbaum.

Gregory R. Hancock Gregory R. Hancock Hancock, Gregory R.

Convergence

Convergence

Convergence

Convergence is a process in statistical analysis describing a series of calculations or guesses for the purpose of ultimately producing a very precise estimate. A simple example of several related equations illustrates the process. Imagine being tasked with solving the following set of equations for *X* and *Y*:

$$X + Y = 5, X - Y = 3, 2X + 3Y = 12.$$

Given only Equations 1 and 2, the task easily yields *X* = 4 and *Y* = 1. However, these values do not satisfy Equation 3, for which substituting *X* = 4 and *Y* = 1 would yield 11 rather than 12. In fact, there is no exact solution for *X* and *Y* that satisfies all three equations. Still, one may wish to know what estimates of *X* and *Y*, and Ỹ, respectively, make all three equations as close to true as possible.

What “close” means in one equation is clear; for Equation 1, for example, should yield a value close to 5. For more general purposes, however, close must be operationalized across the set of equations by specifying a function that yields a single numerical value operationalizing the discrepancy between the equations’ outcome values (5, 3, 12) and the outcome values expected based on the estimates and Ỹ. Such a *discrepancy function*, or *fit function*, may then be used to guide the derivation of optimal values for those estimates.

A simple discrepancy function example, representing an unweighted least squares criterion, would be:

$$F = [5 - (\tilde{X} + \tilde{Y})]^2 + [3 - (\tilde{X} - \tilde{Y})]^2 + [12 - (2\tilde{X} + 3\tilde{Y})]^2.$$

Using this function, we seek and $\tilde{Y}$ values minimizing *F*. Readers familiar with multivariable calculus recognize that this could be accomplished analytically, setting to zero the partial derivatives of *F* with respect to and $\tilde{Y}$ and solving. Computers, however, are less adept at analytical solutions; fortunately, they are good at using algorithms that employ iterative strategies to derive estimates for unknown quantities.

After choosing initial *start values* for and $\tilde{Y}$, a computer changes those estimates incrementally, moving in those directions that make *F* smaller. The process continues adaptively through several *iterations*, altering the estimates in typically smaller increments until *F* reaches *convergence*. That is, the algorithm stops when *F* can no longer meaningfully decrease given its incremental changes in and $\tilde{Y}$, ideally reaching a value close to the analytical minimum; the resulting empirical values of and $\tilde{Y}$ constitute the estimates according to the criterion used to define *F*.

Although this is just a simple example, it represents a process that occurs throughout much of statistics. *Maximum likelihood estimation*, for example, employed across many applications (e.g., logistic regression, item response theory, structural equation modeling), seeks estimates for model parameters that optimize a discrepancy function characterizing the likelihood of all observations within a sample of data.

Statistical methods employ many different types of discrepancy functions as well as different search algorithms to optimize them. An algorithm might also fail to converge upon a solution, a result more common with complex models and models with very poor fit to the data. Alternatively, an algorithm might converge but reach a *local minimum* in the discrepancy function rather than the global minimum. This underscores the importance of choosing multiple sets of start values to ensure convergence occurs and that it is to a globally optimum solution for the parameter estimates of interest to the researcher.

Gregory R. Hancock

See also Item Response Theory; Logistic Regression; Maximum Likelihood Estimation; Model–Data Fit; Structural Equation Modeling

Further Readings

Argyros, I. K. (2008). Convergence and applications of Newton-type iterations.

New York, NY: Springer.

Eliason, S. R. (1993). Maximum likelihood estimation: Logic and practice. Newbury Park, CA: Sage.

Lori Kupczynski Lori Kupczynski Kupczynski, Lori

Cooperative Learning Cooperative learning

404

408

Cooperative Learning

Cooperative learning, sometimes called small group learning, is a teaching strategy that utilizes and emphasizes small learner groups as a core unit that must work together and complete tasks collectively in order to achieve the desired academic goals while providing the learner members with both academic and social learning experiences. Groups typically consist of two to four members. Each learner in a group is responsible for his or her own learning as well as the learning of his or her group mates, creating an atmosphere of cooperative achievement.

Cooperative learning has been described as structuring positive interdependence. Cooperative learning can be conceptualized as placing learners on a team in which the goal is for all members to achieve academic success. This entry explains the differences between cooperative learning and individual learning, then discusses the elements of successfully incorporating cooperative learning in the classroom, the types and techniques of cooperative learning, and its benefits and disadvantages.

Cooperative Learning Versus Individual Learning

Cooperative learning can be understood as contrasting with individual learning teaching approaches. The key difference between cooperative learning and individual approaches to learning focuses on how the learners' learning goals are structured. Learner learning goals specify that how learners are expected to interact with each other and with the instructor. Learners in a cooperative learning environment are able to capitalize on each other's resources and skills by sharing information, evaluating ideas, monitoring work, and checking answers for group members. In cooperative learning, the instructor’s role changes from giving information to learners to help them learn to that of a learning facilitator.

In cooperative learning, learners are encouraged and expected to focus on outcomes that are not only beneficial to themselves but also to the other members of the group. In individual learning learners work against each other and toward attainment of an academic goal such as receiving an “A” in the course. Additionally, in individual learning, learners work individually by themselves to accomplish the set learning goals, whereas in cooperative learning, learners work together to improve their chances of success and the success of their teammates.

Elements of Successful Incorporation of Cooperative Learning

Five essential elements have been identified for the successful incorporation of cooperative learning in the classroom: positive interdependence; individual and group accountability; promotive interaction; interpersonal and small group skills; and group processing.

Positive interdependence

occurs when members of a group identify a link to each other that connects the success of one with the success of all others. Group members recognize, then, that the efforts of each member are beneficial to everyone and this creates a commitment to all members of the group. In order for positive interdependence to be accomplished:

- Each learner must completely participate within the group.
- Each group member must have a task, role, or responsibility.
- Each group member must understand that each person is responsible for his or her own learning as well as that of the group.

Individual and group accountability

requires that the group be certain about goals and able to determine not only the progress toward achievement but also the distinct efforts of each individual member. This includes accountability for each member to contribute his or her share of the work required. In order to have this type of accountability:

Each learner must achieve and demonstrate content mastery.

Each learner must be accountable for his or her own work and learning.

Each learner must actively contribute and not expect the group to “carry” them.

Promotive interaction

exists when all members of the group share resources and focus on providing assistance and encouragement for each member’s endeavors to learn. Examples of promotive interaction may include oral explanation of how to solve a problem, explanations of how to solve a problem, or peer teaching to classmates. To achieve promotive interactions:

Face to face interaction is necessary.

Members must foster the success of other group members.

Learners must be able to explain to one another what they have learned or are learning.

Learners must be able to assist one another with understanding and completing assignments.

Interpersonal and small group skills

are the skills necessary to perform successfully as part of a group. In cooperative learning groups, students are expected to understand these skills (teamwork) along with acquiring academic content (taskwork). These skills should be seen as a goal and result of cooperative learning and include effective communication skills, interpersonal skills, and group skills. Examples of skills gained through cooperative learning include:

- leadership,
- decision making,
- trust building,
- friendship development, and
- communication
- Conflict management.

Group processing

Group processing

occurs when group members discuss how well they are achieving their goals. Groups need to determine what member actions are beneficial and what actions are counterproductive and decide what behaviors to continue or change. The focus of group processing is to elucidate and advance the group’s efficacy and the effectiveness of individual members.

Ultimately, in order for cooperative learning to be successful, two main characteristics must be present: 1. Tasks and reward structures, responsibility of each individual, and accountability of each participant must be clearly distinguished at the beginning and presented to all group members. Individual group members must know exactly what their responsibilities are. There must be individual accountability within the group for each group member’s responsibilities.

2. Each group member’s responsibility must be such that it can only be completed by that individual member, and as such, each member must actively participate to ensure success for the whole group.

Cooperative Learning Types

Formal Cooperative Learning

Formal cooperative learning is designed, facilitated, and overseen by the instructor over time. Formal cooperative learning is composed of learners working cooperatively to accomplish shared learning goals and to complete defined assignments and tasks. This may occur for as little as one class period or continue for as long as the entire term of the class. In this type of cooperative learning, groups are set for a specific duration and learners contribute to each other’s knowledge on a continuous basis. This type of cooperative learning is usually more appropriate for larger groups of learners.

The instructor’s role in formal cooperative learning includes various elements such as preinstructional decisions regarding tasks, objectives, roles, materials, groups, and room assignments; explanations of the tasks and the cooperative learning structure for the students; monitoring the learning and intervening as needed to ensure the task remains on track; and assessing the learning of the individuals and group as it functions.

Informal Cooperative Learning

Informal cooperative learning consists of having learners work together to achieve a joint learning goal in groups formed to last for a brief period, from a few minutes to a class session. In informal cooperative learning, the learners participate in focused discussions before and after the lesson while also participating in pair discussions throughout the lesson. Informal cooperative learning focuses on active involvement of learners in understanding what is being presented as each learner has an individual responsibility to be an active participant in the paired discussions. Informal cooperative learning sets a mood supportive of learning, assists in creating expectations, and provides a closing to an instructional gathering.

During informal cooperative learning, instructors are afforded more time as well as more flexibility to move around the classroom and focus on the learning that is occurring. This can offer instructors additional awareness of how well learners are grasping the concepts presented. When using informal cooperative learning groups and techniques, it is important to ensure that the assignment and the directions are clear and precise and to require groups to produce a specific tangible outcome, such as a written answer. Informal cooperative learning is made up of three types of discussions: 1. Introductory-focused discussion: Instructors assign learners to pairs or triads and explain the task.

2. Intermittent-focused discussions: Instructors divide the lecture into segments. After each segment, learners are directed to partner up with a peer near them and work cooperatively to answer a question. Each learner is expected to first formulate an individual answer. Then, the two learners share and synthesize ideas into what most likely is a more accurate answer.

3. Closure-focused discussion: Learners participate in a final discussion task in which they summarize what they have learned with a focus on integrating the information into the existing conceptual frameworks.

Cooperative Base Groups

Cooperative base groups are continuous, diverse cooperative learning groups with consistent involvement by members. Characteristically, these groups meet regularly and last for the duration of the class. A good example is a course-long study group. Base groups can participate in and undertake many tasks, including academic tasks, such as reviewing each other’s work; personal tasks, which could be assisting each other with resolution of nonacademic issues; routine tasks, such as preparing the classroom for a lesson; and assessment tasks, such as checking each other’s perception of the material presented. Members’ primary responsibilities include ensuring that all members are progressing academically, maintaining accountability, and providing support, assistance, and encouragement for assigned work.

Cooperative Learning Techniques

There are many varied cooperative learning techniques available. Some cooperative learning techniques utilize learner pairs and other techniques operate with small groups consisting of four or five learners. Cooperative learning techniques have been designed and adapted for any content area making this a versatile teaching approach in both formal and informal cooperative learning settings. Some examples are as follows:

Think–Pair–Share

Think–pair–share requires learners to reflect individually on a presented question or problem. The learner may write down thoughts or simply brainstorm; however, the learner is expected to have developed thoughts or ideas on the posed subject. When prompted, each learner pairs up with a peer to discuss his or her idea(s) and listen to the ideas of his or her partner. In this way, learners are challenged to evaluate their own ideas and that of their peer. Following the pair discussion, the teacher may solicit replies from everybody.

When a teacher uses this method, it alleviates the worry about learners not volunteering because each learner will have already been required to contemplate and discuss at least two ideas (the learner’s and learner’s partner’s) allowing for dialogue and discussion expansion. This technique is particularly well suited for informal cooperative learning as it does not benefit from having set groups.

Jigsaw

The jigsaw technique requires each learner to become not only a learner but also a teacher. Learners are participants of two groups, a primary “home” group and a secondary “expert” group. One way to set this up is to assign each member of the home group a number that corresponds to each member of the other home groups such that each home group has the same numbers (e.g., 1–5). In this way, all of the number 1s will break away and become an expert group (all 1s). Learners then study the assigned material together with the other members of the “expert” group so that all members of the group have learned the assigned topic. The learners then return to their home groups and each learner is responsible for teaching the topic on which the learner became an “expert” while learning about the topic with learner’s homogenous group. This cooperative learning technique is most useful in formal cooperative learning groups.

Inside–Outside Circle

This particular cooperative learning strategy is more likely to be used in informal cooperative learning. The physical setup requires learners to form two concentric circles one inside of the other. As the activity takes place, learners in one circle will rotate and face new partners with each rotation which may coincide with each new question or topic. This method is useful when generating new ideas and solving problems.

Reciprocal Teaching

This technique allows for learner pairs to create and hold a dialogue about a text. In this technique, partners take turns reading the text and asking each other questions about the text. This technique allows learners to receive immediate feedback from their partner. Reciprocal teaching allows for learners to use and practice important collaborative learning skills such as clarifying, questioning, predicting, and summarizing. This technique is well suited for informal collaborative learning groups.

Benefits of Cooperative Learning

The benefits of cooperative learning are extensive. This form of learning allows students to develop higher order thinking skills, increase self-esteem, raise satisfaction with the learning environment, foster a positive attitude toward the content, develop more advanced communication skills, and increase social content, develop more advanced communication skills, and increase social interaction. It also increases student retention and enhances self-management skills.

Cooperative learning strategies contribute to an environment of exploratory and active learning where students are able to clarify concepts and ideas through discussion and debate. Students learn to critique ideas rather than people, learn interpersonal relationship skills, and meet high expectations. As well, this technique allows for task-oriented instruction that is less disruptive and that allows for differentiated instruction to reach students with various learning styles, reduces classroom and test anxiety, and mirrors real-life social and business situations for students to better prepare them for life beyond the classroom.

Disadvantages of Cooperative Learning

Cooperative learning does have its disadvantages. In cooperative learning environments, it is common for low-achieving students to become passive and not focus on the task at hand. There is an increased chance of conflict and an increased need for conflict resolution. Groups may get off task and begin discussing irrelevant information. Higher ability students may not be challenged, whereas lower ability students may always find themselves in need of help and may never experience leadership or “expert” status. There is the risk that one student will take over the group; conversely, there is the possibility of someone not actively contributing and expecting the group to “carry” them.

Lori Kupczynski

See also College Success; Curriculum; Learning Styles; Learning Theories

Further Readings

Aldrich, H., & Shimazoe, J. (2010). Group work can be gratifying: Understanding and overcoming resistance to cooperative learning. College Teaching, 58(2), 52–57.

Baker, T., & Clark, J. (2010). Cooperative learning—a double edged sword: A cooperative learning model for use with diverse student groups. Intercultural Education, 21(3), 257–268.

Johnson, D. W., & Johnson, R. (1989). Cooperation and competition: Theory and research. Edina, MN: Interaction Book.

Johnson, D. W., & Johnson R. T. (2008). Social interdependence theory and cooperative learning: The teacher's role. In R. B. Gillies, A. F. Ashman, & J. Terwel (Eds.), The teacher's role in implementing cooperative learning in the classroom (pp. 9–37). New York, NY: Springer.

Johnson, D. W., Johnson, R. T., & Holubec, E. (2008). Cooperation in the classroom (8th ed.). Edina, MN: Interaction Book.

Tsay, M., & Brady, M. (2010, June). A case study of cooperative learning and communication pedagogy: Does working in teams make a difference? Journal of the Scholarship of Teaching and Learning, 10(2), 78–89.

Elizabeth T. Gershoff Elizabeth T. Gershoff Gershoff, Elizabeth T.

Corporal Punishment

Corporal punishment

408

412

Corporal Punishment

Corporal punishment is the use of physical force, no matter how light, with the intention of causing a child to experience bodily pain so as to correct or punish the child’s behavior. Corporal punishment remains a commonly used practice by parents around the world, but it is also used by teachers throughout the world as a means of punishing children for their misbehaviors. This entry first describes corporal punishment in schools and looks at its legality and prevalence. It then discusses disparities in the use of corporal punishment, the outcomes for children who are subjected to corporal punishment, and efforts to reduce corporal punishment.

School corporal punishment often includes hitting children with an object, such as a wooden paddle, stick, or whip, but also takes the form of slapping, pinching, hair pulling, and ear pulling. Corporal punishment does not refer only to hitting children as a form of discipline; it also includes other practices that involve purposefully causing the child to experience pain in order to punish the child, including washing a child’s mouth out with soap, forcing a child to stand in a painful position for long periods of time, making a child kneel on sharp or painful objects (e.g., rice, a floor grate), placing hot sauce on a child’s tongue, and forcing a child to engage in excessive exercise or physical exertion. The term *corporal punishment* is synonymous with the term *physical punishment*.

Legality

As of 2016, corporal punishment in schools was prohibited in 128 countries. School corporal punishment is banned from all of Europe and most of South

America and East Asia; it is permitted in most countries in Africa and Southeast Asia and in the United States. In the United States, corporal punishment in public schools is legal in 19 states, while corporal punishment in private schools is legal in 48 states (the exceptions are Iowa and New Jersey). Australia, South Korea, and the United States are the only industrialized countries that allow school corporal punishment. In many of the countries that allow school corporal punishment, corporal punishment has been banned from prisons and from the armed services, leaving schools as the last public institutions where corporal punishment is legal, and children the last group of people it is legal to hit.

There is a growing international consensus that corporal punishment of children, whether by teachers or parents, is a violation of children's human rights under the United Nations Convention on the Rights of the Child. The United Nations Committee on the Rights of the Child has concluded that all corporal punishment of children violates children's right to protection from physical and mental violence (per Article 19 of the Convention on the Rights of the Child) and should be banned. As a result, a total of 49 countries have banned all corporal punishment of children, including that by parents.

Prevalence

The United Nations Children's Fund and other organizations have documented that, in some countries, nearly all children (upward of 80% of students) are subject to school corporal punishment on a regular basis; this is true, for example, in Egypt, India, Jamaica, Myanmar, Uganda, and Yemen, among many other countries. Interviews with children and teachers have revealed that school corporal punishment continues even in countries where it is banned, such as in Cameroon, Kazakhstan, Kenya, and South Africa. In the United States, the U.S. Department of Education reported that more than 110,000 children were subject to school corporal punishment in the 2013–2014 school year.

Disparities in Use of Corporal Punishment

Around the world, school corporal punishment is not used equally across all groups of children. Boys, children from racial and ethnic minorities, and children with disabilities are more likely to experience corporal punishment than their peers. In United Nations Children's Fund's Young Lives study, boys were more likely than girls to experience school corporal punishment in each of the four countries studied: Ethiopia, 44% of boys versus 31% of girls; India, 93%

countries studied: Ethiopia: 44% of boys versus 31% of girls, India: 83% versus 73%, Peru: 35% versus 26%, and Vietnam: 28% versus 11%. Indeed, in both Singapore and Zimbabwe, gender discrimination is written into law: Only boys can be subject to school corporal punishment in those countries.

Disparities in school corporal punishment by gender, race, and disability status have been documented in the United States. An analysis of data from all 95,088 public schools in the United States revealed that Black children were much more likely than White children and that children with disabilities were more likely than children without disabilities to be corporally punished in school. The most systematic disparities were for gender, as nearly every school district reported corporally punishing boys at a rate of 3 times that for girls and often times at a rate of 5 times that for girls. These disparities are in contravention of several U.S. federal laws that protect schoolchildren from discrimination on the basis of race, gender, and disability status.

Disparities in the use of school corporal punishment are concerning both because they are unfair and potentially illegal, but also because students who perceive they are being treated in a discriminatory fashion are more likely to engage in negative school behaviors, to have low academic achievement, and to have mental health problems.

Outcomes for Children

Although there is an extensive research literature on the child outcomes linked with parents’ use of corporal punishment, whether and how school corporal punishment affects children has not been extensively studied. The studies that do exist have occurred outside the United States.

Some educators may use corporal punishment in an effort to improve children’s academic performance and achievement, sometimes indirectly by reducing problem behavior. Yet there is no evidence that school corporal punishment promotes learning and in fact some evidence that it is a hindrance. Research studies conducted in Jamaica and Nigeria each found that children who receive corporal punishment score lower on literacy skills, math skills, executive functioning, and intrinsic motivation.

The strongest evidence of links between school corporal punishment and children’s achievement comes from United Nations Children’s Fund’s Young Lives study of children in four developing countries noted earlier. The study

Lives study of children in four developing countries noted earlier. The study surveyed children in 2011 and again in 2013 and was able to link corporal punishment at age 8 to school performance at age 12, thus eliminating the possibility that the association is a result of children, with low scores eliciting corporal punishment as children’s later school performance cannot predict their corporal punishment earlier in time. Children from each country reported high rates of school corporal punishment (from 20% to 80% of children) when they were 8 years of age, and the more corporal punishment they received at age 8, the lower their math scores were in two samples (Peru and Vietnam) and the lower their vocabulary scores in Peru. Importantly, in none of the countries did school corporal punishment at age 8 predict better school performance at age 12.

One reason that corporal punishment may interfere with children’s learning is that children avoid or dislike school because it is a place where they are in constant fear of being physically harmed by their teachers. In the Young Lives study, 5% of students in Peru, 7% in Vietnam, 9% in Ethiopia, and 25% in India who reported a reason for not liking school listed being beaten by teachers as their most important reason. Studies in a variety of countries have revealed that students are afraid of corporal punishment in school and skip days of school or drop out of school altogether to avoid being beaten by teachers.

Students who are corporally punished in school are also more likely to suffer from mental health and behavioral problems. Children who are corporally punished are more likely to have depressive symptoms, to be hostile, and to be aggressive. In the Young Lives study, school corporal punishment at age 8 predicted less self-efficacy 4 years later in Ethiopia and Peru and lower self-esteem 4 years later in Ethiopia and Vietnam.

An important concern with the use of corporal punishment is that it may cause serious injury to children, in large part because objects are so often used to hit children in schools. Some injuries are relatively minor, such as bruises, bumps, and cuts, but others are more major, including broken bones, hematomas, nerve damage, and in rare cases, death. Injuries from school corporal punishment are not restricted to developing countries; in the United States, court cases have documented these same physical injuries from school corporal punishment, including cases of death by excessive exercise as punishment.

Efforts to Reduce Corporal Punishment

Advocates around the world have called for school corporal punishment to be banned because of the research indicating it is ineffective and potentially harmful to children and the fact that it is considered a violation of children’s human rights. In the United States, prominent professional organizations such as the American Academy of Pediatrics, the American Bar Association, the American Civil Liberties Union, the American Medical Association, the American Psychological Association, the National Association of Elementary School Principals, and Prevent Child Abuse America have publicly called for school corporal punishment to be abolished in the United States.

If concern for children’s welfare is not enough of an incentive, yet another aspect that may motivate countries to consider bans is the costs associated with corporal punishment. In a report prepared for the nongovernmental organization Plan International, researchers calculated that countries lose millions and sometimes billions of dollars each year as a result of various forms of school violence, including corporal punishment. These costs include the long-term costs associated with lower achievement and higher dropout rates, such as lower earnings, higher physical and mental health needs, and higher reliance on social services.

National bans on school corporal punishment are an important step toward reducing the practice, but as noted earlier, corporal punishment continues even in countries where it is illegal, in large part because teachers, parents, and children are often convinced that corporal punishment is necessary for disciplining children. Eliminating corporal punishment will require interventions that teach both adults and children about the harms associated with corporal punishment and about more effective and nonviolent methods of discipline.

There are examples from around the world of successful educational campaigns to reduce school corporal punishment. The Council of Europe has an ongoing campaign called “raise your hand against smacking” (*smacking* is a term used for spanking in Europe) that is focused on changing public attitudes about corporal punishment. Similarly, a campaign by Plan International known as Learn Without Fear has trained teachers throughout the world in nonviolent discipline and has advocated for bans on corporal punishment in a number of countries. In Uganda, an intervention called the Good Schools Toolkit has been used successfully to train teachers in positive disciplinary methods and thereby reduce the incidence of corporal punishment by 42%.

Changing attitudes about corporal punishment and providing teachers with

Changing attitudes about corporal punishment and providing teachers with methods with which they can replace corporal punishment are necessary steps in eliminating school corporal punishment. Schoolwide interventions such as social–emotional learning and positive behavioral interventions and supports are considered effective at reducing students' problem behaviors and creating a positive learning environment for students. Such interventions work to improve student behavior at the school level, thereby obviating the need for school corporal punishment in the first place.

Elizabeth T. Gershoff

See also Adultism; Childhood; Compliance; Cross-Cultural Research; Educational Psychology; Parenting Styles; Punishment; Social Justice; U.S. Department of Education

Further Readings

Committee on the Rights of the Child. (2006). General Comment No. 8 (2006): The right of the child to protection from corporal punishment and or cruel or degrading forms of punishment (articles 1, 28(2), and 37, inter alia) (CRC/C/GC/8). Geneva, Switzerland: United Nations.

DeVries, K. M., Knight, L., Child, J. C., Mirembe, A., Nakuti, J., Jones, R., & Naker, D. (2015). The Good School Toolkit for reducing physical violence from school staff to primary school students: A cluster-randomised controlled trial in Uganda. Lancet Global Health, 385, e378–e386.

Gershoff, E. T. (2013). Spanking and child development: We know enough now to stop hitting our children. Child Development Perspectives, 7, 133–137. doi:10.1111/cdep.12038

Gershoff, E. T., Purtell, K. M., & Holas, I. (2015). Corporal punishment in U.S. public schools: *Legal* precedents, current practices, and future policy. Springer Briefs in Psychology Series, Advances in Child and Family Policy and Practice Subseries, 1, 1–105. doi:10.1007/978-3-319-14818-2

Global Initiative to End All Corporal Punishment of Children. (2015). Towards nonviolent schools: Prohibiting all corporal punishment. Global report 2015. Retrieved from http://www.endcorporalpunishment.org/resources/thematic-reports/schools-report-2015.html

King, J. B. (2016, November 22). Letter to states calling for an end to corporal punishment in schools. Washington, DC: U.S. Department of Education. Retrieved from https://www.ed.gov/category/keyword/corporalpunishment

Ogando Portela, M. J., & Pells, K. (2015). Corporal punishment in schools: Longitudinal evidence from Ethiopia, India, Peru, and Viet Nam (Innocenti Discussion Paper No. 2015-02). Florence, Italy: UNICEF Office of Research. Retrieved from https://www.unicef-irc.org/publications/series/22/

Pereznieto, P., Harper, C., Clench, B., & Coarasa, J. (2010). The economic impact of school violence. London, UK: Plan International … Overseas Development Institute. Retrieved from plan - international.org/learnwithoutfear

Hyun Joo Jung Hyun Joo Jung Jung, Hyun Joo

Jennifer Randall Jennifer Randall Randall, Jennifer

Correlation

Correlation

412

413

Correlation

If one wants to know the degree of a relationship, the correlation between two variables can be examined. Correlations can be quantified by computing a *correlation coefficient*. This entry first describes a concept central to correlation, *covariance*, and then discusses calculation and interpretation of correlation coefficients.

Covariance indicates the tendency in the linear relationship for two random variables to covary (or vary together) that is represented in deviations measured in the unstandardized units in which *X* and *Y* are measured. Specifically, it is defined as the expected product of the deviations of each of two random variables from its expected values or means.

The population covariance between two variables, *X* and *Y*, can be written by:

$$\begin{aligned} cov(X,Y) &= E[(X-E(X))(Y-E(Y))] \\ &= E[XY - XE(Y) - E(X)Y - E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) - E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

where *E* is the expected value or population mean.

Similarly, the sample covariance between *x* and *y* is given by:

$$s(x,y) = \frac{1}{N-1}\sum_{i=1}^{n}(x)_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{N}\right]$$

where *N* is the number of observations; are the sample means of *x* and *y*.

When one interprets covariances, zero covariances indicate that variables are not linearly related. If they are nonlinearly associated or statistically independent, the covariance is zero. On the other hand, a nonzero covariance indicates the tendency of covarying. If the sign of covariance is positive, the two variables tend to vary in the same direction. If a covariance value is negative, the two variables tend to move in the opposite direction. The covariance is not independent of the unit used to measure *x* and *y*, and so magnitude of the covariance depends on the measurement units of two variables. Note that the nonzero covariance does not indicate causation and how strong the association is between two variables.

When considering a variance–covariance matrix, covariances or correlations among observed variables depend on the relationship between latent variables and linear composite variables (i.e., tests consisting of more than one item). The covariance between two composite variables is the sum of the elements of the covariance matrix. It can be written by:

$$\sigma(X,Y) = \sigma(X,Y) = \sum_{i=1}^{p}\ \sum_{j=1}^{q}\sigma(X_i, Y_j)$$

where *p, q* are the numbers of variable in *X* and *Y*, respectively.

This most commonly computed correlation coefficient is a standardized index of linear association. From the covariance, the correlation coefficient for *X* and *Y* is calculated using the following equation:

$$r = \frac{\sum_{i=1}^{N} \frac{x_i\, y_i}{N-1}}{s(x,y)} = \frac{\sum_{i=1}^{N} x_i\, y_i}{(s_X s_Y)}$$

The *product moment correlation coefficient* was originally invented by Karl Pearson in 1895 based on the studies conducted by Francis Galton and J. D. Hamilton Dickson in the 1880s. The correlation coefficient ranges from −1 to 1. If its value is 0, the variables have no linear relationship; and if its value is −1 or 1, each variable is perfectly predicted by the other. Its sign indicates the direction of the relationship.

Hyun Joo Jung and Jennifer Randall

See also Autocorrelation; Pearson Correlation Coefficient; Phi Correlation Coefficient

Further Readings

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. Mahwah, NJ: Erlbaum.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London, 45, 135–145.

Galton, F. (1889). Natural inheritance. London, UK: Macmillan.

Galton, F., & Dickson, J. H. (1886). Family likeness in stature. Proceedings of the Royal Society of London, 40, 42–73.

Mulaik, S. A. (2010). Foundations of factor analysis. New York, NY: CRC Press.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II. Philosophical Transactions of the Royal Society of London, A, 186, 343–414.

Jana Craig-Hare Jana Craig-Hare Craig-Hare, Jana

Cost–Benefit Analysis

Cost–benefit analysis

413

416

Cost–Benefit Analysis

Cost–benefit analysis (CBA) is a systematic approach used to evaluate the strengths and weaknesses of available options through a critical comparison of benefits and costs. Within the educational environment, this is a strategy used to evaluate the worth of educational programs and determine the added value in relationship to the monetary cost of the program. CBA has two overarching purposes: (1) to determine whether a program, investment, or decision is sound through verifying if its benefits overshadow the costs and by how much and (2) to provide a basis for comparing programs or projects by matching the total expected costs of each option against the total cost of benefits. Although one might assume that money is the driving force behind CBA, it is primarily used as protection for the well-being of individuals affected by the project or program. True economists want to measure the welfare, which is often a challenge because welfare cannot be directly measured. Instead, they use money as an expression of welfare and to assign "worth" to the program, allowing people to make decisions about programs based on this worth. Education is a form of investment in human capital, yielding economic benefits and contributing to the productive capacity of society. In this entry, how CBA is used in educational settings is analyzed. A comparison of CBA and cost-effectiveness analysis (CEA) is also conducted. The steps of conducting a CBA are also provided. Finally, the entry examines how CBA research has supported the importance of quality preschool education programs and concludes with a debate about the overall benefits of CBA in education.

CBA and Education

Educational leaders and policy makers are often faced with determining the best educational programs to serve their students. Investment at any level involves a sacrifice of some type to secure future benefits. CBA is a particularly useful tool for examining these educational programs and interventions. The framework for CBA allows one to compare the costs and benefits to the various policies and/or program alternatives. This technique can be used as an attempt to compare the monetary value of benefits with the monetary value of costs. By calculating the costs and benefits of alternatives in terms of monetary values, it becomes easy to compare components such as rates of return on investment, net differences between costs and benefits, and benefit to cost ratios. The goal, however, would be for educational leaders to select programs that maximized the total benefits relative to costs.

To perform a CBA of alternatives, it is critical to assume that the benefits or outcomes can be valued by their market costs or comparable alternatives. Yet oftentimes programs in the social sciences do not have a market counterpart. For example, if a program is designed to improve student learning, how can one get a market price for student achievement? Benefits of a program, however, can be reflected by increased graduation rates and the added value of students being prepared for future college and career opportunities; it’s often difficult to quantify and place a monetary value on all of the benefits of a program.

CBA Versus CEA

The effectiveness of a strategy may be expressed in terms of its actual outcome instead of its monetary value. Monetary measures of resource costs, in this instance, are related to the effectiveness of a program to produce a particular impact or outcome. In cases in which the effectiveness of a program to achieve a particular goal is linked to costs, the method is considered to be a CEA, rather than a CBA. CEA is gaining traction in educational program evaluation, as one might examine various alternatives for raising the literacy level of a population, increasing attendance rates for secondary students, reducing dropout rates, and so on. CEA allows us to rank potential programs according to the significances of their effects relative to their costs but prevents us from equating the costs directly to the benefits. CEA has been utilized to compare educational alternatives related to class sizes, the length of the school day, computer-assisted instruction, and peer tutoring. Oftentimes, however, these analyses only compare the effectiveness of the alternative programs, neglecting to consider the costs associated with the alternative programs.

associated with the alternative programs.

CBA and CEA are both valuable tools for program evaluation. Although CEA is a method that relates the costs of a program to its key outcomes or benefits, CBA takes that process further by attempting to compare costs with the dollar value of an analysis and can be applied at any time before, during, or after a program implementation. Both CBA and CEA can greatly assist decision makers in assessing the efficiency of a program.

Steps of CBA

The actual process of conducting a CBA is rather sophisticated, with inherent challenges in estimating and calculating program costs and benefits. There are, however, practical steps that can be utilized in this type of analysis. Henry Levin, a leading advocate for the use of CBA in program evaluation, advocates a step-by-step approach to CBA: (a) determine the resources (ingredients) used by the program, (b) determine the costs of the resources on a common metric, (c) measure the monetary costs of all products or outputs from the program, and (d) develop different cost–benefit ratios for all groups involved. Determining costs for individual programs with an educational system, however, has unique challenges. Educational programs are typically not funded by external agencies; therefore, they usually operate within the total district or school budget. It is difficult to determine the actual cost of a program from the overall budget. Personnel, a primary cost within a school district's budget, involves a variety of individuals who frequently work on many projects each day. It is difficult to accurately account for personnel time when individuals are working on multiple projects within environments such as this.

The first step in the CBA process is to identify all the costs and benefits associated with the project or decision. This list should be comprehensive, including all direct costs as well as indirect costs, and other costs such as intangible, opportunity, and the cost of potential risks. Benefits, as well, should be comprehensive, including all direct and indirect revenues and intangible benefits. All items on the list should then be assigned a common unit of monetary measurement. Typically, it's best to take a conservative approach with a conscious effort to avoid bias. Finally, the aggregate costs and benefits should be compared to determine whether the benefits outweigh the costs. If the benefits are favorable, stakeholders might choose to continue with the project or program. If not, they might review aspects of the project to determine whether adjustments such as increasing the benefits and/or decreasing the costs can be adjustments such as increasing the benefits and/or decreasing the costs can be made to make the project worthwhile. If adjustments cannot be made, the project or program may be discontinued in the future.

CBA of Preschool Education

One significant area of study has been surrounding preschools. It has been believed that quality early childhood education, particularly for children from low-income households, improves a child’s foundation for learning, as well as has the potential to reduce children needing to repeat a grade, being placed in special education, and committing juvenile crimes. As such, this strong foundation may also improve high school graduation rates and students participating in postsecondary educational opportunities, lead to fewer teen pregnancies, and lower the need for public assistance. These outcomes, which may in part be due to receiving a quality preschool education, contribute to the overall benefit of a society. The benefits, however, can be compared with the costs of preschool, to the degree that one can put monetary values on these benefits. Over two decades of research has been conducted through experimental and quasi-experimental research to weigh these results within a cost–benefit framework.

The evaluation of Perry Preschool represents one of the most extensive studies of preschool programs using a cost–benefit approach. In 1963 and 1964, innercity children were randomly assigned to either the preschool intervention treatment group or a control group that did not receive the preschool intervention. During their academic careers and into adulthood, students were periodically surveyed and follow-up evaluations were completed regarding the educational and life outcomes of both the treatment (Perry Preschool) and control (no preschool) groups. Findings revealed that students in the Perry Preschool program were less likely to repeat grade levels or need special education services. Additionally, it was more common for these students to graduate from high school and continue on to postsecondary education. Later, they earned more money and paid more taxes. For every dollar invested in the preschool intervention program, the investment paid almost US$13. This created a cost–benefit ratio of 13:1. Higher tax revenues and lower government support costs associated with the treatment group were attributed to the benefits received.

This method has also been applied to increasing graduation rates for high school students in the United States. Five interventions were identified that reduced students in the United States. Five interventions were identified that reduced dropout rates, thereby increasing the number of students graduating from high school, which included two preschool/early childhood interventions, reducing class sizes in the early grades, increased teacher salaries, and a high school educational reform program. Each intervention included an associated cost for each additional high school graduate, allowing for the cost-effectiveness comparison to be applied. The analysis was extended to a CBA by comparing the fiscal costs to the fiscal benefits, associating the high school completions as a taxpayer benefit from this public sector investment in education. Overall, results indicated that all five interventions benefited the taxpayer and exceeded the costs of the investments made into the programs. By using a method such as CBA, one can obtain research-based evidence that educational interventions are effective but also can be a sound investment for society.

Debating the Benefits of CBA in Education

There remains some controversy regarding the use of CBA in education. CBA requires that all alternative uses of the resources must be known to place value on the resources and identify the cost of the program. In education, however, one may not know all the possible alternative uses of resources. This goes back to the difficulty in obtaining utilization data for personnel to assign costs to the percentage of time they spent on the project.

CBA within educational research serves two primary purposes. First, it is important for school districts to have a greater understanding of how and where money is spent, including activities that receive the most time, money, and/or attention. Second, CBA allows for alternative education reforms or interventions to be compared based on economic costs. Oftentimes, cost analyses seek to expose the hidden costs of school initiatives to help districts implement them with fidelity. The importance of this purpose is relevant, because without the required monetary and personnel support, any school reform initiative is likely to fail. As the per pupil expenditures continue to rise over the past 30 years, education research stands to gain a great deal from the tools of economic analysis such as CBA.

Jana Craig-Hare

See also Policy Evaluation; Policy Research; Program Evaluation; Utilization-Focused Evaluation

Further Readings

Belfield, C., Nores, M., Barnett, S., & Schweinhart, L. (2006). The high/scope Perry preschool program: Cost-benefit analysis using data from the age-40 follow-up. The Journal of Human Resources, 41(1), 162–190.

Hummel-Rossi, B., & Ashdown, J. (2002). The state of cost-benefit and cost-effective analyses in education. Review of Educational Research, 72(1), 1–30.

Levin, H. (2001). Waiting for Godot: Cost-effectiveness analysis in education. New Directions for Evaluation, 2001(90), 55. doi:10.1002/ev.12

Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. Journal of Research on Educational Effectiveness, 8(3), 400–418.

Levin, H. M., & McEwan, P. J. (Eds.). (2002). Cost-effectiveness and educational policy. Annual Yearbook of the American Education Finance Association. Routledge.

National Research Council and Institute of Medicine. (2009). Strengthening benefit-cost analysis for early childhood interventions: Workshop summary. In A. Beatty, (Rapporteur), Committee on strengthening benefit-cost methodology for the evaluation of early childhood interventions, board on children, youth, and families. Division of behavioral and social sciences and education. Washington, DC: The National Academies Press.

Covariance

***See* Analysis of Covariance**; **Correlation**

Mary L. McHugh Mary L. McHugh McHugh, Mary L.

Cramér’s *V* Coefficient

Cramér’s *V* coefficient

416

418

Cramér’s *V* Coefficient

The Cramér’s *V* (also known as Cramér’s φ) is one of a number of correlation statistics developed to measure the strength of association between two nominal variables. Cramér’s *V* is a nonparametric statistic used in cross-tabulated table data. These data are usually measured at the nominal level, although some researchers will use Cramér’s *V* with ordinal data or collapsed (grouped) interval or ration data. Although an italicized capital *V* is most often used as the symbol for the statistic (*V*), the lowercase Greek letter φ with a subscripted c may also be used as follows: φ_c. This entry further describes the Cramér’s *V* and discusses its assumptions, calculation, and interpretation. It concludes with an example of the use of the Cramér’s *V*.

The *V* is a nonparametric inferential statistic used to measure correlation (also known as effect or effect size) for cross-tabulated tables when the variables have more than two levels. It is the effect size statistic of choice for tables greater than 2 × 2 (read two-by-two). Typical significance statistics for those tables include the chi-square and the maximum likelihood chi-square. The data in columns and rows should be nominal, although the *V* is frequently used with ordinal variables and collapsed interval/ratio data. Unlike the contingency coefficient, the *V* can be used when there are an unequal number of rows and columns. For example, the researcher should choose the *V* when the table has two columns and three rows.

The Cramér’s *V* was developed by Carl Harald Cramér, a Swedish mathematician known for his work on analytic number theory and probability theory. Based on Karl Pearson’s chi-square statistic, the *V* was developed to measure the size of the effect for significant chi-square tables.

The *V* is a correlation statistic, and as such, it measures the strength of an association between two variables. The *V* statistic provides two items of information:

> First, it answers the question, “Do these two variables covary?” That is, does one variable change when the other changes? (i.e., are the two variables independent?)
> Second, the size of the *V* describes the strength of the association. As the *V* approaches one level, the association is stronger. In a perfect correlation, for every one level of rise in one variable, the other variable would change exactly one level. The value of a *V* statistic can range only from 0 to +1.0; it cannot be a negative number. (Given that the calculation requires the square root of a number, the result cannot be negative with the standard formula.)

Many statistical computer programs (e.g., STATA, SPSS, and SAS) compute the *V* statistic as an option to accompany the output of the chi-square statistic, and the significance of *V* is the same as the significance of the chi-square.

Assumptions

Cramér’s *V*, like virtually all inferential statistics not specifically designed to test matched pairs or related measures, assumes that the sample was randomly selected from a defined population. It assumes subjects were independently sampled from the population. That is, selection of one subject is unrelated to selection of any other subject. Like the chi-square, there must be an adequate sample size for the computed φ statistic to be useful. The chi-square demands that 80% or more of the cell-expected values must be at least 5, and if this assumption is violated, neither the chi-square nor a φ calculated on the basis of that chi-square can be relied upon. It should be noted that samples smaller than 30 are considered to be very small samples, and small samples are less likely to be representative of the population of interest than larger samples. A sample size of 30 will, in most studies, provide a minimum of 5 for the expected values in all four cells.

Calculation

A great advantage of the *V* is that it is so easily calculated from the chi-square result. The calculation is as follows:

$$V = \sqrt{\frac{\chi^2 / n}{\min\,(r-1)\,\text{or}\,(c-1)}}.$$

Where $r - 1$ means the number of rows −1, $c - 1$ means the number of columns −1, and min means select the minimum of the two values.

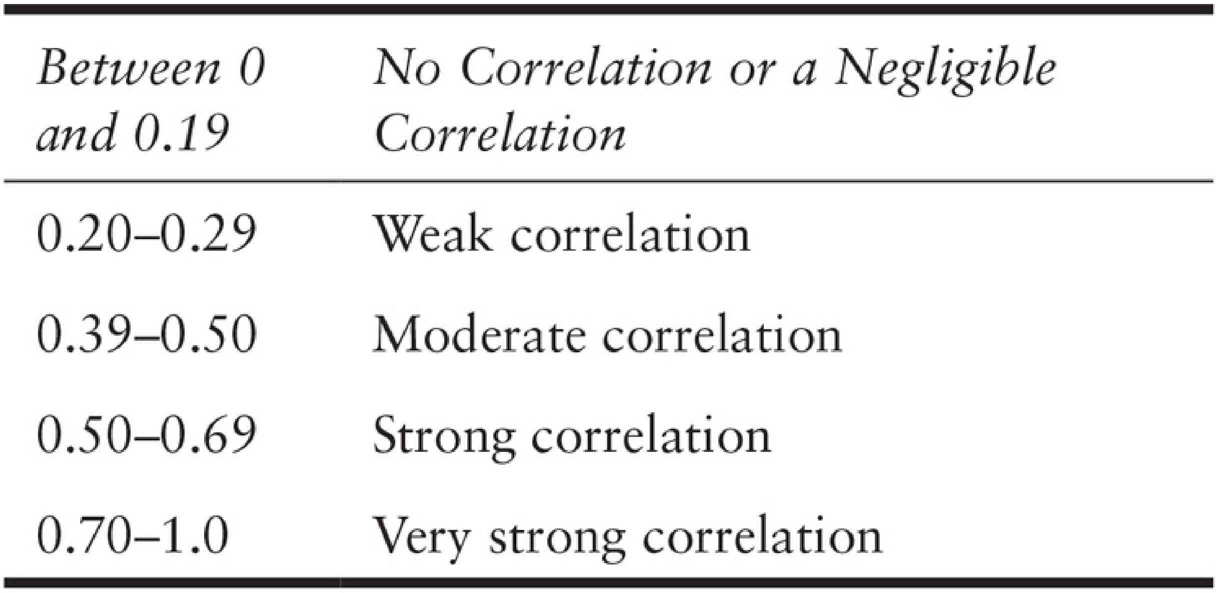
For example, if there are three rows and four columns, $r - 1 = 2$, and $c - 1 =3$. Thus, the chi-square ÷ n will be divided by the value, 2.

It is important to remember that an effect size statistic is useful only if the original chi-square was statistically significant. It is a mistake to conduct further analysis, such as effect size testing, if the original test of independence on the table fails to produce a significant result. When the chi-square (or Fisher's exact) on a 2 × 2 table is nonsignificant, the range of the confidence interval about the obtained *V* will contain the value of zero. Thus, calculation of the *V* is unnecessary because it is, by definition, not significantly different from zero.

Interpretation

Values for the *V* can range from 0 to +1.0. A value of 1.0 means there is a perfect 1 to 1 correlation between the two variables. Like the Pearson *r*, the *V* can be squared to obtain a measure of the amount of variance in the dependent variable that is explained by the independent variable. A *V* of 0.68 squared results in a value of 0.46, which means that the independent variable accounts for 46% of the variance in the dependent variable.

Although different authors may use different values for weak, moderate, and strong correlation measures, Table 1 can be used as a general guide to interpretation of the strength of effect size represented by various values of the *V*.

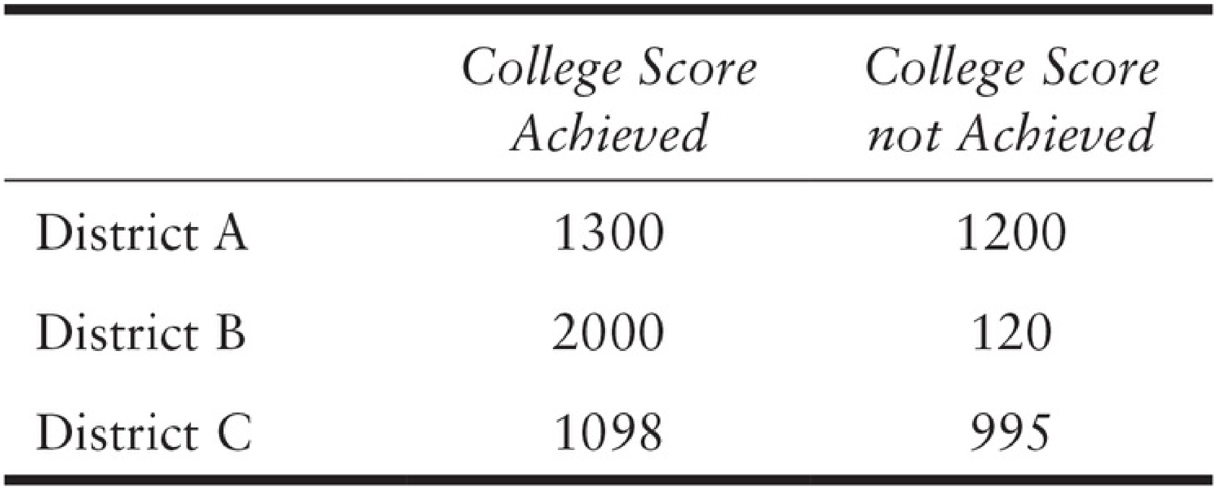


Between 0 and 0.19	*No Correlation or a Negligible Correlation*
0.20–0.29	Weak correlation
0.39–0.50	Moderate correlation
0.50–0.69	Strong correlation
0.70–1.0	Very strong correlation

These interpretations are based on the amount of variance in the dependent variable explained by the independent variable. A correlation of +0.29 means that even if statistically significant, only about 8% of the variance in the dependent variable is explained by the independent variable.

Example of Use of the Cramér’s *V*

In this hypothetical example, three large school districts’ populations are examined for the number of students achieving high enough scores for college admission on the SAT or ACT. In Districts A and C, students may take their choice of courses so long as they meet the state’s minimum graduation requirements. However, in District B (a more affluent district), college preparatory courses are mandatory for graduation. The results are presented in Table 2.



	College Score Achieved	College Score not Achieved
District A	1300	1200
District B	2000	120
District C	1098	995

The chi-square test of this table produces the following results: chi-square = 1139.62, $df = 2$, $p < .0001^2$. There are three rows and two columns.

Using the formula for Cramér’s V, the following values are required:

Chi-square = 1191.63

Sample size (n) = 6713

Number of rows = 3, Rows − 1 = 2

Number of columns = 2, Columns −1 = 1

$$V = \sqrt{\frac{\chi^2/n = 1191.63/6713 = .177417}{\min(3-1)\text{ or }(2-1) = 1}}$$

$$V = \sqrt{.177414} = .42.$$

The result is interpreted as follows: There was a significant difference among the three districts in the number of students achieving SAT or ACT scores high enough for college admission (chi-square = 1192, $df = 2$, $p < .0001$). The effect size of the relationship was 0.43, which is a moderately strong effect size.

Mary L. McHugh

See also Chi-Square Test; Pearson Correlation Coefficient; Phi Correlation

Coefficient; Spearman Correlation Coefficient; Two-Way Chi-Square

Further Readings

Norton, B. T. (1978, February). Karl Pearson and statistics: The social origins of scientific innovation. Social Studies of Science, 8(1), 3–34.

VassarStats. (2016). Chi-Square, Cramer’s V, and Lambda online calculator. Retrieved from June 22, 2016, VassarStats website: http://vassarstats.net/newcs.html

Jonathan A. Plucker Jonathan A. Plucker Plucker, Jonathan A.

Lorraine Blatt Lorraine Blatt Blatt, Lorraine

Creativity

Creativity

418

422

Creativity

Over the past 50 years, theory and research on creativity have advanced significantly. These advances can be seen across a number of domains and fields, including business, technology, health care, and design, and the implications for education have been significant. Indeed, scholars now have rich, detailed definitions and conceptions of creativity, considerable knowledge about enhancement of creativity, and comprehensive assessments for use in educational settings. The increased emphasis on creativity in education and the corresponding surge in creativity research have important implications for the definition of creativity and enhancements to creativity.

Scholars agree that a universal definition of creativity needs to encompass more than just the traditional notions of uniqueness and utility, expanding to include ideas such as tangibility, context, and surprise. Furthermore, there are a plethora of established strategies for enhancing creativity in the classroom, from pedagogical techniques such as divergent thinking training and modeling to external resources such as technology use and exposure to and interaction with outside communities. Finally, assessing creativity in the classroom is possible through a variety of instruments and techniques designed to measure creative products, process, people, and environments. This entry reviews definitions of creativity, research on enhancement efforts, and creativity assessments.

Definitions

Creativity is a term embedded in the lexicon, to the extent that it becomes difficult to detach the construct from the widely held myths and stereotypes surrounding it. These misconceptions include the notion that creativity is something we either are or are not born with. Additionally, creativity is often associated with socially isolated and dark or carefree and irrational behavior (such as the "loner" or "hippie" archetypes). Scholars believe many of these misconceptions result from imprecise definitions of creativity. Consequently, in order for research on creativity to be dissociated from these myths, it is important for researchers to clearly define creativity as they intend it to be used. Yet, in one study that examined published creativity research, only a third of articles included explicit definitions of creativity. As a result, scholars in the fields of education and psychology have published several articles attempting to resolve the absence of a definition for creativity.

The traits most commonly and overtly associated with the study of creativity are uniqueness and usefulness. Jonathan Plucker, Ronald Beghetto, and Gayle Dow, in their widely cited definition of creativity published in 2004, added the criteria of tangibility (e.g., an observable product) and the situation or context (acknowledging that whether something can be considered unique or useful is dependent on the existing environment and social framework). Their definition combines these criteria into one comprehensive definition:

> Creativity is the interaction among *aptitude, process, and environment* by which an individual or group produces a *perceptible product* that is both *novel and useful* as defined within a *social context* (p. 90).

Other scholars have provided similar definitions. Using U.S. Patent Office criteria, Dean Simonton asserts that something is creative if it includes some proportion of novelty, usefulness, *and* surprise. Although Plucker and colleagues and Simonton used different approaches to define creativity, their definitions are not mutually exclusive. The most obvious overlap in the two definitions is the emphasis on novelty and usefulness with the shared idea that those two traits alone are insufficient to define creativity.

Given these definitions, there are several models for how to conceptualize creativity, including the "4P model" and "5A model." In the 4P model, creativity includes *people* (the creators themselves and their corresponding personalities and attitudes), *process* (the actual procedures through which original, useful ideas are produced), the creative *product* itself, and *press* (the creators' context and how they interact with it). The 4P conceptualization has been used for several decades, and Vlad Glaveanu expanded and modified that model from a sociocultural perspective to create his 5A framework: *actors* interacting with their social context, creative *action or activity, artifacts* (products in cultural context), *audiences* as a component of press, and *affordances*, those activities facilitating interactions between actors and audiences. Regardless of the conceptualization, the key theme across these and other popular models is that creativity is a complex, multifaceted construct, providing multiple pathways for teachers to foster and assess student creativity.

Enhancement

Researchers have identified a number of strategies for enhancing creativity—in addition to barriers within schools that may serve as barriers to the development of student creativity. This section reviews the use of specific teaching strategies, game-and play-based interventions, collaboration, technology, and interaction with outside-of-school communities.

Pedagogical Techniques

First and foremost, teachers can equip students with the correct attitudes about creativity. As previously mentioned, there are several myths surrounding creativity. If teachers debunk these myths for students and teach them that creativity is something that can be learned and achieved deliberately, students will understand that their creative potential is not predetermined.

Teachers can also enhance creativity by encouraging creative ideation, also known as divergent thinking, which is the idea that any given problem has multiple solutions. For example, teachers can train divergent thinking by demonstrating a few different methods for arriving at a solution to a math problem, having students come up with their own hypotheses for science experiments, and instructing students to interpret texts from multiple angles and perspectives. These divergent thinking training approaches enhance creativity by improving the fluency, flexibility, originality, and elaboration of students' ideas.

Another area in which educators can foster student creativity is the area of creative articulation, a concept designed to help explain how creators select potential audiences for their creative work and use communication and persuasion to maximize the value of their creative work in the eyes of those audiences. In the real world, creativity does not stop at the idea or product stages rather continues in a cycle of feedback and revision as the creator or creative team shares their work with various audiences and receives constructive criticism. Yet few opportunities are provided within schools for students to share their creative work and receive feedback. Perhaps more to the point, opportunities for students to learn how to provide constructive feedback are limited. Whenever possible, teachers should provide opportunities for students to share their work with their peers, other educators, and community members, and students should receive guidance and practice as they learn how to provide constructive criticism when evaluating others' creative products.

Teachers can also exhibit and teach creative ideation with a willingness to deviate from the lesson plan to indulge in an unplanned opportunity for learning. For example, when a teacher is explaining themes in a novel, a student might ask a question about how one of the themes directly relates to a current event affecting the school's community. Then, the teacher can choose to disregard the lesson plan, at the risk of not covering all of the novel's themes, to instead facilitate a discussion inspired by the student's question. By taking advantage of this unplanned opportunity, the teacher is providing a more creative and enduring learning experience for the students. When students learn from a teacher who engages with their questions and relevant context in this way, they are also learning from the teacher's willingness to be spontaneous. In this sense, the teacher is modeling creativity.

Creative modeling is one of the most powerful strategies teachers can use to cultivate students' creativity. Recent studies suggest that students who experience creative modeling are more likely to exhibit creative behavior themselves. For example, Xinfa Yi, Plucker, and Jiajun Guo found that when they exposed students to creative models (such as collages and drawings with the same subject matter the students would have to replicate), the students demonstrated significantly more creativity, technical quality, imagination, artistic level, and elaboration, and gave a better general impression on a series of creativity assessments than students who were not exposed to creative models. Modeling creativity is especially effective for disadvantaged students who are less likely to have exposure to creative models outside the classroom. When teachers demonstrate creative behavior themselves, or even when teachers expose students to the creative work of others, students are more likely to be expose students to the creative work of others, students are more likely to be open to new ideas and risk-taking.

Research also shows that game-based and play-based interventions are successful strategies for developing creativity. These types of activities merge real world situations with imagination. This forces students to engage with material creatively yet comfortably, given that games and play are already so integral to students’ everyday lives. For example, pretend play, where students are forced to take on roles and perspectives that may not match their own, can advance students’ creativity by helping them process their emotions and practice the real-world scenarios. Lessons executed through games and play also provide opportunities for peer collaboration which allows students to enhance creativity in one another.

Collaboration among peers is a mixed blessing for creativity. On the negative side, social pressures can make some students reluctant to contribute to the group’s creative work, and group members’ ideas may serve as constraints on subsequent idea generation within the group. On the positive side, under the right framework, peers can be an asset to one another’s creative processes. When teachers provide a structured environment for collaboration, such as assigning roles within a group or laying out expectations for peer feedback, students are more likely to exchange ideas and problem solve in creative ways that they may not have been able to achieve individually.

External Resources

In addition to teachers and peers enhancing creativity in the classroom, external resources can provide students with distinctive opportunities that facilitate creativity. Teachers often shy away from some of the pedagogical techniques mentioned earlier because of outside pressure to focus on more standardized and concrete achievement; practices that enhance creativity are not necessarily synonymous with those that improve test scores or maximize factual knowledge. So, external resources such as technology and creative communities of practice can help supplement the creativity enhancements the teachers may be unable to directly provide.

Technology can be a helpful tool for fostering student creativity. For example, with video editing software, students are able to explore and express their story-telling abilities through animation or filmmaking, which introduce elements of complexity that can elevate the creative process. Additionally, technological complexity that can elevate the creative process. Additionally, technological tools like geographic information systems give students a chance to apply and improve their visual–spatial skills. This type of technology has the ability to unlock and develop creativity in areas that students would not normally be exposed to. Therefore, by integrating technology with the curriculum, students can make connections and tap into ideas that traditional teaching methods may not allow for. This can also occur when teachers take students outside the physical setting of the classroom or invite creative practitioners into the classroom.

Students’ creativity benefits from exposure to creative and professional communities beyond their schools. For example, if students visit an art gallery to see original artwork and witness the process of the creation of a piece of art, they will develop a new understanding and perspective about how art can be created. Similarly, if a bank manager visits a math class to teach students about how knowledge of interest rates affects loans in the course of solving a real-word problem, students may develop a deeper understanding of the practical uses of math that may lead them to approaching a calculus problem in a different way. Thus, external environments and practitioners offer new and diverse circumstances and scenarios for students to engage with educational material, helping students fulfill their creative potential.

Assessments

Although the conventional wisdom is that high-quality assessments of creativity are not available for use in educational settings, researchers have developed extensive measures of creativity that are appropriate for a variety of uses in classrooms and schools.

Creativity is traditionally assessed from four different perspectives: the assessment of creative products, creative processes (cognition), creative people, and creative environments. Each perspective is marked by a rich history of instrument development and applied assessment within educational settings.

Creative product assessments allow teachers to focus on the general level of creativity of student products or specific characteristics that are associated with creative products. From the specific characteristic perspective, there are instruments such as the Student Product Assessment Form, which allows teachers and students to evaluate products along several dimensions, including originality, attention to audience concerns, and problem focusing. At the other end of the spectrum, the consensual assessment technique involves a general evaluation of a product’s creativity by outside raters. These types of assessments lie on different ends of a broad continuum, but these and related product assessments allow educators and students to provide formative and summative feedback to students regarding the creativity of their work.

Creative process measures have traditionally focused on divergent and convergent thinking. Divergent thinking measures, such as the Torrance Tests of Creative Thinking, are traditionally the most popular creativity assessment in schools. Convergent thinking measures include the Remote Associates Test, in which students are provided with three, seemingly unrelated terms and asked to identify how they are related. For example, if the three unrelated terms are *cube, skate,* and *cream,* the student should identify *ice* as the word that relates them.

In educational settings, both divergent and convergent thinking measures have been used primarily to identify creative talent. Although convergent thinking measures have fallen out of favor since the 1970s, divergent thinking assessments are experiencing a resurgence, in part due to advances in scoring and score interpretation. However, despite the rich research base on divergent thinking measures, researchers’ understanding of creative cognition has expanded beyond the divergent–convergent distinction, and assessments based on these measures have not received as much attention as they deserve, both in experimental and educational settings.

A number of creativity instruments concerning individual characteristics of creators have been developed, and these may be useful for educators in a few different ways. For example, creative personality scales, such as the Gough Creative Personality Scale, may help identify students who have already developed the necessary attitudes for long-term creative productivity. Such measures have also been used as pre-post measures for creativity interventions. For the purposes of identification, instruments such as the Hocevar Creative Behavior Inventory, in which students identify their key accomplishments from a list of possible creative activities, can be useful as they are based on the belief that the best predictor of future creative behavior is past creative behavior. These “person measures” tend to be short and easy to administer, as most are self-report scales.

Given the importance of environmental factors in creative development, the lack of widely used environmental measures (e.g., a creative classroom environment scale) is surprising. Such instruments have been developed for workplace environments, but they have not been studied or widely used in K–12 education settings.

One important limitation of all of these instruments is that many have not been normed with diverse populations of students. The Torrance Tests are an exception, but most other measures have limited evidence of psychometric integrity with economically disadvantaged, non-White students. Given that such students now constitute over half of the K–12 student population in the United States, there is a need for research to address this limitation.

Jonathan A. Plucker and Lorraine Blatt

See also Alternate Assessments; Motivation; Personality Assessment; Surveys; Torrance Tests of Creative Thinking; Triarchic Theory of Intelligence

Further Readings

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. Educational Psychologist, 39(2), 83–96. doi:10.1207/s15326985ep3902_1

Plucker, J. A., Guo, J., & Dilley, A. (2017). Research-guided programs and strategies for nurturing creativity. In S. Pfeiffer (Ed.), APA handbook of giftedness and talent. Washington, DC: American Psychological Association.

Simonton, D. K. (2012). Taking the U.S. Patent Office criteria seriously: A quantitative three-criterion creativity definition and its implications. Creativity Research Journal, 24(2–3), 97–106. doi:10.1080/10400419.2012.676974

Simonton, D. K. (2016). Defining creativity: Don't we also need to define what is not creative? The Journal of Creative Behavior, 0(0), 1–15. doi:10.1002/jocb.137

Yi, X., Plucker, J. A., & Guo, J. (2015). Modeling influences on divergent thinking and artistic creativity. Thinking Skills and Creativity, 16, 62–68. doi:10.1016/j.tsc.2015.02.002

Credential

422

Credential

See Certification

Jill Hendrickson Lohmeier Jill Hendrickson Lohmeier Lohmeier, Jill Hendrickson

Criterion-Based Validity Evidence Criterion-based validity evidence

422

425

Criterion-Based Validity Evidence

Criterion-based validity evidence is frequently referred to as criterion-based validity, criterion-related validity, or simply criterion validity. In social science research, understanding the psychometric properties of an instrument is essential. These important psychometric properties include reliability and several types of validity. Thus, a researcher must often assess the evidence for the face validity, construct validity, content validity, and/or criterion-based validity of the instruments used in research. Although all forms of validity evidence indicate how well a measure measures what it is supposed to measure, criterion-based validity evidence is related to how accurately one measure predicts the outcome of another criterion measure. If a measure is a valid indicator of a construct of interest, then that measure could be used to predict the values of other measures related to that construct. Therefore, a measure that has high criterion validity would be one in which knowing the value of the predictor variable would allow the researcher to predict the value of the other criterion measure with high accuracy. Furthermore, the criterion-based validity applies to the validity of the predictor variable, not the criterion variable.

Criterion-based validity evidence is often of primary concern in educational research and assessment because educators are frequently looking for ways to determine whether assessment measures will be able to predict success or failure in later educational endeavors. Although other measures of validity are assessed using the opinions of experts and the similarity of a measure to other useful measures of the same construct, criterion-based validity is generally calculated and reported in quantitative measures from correlation and regression analyses. The general term *criterion validity* can include measures of predictive validity, concurrent validity, and postdictive validity. It is important to note that there are several threats to the criterion-based validity of conclusions drawn from the use of instruments that researchers must account for when using measures to predict criterion variable values. The remainder of this entry further describes the three types of criterion-based validity, highlights limitations of criterion-based validity evidence, and demonstrates how to calculate a measure of criterion-based validity.

Types of Criterion-Based Validity

The differences among the three types of criterion-based validity differ primarily in terms of whether the predictor variable precedes, occurs concurrently, or follows the criterion variable. Experiments can be designed to assess any of these three types of criterion-based validity, although the three types are not generally assessed with equal frequency in educational research.

Predictive Validity

Predictive validity is specifically related to how well a predictor variable predicts the values of a future criterion measure. Educational researchers are most often concerned with this type of criterion-related validity because they often want to know how well one can predict whether a student will succeed in later educational endeavors or in their careers. For example, college admissions were, at one time, focused on how well a high school student’s standardized test scores and grade point average (GPA) could predict the likelihood of that student graduating from their college. Thus, the predictive validity of those standardized test scores and GPAs has long been of interest to college admissions departments.

Concurrent Validity

Concurrent validity considers whether two separate measures taken at essentially the same time can predict the value of each other. For example, one might be concerned with whether the interest levels for a social media application of students from one geographic region could be used to predict the interest levels of students from a different region in that same application. Although concurrent validity might often be considered in marketing, it is also important in education. It is often useful to understand issues like how well students’ feelings of connectedness to their school might predict the graduation rates and how well connectedness to their school might predict the graduation rates and how well graduation rates might predict the sense of connectedness at a school. If they predict each other well, then the two measures have high concurrent validity.

Postdictive Validity

Postdictive validity is an indication of how well a predictor variable can be used to predict the value of a criterion measure taken previously in time. This type of validity is primarily used in studies in which the impact of educational policy might be considered. For example, a researcher might consider how well college GPA could predict whether students attended preschool. Although this postdictive validity of measuring college GPA may not be useful in the ways that predictive validity can be used for selection criteria or to implement programs to change predicted outcomes, postdictive validity can still be useful in determining whether a measure is a good indicator of a construct of interest. So in the aforementioned example, if the construct of interest is writing skills and a researcher knows there is a relationship between writing skills as an adult and college GPA and that there is a relationship between writing skills as a child and preschool attendance, then the criterion-based validity of a measure like college GPA could be considered in terms of the postdictive validity of predicting preschool attendance because it is an indicator of early writing skills. Postdictive validity might also be important when trying to determine which criterion variables might be related to later predictive variables. If a researcher wanted to consider variables related to failure at the college level, the researcher may look at the postdictive validity of college GPA before dropping out for predicting several childhood variables. Once postdictive validity was established for GPA on some of those variables, the relationship between those variables could be further explored.

Limitations of Criterion-Based Validity Evidence

As with all measures of validity, there are threats to criterion-based validity evidence. Several issues surrounding criterion-related validity evidence are important for researchers to keep in mind when making assumptions based on this type of validity. The first is that although criterion-related validity looks beyond simple correlations, it still does not allow one to assume causation between the measures. Although a measure may accurately predict the value of another measure, it cannot be assumed that changes in one measure will therefore cause the values in the other measure to change. In the aforementioned therefore cause the values in the other measure to change. In the aforementioned example, one can clearly not conclude that earning a high GPA in college causes one to attend preschool, even if there may be high postdictive criterion-based validity evidence for college GPA predicting preschool attendance. More importantly, the less obviously incorrect conclusion that attending preschool causes one to have a higher college GPA cannot be made. Although it may be true, the criterion-based validity evidence is not enough to indicate this causal relationship. Therefore, in order to draw causal conclusions, the researcher would have to conduct a study with an experimental design. In other words, researchers solely using criterion-based validity evidence can only know that the predictive or postdictive relationship exists. Moreover, the relationship could be caused by the impact of numerous other factors, such as parental education levels or interest in writing, on both measures.

A second threat to criterion-based validity evidence is the use of a restricted range of values for the predictor variables in many of the educational measures used for prediction. A common example of this phenomenon is when colleges use their selection criteria to predict college success in terms of GPA or years to graduate. Generally, those students who apply to the most competitive colleges self-select for the first stage of college admissions. That is, for the most part, only those students who have a relatively high GPA and relatively high standardized test scores apply to the competitive colleges. Thus, when looking at how well the GPA or test scores can be used to predict the outcomes of interest, the values of the predictor variables are generally in the top quarter of the true range for those variables. This restricted range can then suggest stronger correlations than truly exist when considering data from the full range.

A third threat to criterion-based validity evidence is measurement error. Although this validity refers to how well the predictor variable predicts the criterion variable, the researcher is generally interested in a construct represented by the criterion variable measure. If the criterion measure does not have good construct validity (i.e., it does not represent the construct of interest well), then the criterion-based validity is not accurate because the predictor variable is actually predicting the values of something other than a measure of the construct of interest.

Similar to measurement error and a restricted range, selection bias can also lead to false measures of criterion-based validity. If a researcher uses data from a biased sample to determine criterion-based validity, the conclusions may not be transferable to other populations. For example, if a researcher collects data from the five largest school districts in a state to determine the criterion-based validity of state-mandated testing in elementary schools for predicting high school graduation rates, the researcher may find high criterion-based validity. However, the state tests may not actually predict high school graduation rates in small rural districts that were not included in the original study because of their small size.

A final important point to consider is that although criterion-related validity evidence may suggest that one measure can predict the value of another measure, it does not explain why the relationship exists. This concern goes beyond the inability of a researcher to draw causal conclusions about the relationship between the predictor and criterion variables. Regardless of causation, when a measure has high criterion-based validity, it is likely that there are some important relationships between the variables and other related variables. Researchers must look beyond an indication of criterion-based validity to truly understand those relationships.

Calculating a Measure of Criterion-Based Validity

The index of criterion validity is the correlation between the predictor variable and the criterion variable. Although different correlation measures can be calculated depending on whether the variables are continuous, the statistic of primary interest with criterion-based validity evidence is the effect size for the correlation. The effect size of the correlation should be considered in order to understand how well the predictor variable predicts the criterion variable. A statistically significant correlation would only indicate a relationship between the predictor and criterion variables; however, a researcher interested in criterion-based validity evidence is not simply interested in whether the relationship exists. The strength of the relationship, and thus, the effect size, is the statistic of interest with criterion-related validity evidence. If multiple variables are used to predict the value of a criterion variable, the predictor variable with the largest effect size is said to have the most criterion-based validity.

Often researchers are not simply interested in whether a variable has high criterion-based validity but instead are interested in using the variable(s) with high criterion-based validity to calculate predicted values for the criterion variable. In these cases, a regression equation can be used to allow the researcher to calculate the predicted values. When multiple predictor variables can be used, a multiple regression equation can be used to calculate the predicted values of the criterion variable.

Jill Hendrickson Lohmeier

See also Concurrent Validity; Content-Related Validity Evidence; Predictive Validity; Psychometrics; Restriction of Range; Validity

Further Readings

Huang, C. (2012). Discriminant and criterion-related validity of achievement goals in predicting academic achievement: A meta-analysis. Journal of Educational Psychology, 104(1), 48–73.

Murphy, K. R., & Davidshofer, C. O. (1988). Psychological testing: Principles and applications. Englewood Cliffs, NJ: Prentice Hall.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York, NY: McGraw-Hill Education.

Wainer, H., & Braun, H. I. (2013). Test validity. Hillsdale, NJ: Routledge.

Kyle Nickodem Kyle Nickodem Nickodem, Kyle

Michael C. Rodriguez Michael C. Rodriguez Rodriguez, Michael C.

Criterion-Referenced Interpretation Criterion-referenced interpretation

425

428

Criterion-Referenced Interpretation

Criterion-referenced interpretation is the interpretation of a test score as a measure of the knowledge, skills, and abilities an individual or group can demonstrate from a clearly defined content or behavior domain. It is often defined as a contrast to norm-referenced interpretation, where an individual's score only has meaning when it is compared to other individuals' scores. Criterion-referenced interpretations are independent of information based on how the average person performs. This entry further describes criterion-referenced interpretation and its uses, then discusses the design and validation of tests that foster criterion-referenced interpretation. The entry concludes with a look at common misconceptions about criterion-referenced interpretation.

Criterion-reference interpreted scores have been used for a variety of decisions, such as monitoring student achievement, evaluating efficacy of instructional programs, granting licensure and certification, planning individual and group instruction, and identifying possible learning disabilities. Tests that are designed to foster criterion-referenced interpretation of scores include Advanced Placement assessments, driver's license exams, and the Programme for International Student Assessment.

A criterion-referenced interpretation assumes an underlying continuum of content knowledge and behaviors that ranges from none to all encompassing. When the breadth and depth of knowledge and behaviors that comprise the content domain—the criterion—is clearly and completely specified, and a test is constructed with a representative sample of items from the content domain, it is understood that there is a correspondence between an individual's performance on the test and their ability level on the underlying continuum. Thus, if a test is constructed to foster a criterion-referenced interpretation, the inference can be made that an individual who scores 75% on the test knows and is able to demonstrate individual knowledge of 75% of the content domain.

First outlined by Robert Glaser in his 1963 symposium address to the American Educational Research Association, criterion-referenced interpretation gained popularity in the United States in the 1970s, as the development of theories of measurement and test design refined the distinctions between criterion-referenced and norm-referenced interpretations and their uses. Although it is possible under certain conditions to interpret scores from a single test in reference to both a criterion domain and a norming group, doing so rarely leads to satisfactory interpretations because different score interpretations require different test designs. It is important to note that the nature of test score interpretation (criterion-or norm referenced) is a characteristic of the interpretation as enabled by test design, not the test itself. There is a tendency in the measurement and assessment literature to refer to anything not explicitly norm referenced as criterion referenced. Here, the description of criterion-referenced interpretation is consistent with the original intent.

Design and Validation

As with all well-developed tests, in tests designed to foster criterion-referenced interpretation, the purpose, content domain, test specifications, and item specifications are defined. A key component of the content domain that supports criterion-referenced interpretation is that it covers a relatively narrow set of cognitive skills (although this is not a technical requirement), so that the resulting test sufficiently measures performance within the domain. This requires the test developer to define the boundaries of skills relevant to the content domain as well as the types and formats of problems and scoring rules that delineate membership of appropriate items and tasks. This recognizes natural variability in item difficulty as a function of conceptual difficulty of items and tasks, the complexity of relevant contexts, and recognition of the natural progress of skill levels in the well-defined domain. The result is a pool of carefully constructed items deeply measuring performance to support criterion-referenced interpretations.

Ideally, once the criterion is well defined, items and tasks are generated that cover the entire expanse of the content domain. From this pool of items, a representative sample is drawn to construct the test. The representative sample of items allows for the correspondence between performance on the test and ability on the underlying knowledge continuum to be established. In practice, however, areas of the domain that are more easily measured tend to be overrepresented, even when they are more peripheral content.

Assuming that items are of high quality, ensuring that the items chosen are a representative sample of the content domain is, theoretically, the only concern regarding item selection for fostering criterion-referenced interpretation. Unlike norm-referenced interpretations, criterion-referenced interpretation does not depend on the variability of scores between test takers. Thus, items that are extremely difficult or extremely easy can be included if they address a fundamental skill or knowledge expected of test takers.

Lack of score variability also means that tests designed to support criterion-referenced interpretation are likely to produce low item-total correlations as measures of item discrimination and result in low internal consistency reliability in a classical test theory sense (thus such estimates are inappropriate for scores intended for criterion-referenced interpretation). Other estimates of score consistency are more appropriate, including decision or classification consistency.

The length of the test is dictated by the scope of the content domain and whether score interpretation is for individuals or groups. The broader the content domain, the longer the test will likely need to be in order for the sample of items to adequately cover the domain. Additionally, individual-level score interpretation requires longer tests because each test taker must respond to items that are representative of the entire content domain. However, group-or program-level score interpretations can be supported with fewer items because the content domain only needs to be appropriately represented when scores are aggregated. This means it is possible for each test taker to only respond to items that cover a portion of the domain as long as the entire domain is covered when aggregated to the group-or program level.

A variety of objective and subjective scoring methods can be used to support criterion-referenced interpretation. Selecting the appropriate scoring method largely depends on the nature of the content domain and the target audience. Although a typical scoring method is to calculate the number or percentage of items answered or tasks performed correctly, this method is not the most meaningful for all criterion domains. For instance, the speed of completing the meaningful for all criterion domains. For instance, the speed of completing the task, such as running a mile or calculating single-digit multiplication, might be of greater importance, especially when the task itself is relatively easy to complete for the intended population. In other contexts, the precision of performance is of greater interest, as when transcribing an interview or using a rubric to score the quality of a test taker’s essay. Many standardized tests employ more sophisticated scoring methods using item response theory or Bayesian estimation along with additional scaling considerations to generate final scores. Regardless of the scoring method employed, the theoretical rational and the procedures used to produce the scores need to be well documented in order to support criterion-referenced interpretation.

Although cut scores or performance standards are not required for criterion-referenced interpretation, they are often set in order to aid decisions based on the criterion-referenced interpretation of scores. Performance standards or cut scores categorize test takers into two or more performance or mastery levels. For instance, when score interpretations are used for granting certification, a cut score might be set at 85% correct, whereby test takers who answered 85% or more of the items correctly are granted certification. Although rationale and evidence must be provided to justify the use of cut scores, testing standards dictate circumstances under which cut scores can be established and defended, where to set the cut score is a policy decision based on judgment, often supported with empirical information.

Tests designed to support criterion-referenced interpretation are considered quota free, meaning that the number of test takers expected to score above or below the cut score should have absolutely no bearing on where the cut score is set. Instead, just as scores are interpreted in reference to what students are expected to know and do in a clearly defined criterion domain, cut scores should be set with explicit references to the criterion domain, not the relative performance of a reference group.

The primary validity evidence for interpreting scores from a test in reference to a criterion is a carefully and completely defined criterion domain. The criterion is the content knowledge and performance tasks an individual or group from a specified population is expected to know and be able to do under specified circumstances. This involves specifying whether certain skills or knowledge are of greater importance to the domain, whereas others might be more peripheral. Common procedures for defining the criterion domain include gathering judgments from experts in the domain, mutual consensus from a variety of people associated with the domain, and analysis of research and published works in the domain.

Although tests designed for criterion-referenced interpretation often are used to assess what individuals know and can do at the end of an instructional period, the criterion domain can be defined for any point in the instructional process where it might be useful to measure test takers’ current achievement. Each component of the test, including purposes, score interpretations, and uses, is subject to validation, where relevant and appropriate evidence is gathered in its defense.

Common Misconceptions

Tests that are specifically constructed to support criterion-referenced interpretation are commonly referred to as criterion-referenced tests; however, this attribution is misleading. Scores from a single test can be interpreted for multiple purposes. For instance, a score could be interpreted both as a measure of what an individual knows and can do (criterion-referenced interpretation) and as a measure of how individual abilities compare relative to other test takers (norm-referenced interpretation). Although some interpretations might be more appropriate than others based on the design of the test, criterion-reference is an attribute of the interpretation of scores and not the test itself.

Another common misconception regards the multiple definitions of the term *criterion*. With the prevalence of tests that utilize cut scores to categorize test takers into performance or mastery levels, many individuals mistakenly refer to the cut score, performance standard, or mastery level as the criterion (e.g., the criterion passing score). However, the criterion refers to the domain of knowledge and behaviors expected from a defined population under specified circumstances. In an attempt to alleviate possible confusion, the term *domain-referenced interpretation* is sometimes used in place of criterion-referenced interpretation (actually, this has been suggested by measurement specialists numerous times but has not been widely adopted).

A third misconception is treating objectives-referenced interpretations or standards-based assessment interpretations as necessarily criterion referenced. Objectives-based and standards-based score interpretations share many of the measurement and score reporting characteristics as criterion-referenced interpretations in that results offer insight into the behaviors and abilities individuals and groups can currently demonstrate. Many take this similarity to mean that tests designed for objectives-referenced and standards-based score interpretation reveal test takers’ knowledge and abilities for specific content domains when, in reality, the scope of the typical standards-based test is far too broad, where many content standards are lightly sampled.

Unlike criterion-referenced interpretation, objectives-referenced and standards-based interpretations do not require as carefully a defined content domain nor items to be a random or representative sample of the domain. Instead, objectives and standards are defined, which themselves are only a subset of the content domain that is expected to be taught. Thus, the inference drawn from the score is no longer what individuals or groups know and can do from the content domain, but what they know and can do from what they were expected to have been taught, in very general terms, because no specific objective or standard is well defined or measured. For school-level accountability, this might suffice as a general indicator; but for individual-level inferences about knowledge, skills, and abilities, this is insufficient.

Kyle Nickodem and Michael C. Rodriguez

See also Achievement Tests; Cut Scores; Instructional Sensitivity; Norm-Referenced Interpretation; Programme for International Student Assessment; *Standards for Educational and Psychological Testing*; Standards-Based Assessment; Trends in International Mathematics and Science Study

Further Readings

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519–521.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10(3), 159–170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48(1), 1–47.

Popham, W. J. (Ed.). (1971). Criterion-referenced measurement: An introduction. Englewood Cliffs, NJ: Educational Technology.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6(1), 1–9.

Rachel Elizabeth Kostura Polk Rachel Elizabeth Kostura Polk Polk, Rachel Elizabeth Kostura

Critical Thinking

Critical thinking

428

432

Critical Thinking

The construct of critical thinking has been widely embraced as a core cognitive skill that should be nurtured and emphasized throughout educational curricula at every grade level. A multitude of definitions have been suggested to describe critical thinking. The general consensus is that critical thinking involves metacognition, or thinking about thinking, to maintain awareness and to reflect and manage one’s own thoughts. Moreover, scholars emphasize the importance of recognizing one’s own biases and being willing to evaluate the validity of arguments that oppose one’s beliefs. Having this willingness implies that a person would have to accept that there is uncertainty about a belief or a solution to a problem and that the person is motivated to examine multiple perspectives or solutions. This entry begins by reviewing the various skills and dispositions incorporated in critical thinking. Next, the entry highlights educational programs designed to teach and improve critical thinking skills. The value and difficulty of assessing critical thinking, as well assessment tools, are then presented. The entry concludes by looking at current and future directions of critical thinking in academia.

Critical Thinking: Skills and Dispositions

With the numerous and variable descriptions of critical thinking, scholars have urged that we must clearly and specifically define critically thinking if we are to systematically assess and promote this construct in educational contexts. In 1990, a leading researcher on critical thinking, Peter Facione, led a panel of expert philosophers in defining critical thinking using the Delphi method. The experts submitted individual definitions of critical thinking, which they analyzed and fine-tuned until they reached a consensus definition. They concluded critical thinking consists of two aptitudes: skills and dispositions on the theory that it is insufficient to expect that a person who has critical thinking skills will simply use them; the critical thinker must also be inclined to practice the skills.

The Delphi consensus definition included six critical thinking skills and seven critical thinking dispositions. The skills are interpretation, analysis, evaluation, inference, explanation, and self-regulation. A critical thinker will interpret, analyze, evaluate, and draw inferences from information to form an evidence-based judgment. Moreover, the critical thinker can explain evidence, theories, concepts, methods, criteria, or contexts that support the judgment. By gathering and evaluating the judgment, the critical thinker can decide what to believe or how to proceed in a given situation. Some scholars call this process analytic reasoning. This type of reasoning involves breaking a concept down into different parts and studying how each part relates to the others. Therefore, analytic reasoning is a cognitive process of coming to an understanding of something believed, through a reasoned process of examining the parts of that belief. If a person is using analytic reasoning, the person is using critical thinking skills.

Critical thinking disposition is categorized as a personality characteristic; disposition indicates how one would approach a problem and use reasoning to solve it. A person with strong critical thinking disposition has high internal motivation to make decisions, solve problems, or evaluate ideas by thinking critically. The Delphi experts identified seven dispositions of critical thinking: inquisitiveness, systemactiy, truth-seeking, open-mindedness, self-confidence, analyticity, and maturity.

These dispositions are derived from a person’s learning style, approach to conflicting opinions, and mindfulness in making decisions. Inquisitiveness describes the individual’s general desire to learn. Systemacity is the tendency to ask questions in an organized and focused manner. Truth-seeking describes an individual’s pursuit of the most accurate knowledge—despite findings that don’t support a person’s own opinions or self-interest—and continual reevaluation of information, remaining honest and objective throughout the pursuit. Open-mindedness, a common factor in most definitions of critical thinking, indicates individuals’ tolerance of views different from their own and sensitivity to their own bias. Self-confidence describes how much trust individuals place in their own reasoning process when arriving at a judgment and their trust in themselves own reasoning process when arriving at a judgment and their trust in themselves to lead others to a solution. Analyticity is characterized by how individuals apply reasoning and evidence to a proposed solution, while anticipating potential difficulties and remaining alert to the need to intervene. Maturity is the disposition of approaching problems while being mindful of their complexity or poor structure, and that multiple solutions may be plausible yet uncertain, depending on the context and supporting evidence.

Critical thinking disposition involves having willingness to suspend judgment and having the patience to ensure judgments are based on contextual considerations or on reasons that are relevant and consistent with an argument. The critical thinker evaluates conflicting claims and personal biases before arriving at a conclusion and continues to evaluate the conclusion when new information is presented. According to Robert Williams, a critical thinking scholar, a person is deemed proficient in critical thinking if the person’s arguments are supported with evidence and if the person can accurately judge whether others’ arguments are sufficiently supported. Diane Halpern, a leading researcher on critical thinking, believes critical thinking skills increase a person’s chance to succeed in creating and adapting to change.

Teaching and Improving Critical Thinking Skills

Many studies suggest that critical thinking skills can be fostered in children, and one of the most prominent goals of education is to teach students how to think critically about complex topics. Educational programs have begun to invest more time and resources in strengthening 21st-century skills, which include critical thinking and analysis. Critical analytic skills are part of the brain’s executive functioning. Executive functioning consists of working memory, inhibition, and cognitive flexibility. Strong critical thinkers use these cognitive processes to transfer reasoning capabilities. For example, the critical thinkers will use their working memory to store and update information. The critical thinkers will use inhibition to maintain an awareness of their own beliefs and biases, which will allow the individuals to identify errors in their thinking and resolve conflicting information instead of making premature judgments. Increased cognitive flexibility means the critical thinker is more able to apply different rules to a situation based on a certain context. Although these brain functions develop naturally in most human beings, research also suggests that games or exercises that target executive functioning can improve critical thinking skills.

Scholars have encouraged tailoring instruction toward verbal reasoning, argument analysis, hypothesis testing, probability assessment, problem solving, and decision making. Additionally, critical thinking can be nurtured outside of formal education. Critical thinking skills and dispositions can be practiced through parent–child or peer–to–peer interactions. For example, parents can challenge their children to take different perspectives or draw inferences by telling stories and asking questions. During play, children can promote critical thinking skills in each other by creating shared goals and achieving goals through communication and cooperation. Scholars have recommended that teachers and caregivers engage children in conversations with the intent of promoting critical thinking and structure play so that children have the authority to pursue goals through peer interactions.

Encouraging children to reflect on their actions and activities also promotes critical thinking. By prompting a child to consider why an activity occurred, and what was gained from it, the child’s metacognitive skills are engaged. Also, children are more likely to have heightened interest in the purpose of the activity if the activity is not imposed on them, and they will be more inclined to think independently in forming arguments about the activity.

Overall, the goal of teaching critical thinking in the classroom is that students use critical thinking skills beyond an in-class exam. The most authentic practice and assessment of critical thinking would take place outside of the classroom, where a student could spontaneously apply critical thinking skills to real-world problems or arguments, without the teacher and other stimuli that might prompt the skills. Students should be observed in a naturalistic setting or simulated scenario, after a semester of instruction, to determine whether the student obtained lasting improvements in skills and dispositions.

Critical Thinking Assessment

A number of articles on the subject suggest that critical thinking is best measured using authentic assessment, where tasks within the assessment are similar to real-world activities. Halpern opined that the most important outcome of critical thinking measures is a person’s ability to transfer thinking skills. Projects that have relevance beyond the classroom provide an opportunity for students to show evidence of their higher-order thinking skills. This evidence often takes the form of completed performance assessments. Performance assessments comprise tasks that elicit observations of student performance.

However, creating standardized, large-scale assessments of this nature is challenging due to the complexity of items, difficulty in operationally defining abstract constructs, and maintaining objectivity in scoring. To meet these challenges, performance assessments require large amounts of time and resources. Student performances are evaluated using rubrics, where observable behaviors are separated into a series of criteria and scored according the description that best matches the student’s performance. Rubric development is challenging because the descriptive adjectives are inherently subjective, and their definitions require extended discussion before a consensus on their meaning can be reached. Moreover, rater training must occur before the raters’ rubric scores can be considered valid.

Performance assessments are widely emphasized as summative measures due to their scope, but they also have strong formative value; by reviewing and reflecting on rubric scores, students can understand their strengths and areas for improvement in thinking skills. Moreover, strong thinking skills could be interpreted as both a cause and an effect of completing of a performance assessment. A commonly utilized method of assessing real-word thinking skills is the use of performance items that are set in authentic contexts. For example, constructed-response items based on a given scenario would elicit analytic thinking that is relevant to real-world contexts. Several assessments were designed to measure critical thinking using this method.

The first test designed to measure critical thinking is the Watson Glaser Critical Thinking Appraisal. It was developed in 1964 to select and promote employees into management positions within their occupations. The test is administered using semi-structured interviews that require the employees to use logic to support their views. Scores have been shown to predict reflective judgment. Since the mid-1990s, the Watson Glaser Critical Thinking Appraisal has been more commonly used in educational contexts. In 1985, another logic appraisal test is the Ennis Weir Critical Thinking Essay Test, developed for students. Students are instructed to read an argumentative letter and evaluate errors in reasoning within each paragraph and within the letter as a whole. The scorer uses a guide to assign points to the student’s logic. Due to the highly structured nature of the test and scoring system, the test has high inter-rater reliability.

During the 1990 Delphi meeting, the task force developed two tests that measured the aforementioned skills and dispositions. Test development began with the shared assumption that critical thinking involves the same elements at each stage of human development, from childhood to adulthood. They believed each stage of human development, from childhood to adulthood. They believed that although the standards for performance differ depending on age, the underlying structure of critical thinking is constant, and these elements can be operationally defined at every level. The two tests that were derived from the Delphi expert consensus are the California Critical Thinking Skills Test, which measures the six critical thinking skills, and the California Critical Thinking Disposition Inventory, which measures the seven dispositions of critical thinking. The basic California Critical Thinking Skills Test has two forms, which make the test suitable for evaluating growth in critical thinking skills. The forms comprise multiple-choice items and span six subscales: interpretation, analysis, evaluation, inference, explanation, and self-regulation. The California Critical Thinking Disposition Inventory has one form, comprising 6-point Likert-type scales that address tendencies toward the seven dispositions. Extensive research has been conducted on the two tests, and they are widely used as a research tool. Advantages of utilizing these tests include increased reliability due to the standardized and objective nature of the tests, and their theoretical basis is founded on the Delphi expert consensus of critical thinking.

Another test measuring critical thinking skills and dispositions is the Halpern Critical Thinking Assessment (2010), which requires students to respond to scenarios through constructed-response items. The test has two forms that consist of 25 everyday scenarios and follow-up items based on five areas of critical thinking: verbal reasoning, argument analysis, hypothesis testing, probability, and problem solving. In the first form, the scenarios are presented with open-ended questions in the first half, followed by a series of multiple-choice questions that pertain to the scenarios. The second form of the test is shorter, comprising only the multiple-choice items. Research on this assessment indicates that the test has high reliability and validity.

Critical Thinking in the 21st Century

With a greater emphasis on teaching 21st-century skills in the classroom and lifelong learning in higher education programs, critical thinking will remain an important focus in the world of academia and in real-world contexts. The research on critical thinking shows that humans are naturally equipped with certain critical thinking skills and dispositions through their executive functioning, and these cognitive processes can be fostered within and outside of the classroom. One of the greatest indicators of strong critical thinking skills and dispositions is whether a person will intentionally apply the same reasoning skills to different contexts, maintaining an awareness of preconceptions and openness to new information. Moreover, subsequent opinions and decisions will be supported with solid evidence, and this evidence will be sought intentionally and continually reevaluated as more information is presented.

Rachel Elizabeth Kostura Polk

See also Constructed-Response Items; Delphi Technique; Metacognition; Multiple-Choice Items

Further Readings

Brown, N., Afflerbach, J., & Croninger, S. (2014). Assessment of critical-analytic thinking. Educational Psychology Review, 26(4), 543–560.

Dexter, P., Applegate, M., Backer, J., Claytor, K., Keffer, J., Norton, B., & Ross B. (1997). A proposed framework for teaching and evaluating critical thinking in nursing. Journal of Professional Nursing, 13(3), 160–167.

Kuhn, D., & Dean, D.Jr. (2004). Metacognition: A bridge between cognitive psychology and educational practice. Theory Into Practice, 43(4), 268–273.

Murphy, P., Rowe, K., Ramani, M., & Silverman, L. (2014). Promoting critical-analytic thinking in children and adolescents at home and in school. Educational Psychology Review, 26(4), 561–578.

Williams, R. (1999). Operational definitions and assessment of higher-order cognitive constructs. Educational Psychology Review, 11(4), 411–427.

Ryan W. Walters Ryan W. Walters Walters, Ryan W.

Cross-Classified Models Cross-classified models

432

435

Cross-Classified Models

Educational outcomes often result from two (or more) clearly hierarchical sampling dimensions, such as when students not only represent themselves but also represent some larger group. In multilevel modeling analyses, the students might be considered “Level 1” but be nested within schools at “Level 2” and would be modeled appropriately using the traditional multilevel model. However, when sampling dimensions are not clearly hierarchical, such as if students at Level 1 are simultaneously nested within more than one Level 2 variable (e.g., both schools and neighborhoods), the traditional multilevel model must be abandoned in favor of a cross-classified model. An understanding of the cross-classified model is critical given the potential curricular and fiscal policy ramifications that could result from incorrectly analyzing nonhierarchical, multilevel educational data. Therefore, this entry provides a brief overview of the cross-classified model and begins by detailing the transition from the traditional multilevel model to the cross-classified model. Then, the unconditional cross-classified model is presented to exhibit how this model partitions systematic variation at all levels, followed by a brief discussion of the random interaction effect. This is followed by a discussion of the complexity involved when estimating and interpreting the effects of predictor variables. The entry concludes with a brief overview of available software to estimate this model.

Before proceeding, it is important to consider that this entry is specific to cross-sectional data with a two-level structure. Specifically, all examples will consider students at Level 1 who are measured at one occasion and belong to one, and only one, Level 2 classification for schools as well as to one, and only one, Level 2 classification for neighborhoods. If, however, a sampling design considers students who attend multiple schools *or* live in multiple neighborhoods, then a multiple membership model would be required. By contrast, if a sampling design considers students to both live in a given neighborhood *and* attend multiple schools, then a multiple membership, multiple classification model would be required.

Transitioning to the Cross-Classified Model

The cross-classified model is similar to the traditional multilevel model in that the primary purpose of both models is to correctly partition all sources of systematic variation to ensure more accurate variance component estimates and, therefore, less biased statistical inference for predictor variables. For example, outcomes from students at Level 1 nested within the same school at Level 2 would be correlated as a result of the systematic variation due to schools. In this nested example, the traditional multilevel model would partition student-level variability from school-level variability prior to including student-and/or school-level predictors. However, when students at Level 1 are sampled from different schools and different neighborhoods at Level 2, it is unlikely that all students who attend the same school live in the same neighborhood or that all students who live in the same neighborhood attend the same school. Thus, the systematic variation due to schools and neighborhoods are not nested but instead are considered *crossed* at Level 2. To estimate the traditional multilevel model on crossed sampling, dimensions would require this model to be severely misspecified (e.g., ignoring either schools or neighborhoods, deleting observations to create a clear hierarchical structure). By contrast, the primary purpose of the cross-classified model is to account for both sources of systematic variation at Level 2 in a single model.

Because estimating the cross-classified model is computationally demanding, especially with increasing model complexity (e.g., student within schools and neighborhoods within cities), it is common to estimate separate traditional multilevel models for each Level 2 classification prior to estimating the cross-classified model (e.g., one model for students within schools, one model for students within neighborhoods). Although the variance components estimated by the traditional multilevel models are biased, they provide useful evidence of whether a cross-classified model could be estimated. Subsequently, the unconditional cross-classified model is estimated to explicitly partition the sources of variability at all levels of analysis, which is where the discussion turns next.

The Unconditional Cross-Classified Model

Consider a continuous outcome such as mathematics achievement, obtained for each student using a nonhierarchical, multilevel sampling design in which students at Level 1 belong to one unique combination of a given school and neighborhood at Level 2. For these data, the unconditional (i.e., no predictors) cross-classified model is:

$$\mathrm{Math}_{i,s,n} = \gamma_{0,0,0} + u_{0,s} + u_{0,n} + u_{0,sn} + e_{i,s,n},$$

$$u_{0,s} \sim N_0, \sigma_{u0,s2},$$

$$u_{0,n} \sim N_{0,} \sigma_{u0,n2},$$

$$u_{0,sn} \sim N_0, \sigma_{u0,sn2}, \text{and}$$

$$e_{i,s,n} \sim N_0, \sigma_{es,n2}.$$

Here, $\mathrm{Math}_{i,s,n}$ is the observed math achievement for student *i* who attends school *s* and lives in neighborhood *n*; commas are used to separate subscripts to ensure double-digit numbers are indicated clearly (e.g., the math score for student 10 in school and neighborhood 1: $\mathrm{Math}_{10,1,1}$). $\gamma_{0,0,0}$ is the fixed intercept representing the average (or grand mean) math achievement across all students, whereas $u_{0,s}$, $u_{0,n}$, and $u_{0,sn}$ are the random effects for school *s*, neighborhood *n*, and the school-by-neighborhood interaction *sn*, respectively. Given that the random interaction effect is included in this model, random effects $u_{0,s}$ and $u_{0,n}$ represent the school-and neighborhood-specific deviations from the average math achievement of all students, respectively, and these random effects are considered main (or marginal) effects that are averaged across the other Level 2 classification (i.e., $u_{0,s}$ is averaged across neighborhoods and $u_{0,n}$ is averaged across schools). The random interaction effect $u_{0,sn}$ indicates the deviation of average math achievement for a unique combination of school *s* and neighborhood *n* from the math achievement that would be predicted by the fixed intercept $\gamma_{0,0,0}$ and random main effects $u_{0,s}$ and $u_{0,n}$. All random effects are assumed independent (i.e., uncorrelated) and normally distributed with a mean of 0 with variances σu_0,s_2, σu_0,n_2, and σu_0,sn_2 for the school, neighborhood, and interaction random effects, respectively; a parenthetical subscript indicates the specific Level 2 classification is held constant (e.g., σu_0,s_2 is the random effect variance based on all *s* schools). Finally, $e_{i,s,n}$ is the residual value representing the deviation of math achievement for student *i* from the average math achievement of the student's unique combination of school *s* and neighborhood *n*. The residual values are assumed independent of the random effects and normally distributed with a mean of 0 and variance $\sigma_{es,n2}$.

As previously stated, the primary purpose of estimating the unconditional cross-classified model is to explicitly partition the sources of variability at all levels of analysis. Once partitioned, intraclass correlations can be calculated to obtain the proportion of variance attributable to each source of variation. In general, these intraclass correlations are calculated as the variance component(s) for the Level 2 classification(s) of interest relative to the total variability in the outcome. For example, the intraclass correlation of math achievement between two students who attend different schools but live in the same neighborhood is:

$$\rho_{s,n,sn} = \text{between school variabilitytotal variablity}$$

$$= \sigma u0, s2\sigma u0, s2 + \sigma u0, n2 + \sigma u0, sn2 + \sigma es, n2.$$

A Note on the Random Interaction Effect

The ability to include the random interaction effect $u_{0,\mathrm{sn}}$ is a direct result of having a cross-classification of schools and neighborhoods at Level 2; this effect cannot exist when sampling dimensions are clearly hierarchical and cannot be estimated using the traditional multilevel model. If the random interaction effect is detectable and significant, it allows students who attend the same school to be influenced by their neighborhood and students who live in the same neighborhood to be influenced by their school.

With that said, the ability to actually estimate the random interaction effect in the cross-classified model is dictated primarily by the number of students who have a specific combination of a school and a neighborhood, known more generally as the number of *within-cell replicates* or *within-cell sample size*. Sampling designs with numerous within-cell sample sizes ≤1 (e.g., only one student in a specific school lives in a specific neighborhood) will fail to estimate the random interaction effect (and its variance component $\sigma u_{0,sn2}$). For example, a public school district may give students the option to attend any high school of their choosing, but even in a large city, it is highly probable that no school will have multiple students from every neighborhood under study. Indeed, small within-cell sample sizes are common in applied educational research and are one of the primary reasons (in addition to unfamiliarity) for why the random interaction effect is estimated infrequently in the literature. However, erroneously omitting a random interaction effect that could have been estimated successfully will result in significantly biased variance component estimates (particularly for the variances of Level 2 classifications), which will in turn result in incorrect statistical inference for predictor effects.

The Conditional Cross-Classified Model

Following the estimation of the unconditional cross-classified model, predictor variables are typically included to explain each source of variability. Including predictor variables in the cross-classified model is similar to the traditional multilevel model with some added complexity, given there are multiple sources of variation at a given level of analysis.

Continuing with the aforementioned example, predicting a student’s math achievement using school-and/or neighborhood-level predictors at Level 2 is fairly straightforward. For example, a school’s sector (i.e., public *vs*. private) could be included to explicitly explain random school variance $\sigma_{u0,s2}$, whereas a neighborhood’s poverty rate could be included to explain random neighborhood variance $\sigma_{u0,n2}$. If the random interaction effect $u_{0,sn}$ was estimated and significant, then an interaction effect between a school’s sector and a neighborhood’s poverty rate could be included to explain random interaction variance $\sigma_{u0,sn2}$. However, if the random interaction effect was not estimable or nonsignificant, it is unlikely that the sector-by-poverty rate interaction would be detected statistically.

The inclusion of student-level predictors at Level 1 is much more complex because student-level predictors also contain school-and neighborhood-level variability at Level 2. Therefore, including a student-level predictor in a cross-classified model without considering its multiple sources of variability will produce an incorrect estimate for the effect of the student-level predictor, termed a *convergence effect*, that assumes the predictor has an equal effect on all sources of variability. The convergence assumption is explicitly testable. For example, consider using students' psychometric IQ to predict math achievement. To test convergence, the student IQ predictor at Level 1 would be included alongside two newly calculated Level 2 predictors—one representing the average IQ of the students attending each school and the other representing the average IQ of the students living in each neighborhood. When all three IQ effects are included in the conditional model, the school-average IQ effect and neighborhood-average IQ effect are both *contextual effects* that provide an explicit test of convergence. A significant contextual effect indicates the effect of IQ differs between the student, school, and neighborhood, and convergence should therefore not be assumed. When retained in the model, the contextual effects also indicate that after controlling for a student's IQ, there is an additional effect on a student's math achievement score from attending a more intelligent school and living in a more intelligent neighborhood.

Finally, additional random effects for any student-level predictor could also be included in the cross-classified model. For example, the effect of student IQ could vary randomly across different schools and different neighborhoods. In this case, a separate random student IQ variance is calculated across schools and across neighborhoods, with the random student IQ variance across schools predicted by the student IQ-by-school sector interaction and the random student IQ variance across neighborhoods predicted by the student IQ-by-neighborhood poverty rate interaction. Although estimating additional random effects may make sense theoretically, they are typically difficult to estimate in practice given the increased computational demands.

Available Software to Estimate the Cross-Classified Model

The cross-classified model can be estimated using frequentist methods (e.g., restricted maximum likelihood) in SAS, R, SPSS, Stata, HLM, and MLwiN. In addition, Mplus will estimate the cross-classified model as a latent variable model, although this option may be esoteric to those trained primarily in multilevel modeling. Increasing the complexity of the cross-classified model (e.g., introducing additional sampling dimensions and random effects) will quickly expose the computational ceiling of frequentist-based estimation. Thus, Bayesian methods (e.g., Markov chain Monte Carlo) are often employed by

Bayesian methods (e.g., Markov chain Monte Carlo) are often employed because they are more computationally efficient for more complex models. Of the software mentioned, only SAS, R, and MLwiN can be used to estimate the cross-classified model in a Bayesian framework. In addition, individuals with a considerable computational computing background could use SAS/IML or R to create a custom Markov chain Monte Carlo estimator that is specific to their needs.

Ryan W. Walters

See also Bayesian Statistics; Generalizability Theory; Hierarchical Linear Modeling; Markov Chain Monte Carlo Methods; Mixed Model Analysis of Variance

Further Readings

Goldstein, H. (2011). Multilevel statistical models (4th ed.). London, UK: Wiley.

Hox, J. J. (2010). Multilevel analysis: Techniques and applications (2nd ed.). New York, NY: Routledge.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. Journal of Educational and Behavioral Statistics, 18, 321–349.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. British Journal of Mathematical and Statistical Psychology, 63(1), 1–15.

Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). London, UK: Sage.

Wilfridah Mucherah Wilfridah Mucherah Mucherah, Wilfridah

Cross-Cultural Research

Cross-cultural research

435

437

Cross-Cultural Research

Cross-cultural research can be defined as an attempt to compare the use of an intervention or a practice in one culture with a similar intervention or practice in another culture. It also provides a context for the comparative understanding of concepts such as intelligence and motivation across cultures. Cross-cultural research can allow researchers to explore individual differences and diversity issues at a deeper level. This entry discusses the types of research questions that can be answered through cross-cultural research and describes how this research is typically approached.

Cross-cultural research can identify practices that are unique as well as those that are similar across different countries and contexts. Such research can show how learning practices vary in schools and indicate what cultural aspects shape these differences. Cross-cultural research permits the study and comparison of outcomes and conditions for success in different contexts. In education, for instance, numerous research studies have been conducted highlighting the role of parental involvement, teacher training, and school funding in student achievement. However, many of these studies have been conducted in the United States and have not been replicated in other countries. Cross-cultural research would allow researchers to answer questions about how these factors influence student achievement in various contexts around the world.

In attempting to determine to what extent parental involvement influences student achievement worldwide, researchers would look at multiple factors, such as whether there are contextual variables that contribute to the nature of parental involvement in some cultures that are not present in others, what the parent–teacher relationship looks like in different cultures, and how dealing with teacher relationship looks like in different cultures, and how dealing with parental involvement and student achievement manifests itself as an issue in different contexts. Cross-cultural research can enhance understanding of the way certain issues are rooted in schools and yet mediated by the social norms within a culture.

Cross-cultural studies have compared the experience of teaching in different countries, for instance, comparing the learning goals for secondary school students in Canada and Japan and the relative emphasis placed by teachers in the two countries on aspects of critical thinking, problem solving, developing relationships, and taking on responsibility. Teacher narratives have addressed issues such as differences in gender roles in different cultures, and biases students may have about teachers from different countries. Well-designed cross-cultural research can be instrumental in school settings with a large percentage of immigrant students, and findings can be of help to teachers and counselors in building cross-cultural bridges between school and home cultures.

Cross-cultural research typically uses an ethnographic approach. In ethnography, researchers usually take an emic perspective; that is, they study a culture from the perspective of the research participants themselves. This may involve researchers participating in activities of a culture or taking on more of a bystander role. Researchers may also interview or survey participants and collect artifacts or documents as a way of validating the data gathered through observations and controlling for their own biases and assumptions.

Wilfridah Mucherah

See also Cultural Competence; Culturally Responsive Evaluation; Ethnography; Multicultural Validity

Further Readings

Banks, J. A. (2006). Cultural diversity and education: Foundations, curriculum, and teaching (5th ed.). Boston, MA: Allyn … Bacon.

Canfield-Davis, K., Tenuto, P., Jain, S., & McMurtry, J. (2011). Professional ethical obligations for multicultural education and implications for educators. Academy of Educational Leadership Journal, 15(1), 95.

Li, D., Bai, X., & Li, H. (2014). On cultivating senior middle school students' cross-cultural awareness in English classes. Studies in Literature and Language, 9(1), 134–139. doi:10.3968/5437

McAllister, G., & Irvine, J. J. (2000). Cross cultural competency and multicultural teacher education. Review of Educational Research, 70(1), 3–24. doi:10.2307/1170592

McDonald, J. K., Goh, M., Brissett, A. A., Yoon, E., & Wahl, K. H. (2007). Working with immigrant students in schools: The role of school counselors in building cross-cultural bridges. Journal of Multicultural Counseling and Development, 35(2), 66–79.

Siu, F. W., Brodwin, M. G., Huang, I., Brodwin, E. R., & Kier, C. (2014). International collaborative cross-cultural teaching project: United States and Taiwan. Journal of Applied Rehabilitation Counseling, 45(2), 39–45.

VanTassel-Baska, J. (2013). The world of cross-cultural research: Insights for gifted education. Journal for the Education of the Gifted, 36(1), 6–18. doi:10.1177/0162353212471451

Cindy Suurd Ralph Cindy Suurd Ralph Ralph, Cindy Suurd

Leandre R. Fabrigar Leandre R. Fabrigar Fabrigar, Leandre R.

Crossover Design

Crossover design

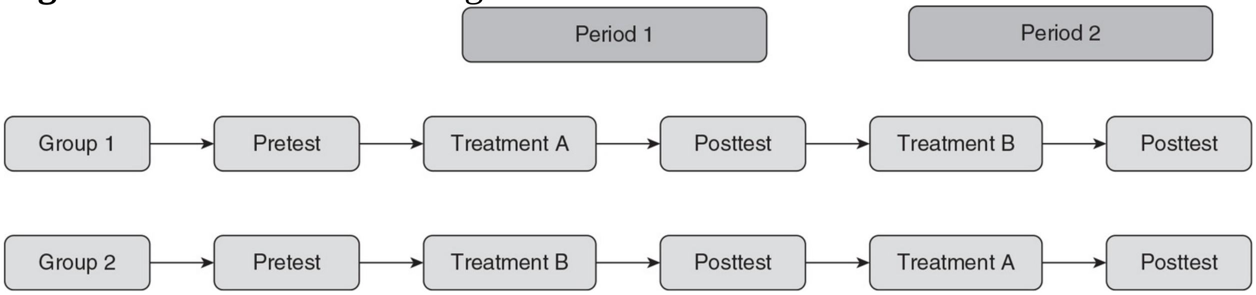
437

438

Crossover Design

The crossover design (also referred to as a replicated Latin square design) refers to a longitudinal study in which participants receive a sequence of treatments that varies based on the group to which the individual is assigned. The groups may be randomly assigned in the case of an experiment or allocated based on some other criteria (e.g., geographic location, classroom) in the case of a nonexperimental study. In its simplest form, the crossover design involves two periods by two treatments. Figure 1 depicts a basic 2 × 2 crossover design with two treatment sequences: AB and BA. As depicted in the diagram, all participants undergo a pretest at the commencement of the study. Then, in the first period of the study, one group of participants receives Treatment A while the other receives Treatment B. At the completion of the first treatment period, participants are administered a posttest. Groups then “crossover,” in the second period, so that individuals who started with Treatment A commence Treatment B and those who began with Treatment B undergo Treatment A. Another posttest is conducted after the second period.

Figure 1 Basic crossover design



More complex variations of the crossover design can include more than two treatments or groups or could involve the use of a treatment in more than one period. The number of waves (or treatment periods) can vary considerably based on the nature of the research question and the length of each treatment. Depending on the design selected, researchers can apply a variety of statistical analyses to determine the impact of period, sequence, carryover, and treatment effects. Designs vary to the extent that they are balanced (i.e., whether or not each treatment is preceded by every other treatment the same number of times) and uniform (i.e., whether each period allocates each treatment to the same number of subjects or whether each subject receives every treatment the same number of times). An example of a balanced uniform crossover design (4 periods × 4 groups) would produce four treatment sequences: ABBA, BAAB, AABB, and BBAA.

Crossover designs have been employed in a wide variety of settings ranging from educational to epidemiological research. The crossover design is primarily used to allow researchers to compare the efficacy or impact of multiple treatments or a combination of treatments and control or placebo conditions. Crossover designs are most appropriate for interventions that are considered temporary, so that multiple treatment options can be tested. Depending on the nature of the treatments, there may be a washout interval between periods to allow the effects of one treatment to wear off before undergoing the next treatment. Clinical trials in medical or pharmaceutical research frequently employ crossover designs to assess the efficacy of different types of medications on a disease or condition. In the social sciences, crossover designs can be used to examine a wide variety of research questions such as the impact of different clinical rotation sequences on examination performance or the effectiveness of different psychotherapy approaches on depression. The remainder of this entry discusses the strengths and limitations of the crossover design.

Strengths

A major strength of the crossover design is that it is within subjects, that is, each participant acts as his or her own control, which helps to account for the impact of individual differences and permits a smaller sample size than comparable research designs (e.g., parallel groups). Crossover and parallel groups designs can address similar research questions; however, the latter does so between subjects using a single treatment period, which requires a larger sample size. Another strength of the crossover design is that it provides control over the

Another strength of the crossover design is that it provides control over the temporal order of treatments or interventions. The inclusion of multiple combinations of treatments allows for a comprehensive examination of the impact of different sequences.

Limitations

One of the major concerns associated with the crossover design is that the effect of an earlier treatment can produce carryover effects into subsequent periods. These effects can be the result of learning, demand characteristics, or psychological or physical reactions to treatments that persist into subsequent periods. For example, an inadequate washout period between treatments could result in the residual effects of a pharmaceutical treatment from one period impacting the next treatment. Similarly, the psychological state of the participants may systematically differ at the commencement of later periods, if, for example, an aspect of one period induces more psychological stress than another period. An example of learning could occur if a group is taught a particular skill during a treatment period, which they could continue to use in subsequent periods. Regardless of their source, carryover or residual effects from earlier treatments may not be equivalent for the different groups, which could result in researchers being unable to determine the independent impact of each treatment period.

Several options have been put forward to account for these concerns. First, researchers are encouraged to design their studies in a way that allows for interactions between treatments and periods to be detected as well as establishes whether or not the interaction was caused by carryover. This typically requires the use of more complex designs, for example, using the balanced uniform crossover design depicted earlier would provide the researcher with the ability to statistically identify and account for the impact of carryover and sequence effects. However, researchers typically need to balance the risk of carryover and sequence effects against the negative impact of increased study duration (e.g., added cost, higher risk of attrition). In circumstances where it is not feasible to conduct a complex design and carryover effects are suspected, researchers may need to restrict their conclusions to the first treatment period.

Cindy Suurd Ralph and Leandre R. Fabrigar

See also Pretest–Posttest Designs

Further Readings

Jones, B., & Kenward, M. G. (2014). Design and analysis of crossover trials. Boca Raton, FL: CRC Press.

Raghavarao, D., & Padgett, L. (2014). Repeated measurements and crossover designs. Hoboken, NJ: Wiley.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton, Mifflin and Company.

Jennifer C. Greene Jennifer C. Greene Greene, Jennifer C.

Cultural Competence

Cultural competence

438

440

Cultural Competence

Cultural competence in educational inquiry refers to a particular stance or sensibility regarding the cultural and sociodemographic diversity of most American (and many international) educational contexts. This stance is one of proactive awareness of and assumed respect for cultural and sociodemographic diversity. This entry first discusses why cultural competence is important in educational research and how it is reflected in each of the steps of doing research. It then discusses an international study on resilience in children and youth and the ways in which the study exemplified a stance of cultural competence.

A stance of cultural competence assumes that the cultural and sociodemographic characteristics of the students, teachers, and other educators being studied—and of the contexts they inhabit—matter to the quality and the equity of the teaching and learning that takes place in those contexts. Cultural competence in educational inquiry is marked by engagement with difference throughout a study or assessment, from conceptualization to reporting. The stance deliberately includes data gathering activities and analyses of the data that directly engage and inform the particular character of the diversity present in the contexts being studied. This understanding of cultural competence makes it more of a sensibility than an acquired body of knowledge and skills about the specific cultures and other markers of demographic diversity that may be present in an educational inquiry context.

Importance of Cultural Competence in Educational Inquiry

Inquiry

The initial decades of the 21st century have borne witness to the deep challenges of cultural, economic, religious, and geographic diversity around the globe. Many developed countries have become havens for migrants from less developed countries seeking sanctuary from political unrest and violence. Educational institutions, historically and today, remain at the forefront of meaningfully and sustainably integrating diverse peoples into existing societies.

Educational institutions provide safe spaces for engaging with diversity. They further enable the next generations to learn about and from their differences and to develop their own sense of respect for and acceptance of other people and cultures. Therefore, educational research, evaluation, and assessment—in the United States and other countries with diverse populations—are themselves compelled to proactively and meaningfully engage with diversity and difference through the lenses of cultural competence.

Cultural Competence in the Process of Educational Inquiry

A proactive stance of cultural awareness, respect, and engagement will permeate all aspects of an educational inquiry. Most significantly, cultural engagement will be reflected in the inquiry questions asked, the definitions and assessments of key constructs, the processes used to gather and analyze data, and the format and distribution of study reports. Each requires respectful attention to the strands of diversity present in the contexts being studied. For each, the inquirer can ask, “How well does this component of my study, or my assessment process, manifest respect for diversity and meaningful engagement with difference?”

For example, researchers might want to study a web-based mathematics program for middle school students that features visual demonstrations of core mathematical concepts so they can determine how the program’s graphics contribute to student learning. Attending to culture, specific study questions could include, “How well do students—who have differing mathematical backgrounds (both quantitatively and qualitatively) and varying levels and kinds of computer experience—develop proficiency in using the program and demonstrate mastery in learning, and how do the program’s graphics contribute to these varied learning pathways?” This question requires differentiated assessments of student engagement and learning, differentiated analyses of data, and reporting that is inclusive of all student participants. A culturally competent study design and instrumentation would respectfully track the multiple pathways of engagement and learning in the mathematics program that are demonstrated by students with varying backgrounds and experiences.

Educational assessment activities can be particularly sensitive to cultural and sociodemographic differences, beginning with the definitions of the constructs being measured. These definitions require cultural respect, possibly invoking thoughtful modification of existing assessment instruments and/or the development of study-and context-specific assessments.

Further, in educational evaluation studies, of central importance to a culturally competent and respectful stance is the specification of the criteria developed to make judgments of program quality and effectiveness. Such criteria, which constitute the definitions of a “good” or “effective” program, require development or selection with a culturally respectful and engaged lens.

Cultural Competence in Studying Resilience Across Cultures and Contexts

The construct of *resilience* is often located in theories of child and youth development, with particular relevance to children and youth who experience adversities, such as family disruptions, persistent poverty, or insufficient love or care. The resilience of children and youth is thus a topic of direct relevance to educational theory, inquiry, and practice.

In the first decade of the 21st century, Michael Ungar and Linda Liebenberg led an international team in an iterative, multiphased study designed to develop a culturally respectful measure of child and youth resilience. The researchers in this assessment study conceptualized resilience not only just as a characteristic of the individual but also as the individual’s responses to environmental affordances of psychosocial resources. Resilience was specifically defined as “doing well despite adversity,” thus encompassing cultural variation in these environmental resources, with specific relevance to the less developed world.

In addition to this explicitly culturally engaged conceptualization of the study’s key construct, this research also evidenced cultural competence in its design and implementation processes. The iterative design was implemented as a conversation between the research team and members of local settings in conversation between the research team and members of local settings in multiple countries. Research team instrument drafts included data from youth and youth worker interviews in local sites.

Further youth interviews accompanied pilot testing of the instrument, invoking a dialogue between resilience as defined on the instrument and resilience as experienced by youth in diverse contexts. The final data analyses were initially performed using Urie Bronfenbrenner’s ecological model of development. When this model did not fit the data, the quantitative analyses were reframed using seven key dimensions of resilience derived from extensive interview analyses. The final analysis developed portraits of youth resilience by gender, country, and the degree of social cohesion present in the contexts the youth inhabited.

Jennifer C. Greene

See also Culturally Responsive Evaluation; Joint Committee on Standards for Educational Evaluation; Minority Issues in Testing; Multicultural Validity

Further Readings

American Evaluation Association Statement on Cultural Competence in Evaluation. Retrieved from http://www.eval.org/p/cm/ld/fid=92

Hood, S., Hopson, R., & Frierson, H. (Eds.). (2015). Continuing the journey to reposition culture and cultural context in evaluation theory and practice. Charlotte, NC: Information Age Publishing.

Kirkhart, K. (2010). Eyes on the prize: Multicultural validity and evaluation theory. American Journal of Evaluation, 31(3), 400–413.

Ungar, M., & Liebenberg, L. (2011). Assessing resilience across cultures using mixed methods: Construction of the child and youth resilience measure. Journal of Mixed Methods Research, 5(2), 126–149.

Dominica McBride Dominica McBride McBride, Dominica

Culturally Responsive Evaluation Culturally responsive evaluation

440

444

Culturally Responsive Evaluation

Culturally responsive evaluation (CRE) is an evaluation approach that places culture and the community of focus at the center of the evaluation, helps to support community empowerment, and has a goal of social justice. This entry explains the reasons for the development of CRE, describes its components, details the process used to conduct CRE and how it contributes to social change, and gives an example of CRE in practice.

Historically, culture has been viewed as “noise” in evaluation—a confounding variable and a subjective factor to be controlled for or discounted. Similarly, the context of an evaluand was considered but treated as separate or distinct. Often, evaluators and program developers indicated that a program was supposed to have the same results regardless of culture and context.

The development of policies and programs has been dominated by people who are not part of the culture and context where those programs have their impact and this disregard and/or ignorance has contributed to perpetuating injustices. Programs have been eliminated because evaluators did not perceive the need for them or were looking at outcomes that were not relevant for the group benefiting from the program. In other cases, context was not considered in decision making around allocation of resources. Some continue to receive programs that may respond to surface needs but fail to solve underlying problems.

CRE has contributed to the field’s recognition of the profound influence of culture and context on the evaluand and its intended beneficiaries. It not only considers context but also uses it as data to understand the evaluand and the participants and a compass to direct the program in ways that lead to justice for that community. Cultural context also provides a way to understand the evaluand that reveals how participants experience it and why. It then builds the understanding necessary to identify what justice would mean to and for that community.

CRE Components

There are four main components to CRE: (1) culture, (2) context, (3) responsiveness, and (4) a commitment to social justice.

Culture

Culture is the shared norms and underlying belief system of a group as manifested and guided by its values, rituals, practices, language, institutions, and artifacts. Culture creates and identifies meaning, delineates values and guides how they are turned into action, and shapes the practices and behaviors of a group. For example, one group may value individualism and another collectivism. These values may be applied through individualistic learning rituals or group-oriented education practices.

Culture influences our perception and, thus, how we perform. It also affects what we view as the perceived totality of options for our behavior. A person who values collectivism may benefit and do quite well in a group-oriented and community-focused program and a person who believes intelligence is inherent may see the first sign of difficulty in answering a challenging question in school as a lack of intelligence and, consequently, quit. Therefore, the program strategy may need to change depending on the cultural belief system. The evaluator must consider these dynamics to assist in achieving programmatic goals.

To help ensure cultural sensitivity, the community or beneficiaries of the evaluand are engaged from the start. The people are at the center of CRE. They are the experts in their own experience and belief system. They know what they value, what they need, what practices they engage in, and how best to address and respond to each factor.

Context

There are many cultural realities that constitute the cultural context of an evaluand. These various layers of context include historical, sociopolitical, community, and organizational levels. Historical context includes what has taken place around the evaluand and intended participants over time. History tells the evaluator about the origins and subsequent changes in the evaluand; it can provide deeper understanding of participant needs, past experiences of the community, and roots of the problem.

The sociopolitical context can influence the evaluand in many ways—from decisions around funding allocations to what types of programs are provided to who implements those programs and how. The community context includes the local community's resources and current, collective experiences that are often directed by cultural perspectives. It may include socioeconomic status, collective assets, possible trauma, and available resources. Finally, organizational context includes more localized facets such as who is facilitating the program and how and where the program is implemented. The intersection of all of these constitutes the current cultural context and is, thus, interdependent.

For example, a city may have had a history of financial misconduct. Education was not a priority in the sociopolitical environment and, therefore, has not received sufficient funding to provide the programming necessary for students to meet the statewide goals for reading and math test scores. The community of focus has experienced high rates of school dropout, unemployment, and violent crime. The violence adds additional stress for the students and interferes in their academic performance. As a result of the funding crisis, the city school board decides to lay off more experienced teachers and hire new and young teachers, many of whom do not know or respond to the culture and context of the students. This combination affects if and how certain programs are implemented and how the students perform.

Responsiveness

Responsiveness encompasses a sense of critical consciousness, intentional action, and flexibility. Ostensibly, these three aspects are guided by values including community, relationships, neutralizing power dynamics, social justice, and critical consciousness. CRE evaluators are, therefore, critically self-reflective and strive to hone their social and emotional intelligence in ensuring sound partnerships and data collection efforts. CRE allows for the necessarily organic-and human-centered structure of the program and evaluation to emerge.

Responsive evaluation places stakeholder engagement, relationships, and dialogue at its center. It includes stakeholders throughout the evaluation process and attends to their issues and needs. It is a democratic, empowerment-focused model where the stakeholders dictate the standards and describe the program practices and meaning from their perspective.

The main goal of responsive evaluation is program improvement, which happens through dialogue and relationship building between the different stakeholders. With a reflective mind-set, the evaluator facilitates these dialogues and working alliances. Ideally, those who have a stake in the evaluand will eventually take ownership and make changes through listening and responding to other vantage points and experiences.

The culturally responsive evaluator also learns and integrates the culture of the community for accurate evaluation. The evaluation questions, data collection tools, interpretations and analysis, and influence of dissemination can all be compromised if culture is ignored. Cultural knowledge and integration increases the validity of the tools, data, and analysis as well as the effect of the evaluation as a whole.

In CRE, the evaluation is a conduit not only for relationship building but also for cultural understanding and social justice. Here, the evaluator is both responsive to the stakeholders, particularly the beneficiary community, culture, and context, and responsible for ensuring the results that benefit the community. Furthermore, the relationship building, cultural responsiveness, and resultant data collection and analysis feed into and support empowerment.

Social Justice

The development of CRE has been guided by a desire to support oppressed and marginalized communities. Thus, the evaluation is a tool for social justice, partnering with the communities the evaluand serves to advocate for and with them. The process of the evaluation, from community partnership to advocacy, is itself an intervention, as the evaluand’s beneficiaries and community members hold power positions within the evaluation. They craft evaluation questions, inform tool development, and have a say as to whether the evaluation is sound. Ideally, they also partner with the evaluator in advocating for changes, be it regarding the evaluand or sociopolitical context, for their community.

CRE Process

CRE Process

CRE has a prescriptive process in order to achieve its goals. Stafford Hood, Pamela Anderson-Frazier, and Rodney Hopson describe a detailed sequence of steps in conducting CRE. The following is a condensed version of the action steps:

Learn the culture and environment

This step includes both formal and informal assessments, such as spending time in and with the community and conducting semistructured interviews and focus groups, as well as gathering secondary data on the historical, sociopolitical, and cultural contexts.

Engage the people in the process

As with all evaluations, there are a variety of stakeholders who should be considered. However, CRE prioritizes the program beneficiaries and their community.

Develop culturally relevant evaluation design and tools

Here, culture and context are a core part of the evaluation, so the questions and tools consider, integrate, and reflect them.

Conduct the evaluation with the community

Community members are equal partners in the evaluation process, informing evaluation questions, tools, and analysis. Ideally, the evaluator also provides technical support, so members can participate as leaders throughout the evaluation.

Disseminate and advocate

Market lessons learned to foster social justice and a thriving community. The results of the evaluation not only contribute to program improvement but they also help to further the community as a whole.

CRE and Social Change

CRE and Social Change

CRE is not only an approach to evaluation, it is also a tool for achieving social change, particularly with and for disenfranchised communities. Both within the process and as a product, the community is supported and encouraged to manifest its empowerment. This focus and partnership helps to enhance validity and provide data for program improvement tailored to that community and sustainable change through advocacy. Thus, it also involves strategy development both on the levels of program improvement and on the broader social change. Although this type of evaluation often requires additional resources, it is designed to yield more significant results for those who are frequently placed in the margins of society.

CRE in Practice

One example took place in Chicago, IL, with a predominantly low-income Latino community on the west side of the city. The evaluator was contracted to conduct a needs assessment of youth resources, such as after-school youth development programs focusing on STEM, sports, cooking, or the arts. The evaluator and team conducted an analysis of the context, including demographics, resources, sociopolitical issues, and community hardships. With the partner organization, they engaged 20 local parents in shaping the needs assessment process and crafting the assessment tools. They also trained the parents in data collection, and the parents applied their new skills and collected data from 1,500 people in the community. The evaluator, partner organization, and parents then came together and analyzed the data. They found, for example, that over half (60%) of youth whose families were surveyed were not involved in after-school programming and 80% of parents expressed a need for more youth programming in the community. The parents developed recommendations. The evaluator synthesized these in the report, and the parents used it as a tool for advocating for what they needed in the community.

From the beginning of the evaluation, the evaluator had to consider language, the vulnerability some community members felt around assessment, and the fear that some members had around their immigration status. Also, a significant cultural asset was the strong sense of community and service of the parents. The survey questions, focus group protocol, and overarching process were created with the consideration of language, sensitivity around possible fears, and knowledge of cultural assets.

In bringing the parents together for this assessment, there was more power and potential for broader community engagement, which was needed to galvanize additional members and advocate for filling the gaps in services discovered through the process. This partnership and responsiveness not only helped to ensure valid data were collected more easily and from more people but also the results could be used to meet needs and support community empowerment.

Dominica McBride

See also Advocacy in Evaluation; Cross-Cultural Research; Cultural Competence; Democratic Evaluation; Multicultural Validity; Social Justice; Transformative Paradigm; Values

Further Readings

Amba, T. A. (2006). The practices and politics of responsive evaluation. American Journal of Evaluation, 27, 31–43.

Frazier-Anderson, P., Hood, S., & Hopson, R. (2011). Preliminary considerations of an African American culturally responsive evaluation system. In S. D. Lapan, M. T. Quartaroli, & F. J. Riemer (Eds.), Qualitative research: An introduction to methods and designs (pp. 347–372). San Francisco, CA: Jossey-Bass.

Greene, J. C. (2005). Evaluators as stewards of the public good. In S. Hood, R. Hopson, & H. Freirson (Eds.), The role of culture and context: A mandate for inclusion, the discovery of truth, and understanding in evaluative theory and practice (pp. 7–20). Greenwich, CT: Information Age Publishing.

House, E. (1990). Methodology and justice. New Directions for Evaluation, 45, 23–36.

Kirkhart, K. (1995). Seeking multicultural validity: A postcard from the road. Evaluation Practice, 16, 1–12.

McBride, D. F. (2015). Cultural reactivity vs cultural responsiveness:

Addressing macro issues starting with micro changes in evaluation. In S. Hood, R. Hopson, H. Frierson, & K. Obeidat (Eds.), Continuing the journey to reposition culture and cultural context in evaluation theory and practice (pp. 179–202). Charlotte, NC: Information Age Publishing.

Stake, R. (2004). Stake and responsive evaluation. In M. C. Alkin (Ed.), Evaluation roots: Tracing theorists’ views and influences (pp. 203–217). Thousand Oaks, CA: Sage.

Linda S. Behar-Horenstein Linda S. Behar-Horenstein Behar-Horenstein, Linda S.

Curriculum

Curriculum

Curriculum

The term *curriculum* is widely used among educators at all levels of education. Because of the many ways in which it has been defined, individuals may not be referring to the same concept when discussing curriculum. This entry discusses the common conceptions of curriculum and how these conceptions have changed in recent years.

Often curriculum is described as something that is planned and expected to be taught and learned. However, what is taught (actual vs. anticipated), how something is taught (the type of instructional strategies and medium used to implement what is to be taught), and the degree to which what is taught is actually learned, executed, accomplished, demonstrated, or observed and how it is evaluated are often questions left to evaluator inquiry or form the basis of researchers’ studies.

In the mid-20th century, common conceptions of curriculum were plan, system, field of study, experience, and content. Curriculum as a plan refers to what content or skills an educator anticipates teaching. Curriculum as a system refers to the people, processes, and organizational structures that guide planning, teaching, and measuring the taught content.

Curriculum as a field of study refers to the disciplinary emphasis of curriculum as a body of content in its own right to be mastered, which is guided by theory, principles, and practice. Curriculum as experience refers to what learners undergo in an educational system or organization either directly or indirectly as a result of what they are taught by the personnel who provide instructional and measurement activities. Curriculum as content refers to the subject or disciplinary matter and/or psychomotor or affective skills that are taught.

Owing to the influence of postmodernism, conceptions of curriculum that emerged in the later 20th and early 21 centuries were the null, hidden, and transformative curriculums. The null curriculum refers to what is not taught within the subject matter or content, such as particular viewpoints, historical events, or nuanced perspectives. The null curriculum in effect restricts the range of perspectives that are offered to students and results from the educational background of the instructor, the reigning political stance of the locale in which curriculum is taught, or that which is influenced by the preferences of the region, the community, or district where the individual school resides.

The hidden curriculum refers to the values and cultural norms that characterize the learning community, the types of interactions that are permitted or excluded within a particular course which in turn are inherently contextualized by location or time, content or material, and student and instructional members of a learning community that remain openly unacknowledged. Nonetheless, the hidden curriculum is believed to exhibit an influence on both curriculum and student outcomes. From another perspective, the hidden curriculum could be described as the nonverbal experiences that are felt by students because they are transmitted through action, though left unspoken. The hidden curriculum can be inferred by the lack of equal treatment and equality of educational opportunities for all students as well as the practice of tracking and unequal implementation of discipline policies that are in direct conflict with the belief that schools provide equality of opportunities for all of their students.

The transformative curriculum refers to instruction that invites students to question the truth capacity of what they are learning or encourages them to use newly learned information or skills to metamorphose their own thinking and actions.

What is actually taught and how it is taught is also influenced by prevalent societal norms and practices of a particular era in history. Curriculum can be conceived as residing on a continuum. It may be thought of as static and unchanging content such as the didactic, teacher-centered practice of teaching the classics. In contrast, a curriculum that is influenced by societal norms is perceived to be dynamic, fluid, and everchanging.

With the implementation of new forms of technology, including computers and

With the implementation of new forms of technology, including computers and other educational media, researchers such as Robert Kozma and Chris Dede have asserted the potential of technology to promote better achievement and to improve student attitudes toward schooling and learning in general. To be sure, there are wide variations in schools' educational technology, and the way it is used can depend upon the subject matter, learning objectives, teacher proficiency, and infrastructure support at school and district levels. Still, there is little doubt that technology has influenced perceptions of what and how curriculum is implemented.

New opportunities to access information through technology have expanded students' ability to learn outside the classroom, in turn changing the nature of what is seen as curriculum. Educational content is now widely and often freely available on the Internet. As a result, curriculum is no longer limited solely to the subject matter concepts or those activities that an instructor brings to the classroom, laboratory, or clinical-learning environment and plans to teach. Nonetheless, the selection of content that is perceived to be worth knowing is typically influenced by the dominant voices of society, including those who are subject matter experts, textbook publishers, and testing companies.

Linda S. Behar-Horenstein

See also Classroom Assessment; Common Core State Standards; Curriculum Mapping; Curriculum-Based Assessment; Curriculum-Based Measurement; High-Stakes Tests; Instructional Objectives; Instructional Theory; Literacy; State Standards

Further Readings

Dede, C. (1996). Emerging technologies and distributed learning. American Journal of Distance Education, 10(2), 4–36. doi:10.1.1.136.1029

Eisner, E. W. (1985). The educational imagination (p. 176). New York, NY: Macmillan.

Henderson, J. G., & Gornik, R. (2007). Transformative curriculum leadership. Englewood Cliffs, NJ: Prentice Hall.

Joughin, G. (2010). The hidden curriculum revisited: a critical review of research into the influence of summative assessment on learning. Assessment … Evaluation in Higher Education, 35(3), 335–345.

Kozma, R. B. (1991). Learning with media. Review of Educational Research, 61, 179–221. doi:10.3102/00346543061002179

Kozma, R. B. (1994). Will media influence learning? Reframing the debate. Educational Technology Research and Development, 42(2), 7–19.

Ornstein, A. C., & Hunkins, F. P. (2012). Curriculum: Foundations, principles, and issues (6th ed.). Boston, MA: Pearson.

Tonya Breymier Tonya Breymier Breymier, Tonya

Curriculum Mapping

Curriculum mapping

445

447

Curriculum Mapping

Curriculum mapping involves aligning specific course-and/or grade-level activities toward attainment of specific learning outcomes, which are basically what it is the instructor expects the student to *do*. Curriculum mapping is a process to initiate, review, and validate curriculum alignment. The process results in curriculum maps that provide the visual linkages between course-or grade-level activities and learning outcomes. Curriculum maps also serve as a method of communication among instructors across courses and/or grade levels within educational programs. This entry discusses the development of curriculum mapping, its purpose and benefits, and how mapping takes place.

The learning outcomes can be based on grade-level objectives, program competencies, or standards. Learning outcomes can also refer to learning objectives, ranging from assignment outcomes to course outcomes and course outcomes to program or level outcomes. Curriculum mapping stems from the 1980s work of Fenwick English which began as mere detailing of what instructors were teaching and how it was taught. In the 1990s, the work of Heidi Hayes Jacobs added more depth and breadth to the focus of curriculum mapping.

Curriculum mapping can occur at the course and/or program level. Various studies have examined the value of curriculum mapping, offering tools for effective curriculum mapping and describing/instructor preferences and perceptions of curriculum mapping, but there has been little published research on specific curriculum mapping processes.

Curriculum mapping provides a road map for curriculum planning to achieve previously identified skills, competencies, and/or learning outcomes. Curriculum mapping is a process used in both K–12 and higher education and can be used both within and across grade levels or specific courses. Mapping can involve lesson plans for individual classes or grade-by-grade programmatic planning for an entire school.

Benefits of curriculum mapping include providing short-and long-term goals to meet educational outcomes and identifying gaps and areas for improvement. Curriculum maps are tools to keep faculty focused and can prevent curriculum drift in addition to identifying and preventing curriculum repetition. The curriculum map can identify when concepts should be introduced, reinforced, and mastered within specific courses and/or levels. Curriculum mapping can identify placement of specific assignments, exams, and projects within specific courses, grades, and/or program levels. Attainment of specific learning outcomes can prepare the student for the subsequent course and/or grade level.

Sustained curriculum mapping efforts can improve faculty buy-in/participation and promote curriculum revision when faculty are provided with the resources for curriculum mapping. Curriculum mapping processes need to be outlined with a concrete plan for review and possible revision. Faculty development regarding curriculum mapping procedures and mapping tools, in addition to leadership support and faculty accountability, must be communicated and reinforced. Scheduled mapping discussions within courses, levels, and institutions or programs will ensure maintaining focus and curriculum alignment toward meeting learning outcomes. Mapping also can serve as a process to monitor what faculty do and as an avenue for data collection that can provide valuable assessment and evaluation data for course and/or program improvements.

Examples

The following examples show how curriculum mapping connects the courses at each level and the levels within a program or school.

A K–12 school district could use curriculum mapping to establish leveled competencies for each grade level. Administrators, curriculum personnel, and/or teachers would map the curriculum for each grade level. For each course, they would identify specific concepts to be introduced, reinforced, and mastered to meet specific course learning outcomes. Each course will denote specific assignments, exams, and activities. Attainment of specific course learning outcomes at each level would prepare the student for the next grade level.

Faculty and school leaders on each level would review the course maps, ensuring course concepts are taught as the map outlines and any gaps in the course curriculum are identified.

In another example, a school of nursing in an institution of higher education would establish program learning outcomes to meet national competencies for becoming a licensed registered nurse. The nursing program consists of three levels; each level is mapped to meet specific program learning outcomes. Each level contains courses with identified concepts to be introduced, reinforced, and mastered to meet specific course learning outcomes. Attainment of these specific course learning outcomes at each level would prepare the student for the next level. Completing all three levels meets the overall nursing program learning outcomes and prepares the student for the national competencies to successfully pass the licensed registered nurse exam. The school of nursing curriculum committee meets annually to examine and review the course maps as well as the overall program maps to ensure leveled concepts are taught as the map outlines and identifies any gaps in the program curriculum.

Tonya Breymier

See also Concept Mapping; Curriculum; Curriculum-Based Assessment; Curriculum-Based Measurement; Instructional Objectives; Learning Maps

Further Readings

Arafeh, S. (2015). Curriculum mapping in higher education: A case study and proposed content scope and sequence mapping tool. Journal of Further and Higher Education, 40(5), 585–611. Retrieved from http://dx.doi.org/10.1080/0309877X2014.1000278

Ervin, L., Carter, B., & Robinson, P. (2013). Curriculum mapping: Not as straightforward as it sounds. Journal of Vocational Education … Training, 65(3), 309–318. Retrieved from http://dx.doi.org/10.1080/18636820.2013.819559

Lam, B., & Tsui, K. (2013). Examining the alignment of subject learning outcomes and course curricula through curriculum mapping. Australian Journal of Teacher Education, 38(12). Retrieved from http://dx.doi.org/10.14221/ajte.2013v38n12.8

Shilling, T. (2013). Opportunities and challenges of curriculum mapping implementation in one school setting: Considerations for school leaders. Journal of Curriculum and Instruction, 7(2), 20–37. Retrieved from http://www.joci.ecu.edu

Spencer, D., Riddle, M., & Knewstubb, B. (2012). Curriculum mapping to embed graduate capabilities. Higher Education Research … Development, 31(2), 217–231. Retrieved from http://dx.doi.org/10.1080/07294360.2011.554387

Theodore J. Christ Theodore J. Christ Christ, Theodore J.

Danielle M. Becker Danielle M. Becker Becker, Danielle M.

Curriculum-Based Assessment Curriculum-based assessment

447

450

Curriculum-Based Assessment

Curriculum-based assessment (CBA) emerged in the 1970s and early 1980s as a novel approach to formative assessment and evaluation. This entry discusses the development of CBA, the two paradigms for CBA, and the four different methods of CBA that fall within those paradigms. CBA was developed for use by teachers to guide educational decisions related to the selection and use of curriculum materials and instructional procedures. Because of its foundation in relevant educational practice, CBA can be a highly useful tool in student evaluation and instructional decisions within a problem-solving framework.

CBA is a collection of assessment and evaluation techniques to test student performance using materials sampled from or based on the local curriculum. CBA was designed to be more instructionally relevant to educators than published assessments because of its high reliance on local curriculum and correspondence with students' daily classroom experiences.

CBA was developed to provide more relevant information to educators because many of the developers perceived that widely used published assessments were too generic to guide local decisions for students. Because these materials were sampled from the local curriculum, CBA was thought to provide the most relevant information to guide decisions about instruction and curriculum. At the time, it was assumed that the most authentic materials for use in assessment were those sampled from the local curriculum and learning environment.

CBA encompasses several assessment and evaluation procedures that use direct observation and other methods to measure student performance with alternate curricula and instructional procedures. There are two related, but distinct, methods to sample and construct CBA materials. First, subskill mastery measurement (SMM) is used to divide curriculum goals into short-term, discrete objectives that are assessed sequentially, often in a hierarchical manner. SMM employs mastery measurement, in which performance on one assessment is used to indicate proficiency within one or a few closely related academic domains. Because it assesses mastery of discrete objectives, SMM is useful to evaluate strengths and weaknesses across specific skills. It is also useful to monitor progress over brief periods of time to evaluate educational programs.

The second method is general outcome measurement (GOM). GOM is distinct from SMM in that it samples from the annual curriculum to assess global proficiency relative to achievement across the entire academic year. Although performance is expected to be very low early in the academic year, student performance on successive measurements should increase. In general, GOM assessment scores tend to be more predictive of performance on published norm-referenced assessments than SMM. GOM assessments also lend themselves to triannual screening and longer durations of progress monitoring, which span months rather than weeks.

There are several different types of CBA, representing the two assessment paradigms just discussed. Researchers identified at least four different methods of CBA: CBA for instructional design (CBA-ID), criterion-referenced CBA (CR-CBA), curriculum-based evaluation (CBE), and curriculum-based measurement (CBM). CBA-ID, CR-CBA, and CBE follow an SMM paradigm, while CBM follows a GOM paradigm. Each of these four models are described briefly in the next section.

CBM

CBM is a type of GOM that quantifies student performance in basic academic skill areas through standardized measurement procedures. It was designed to be a reliable, valid, simple, efficient, and inexpensive method for recording the level and rate of student achievement in reading, math, spelling, and written expression. CBM makes use of a GOM assessment paradigm to track progress through the annual curriculum using a series of equivalent assessments. Because it uses GOM, CBM uses measures that are standardized, valid, and reliable.

CBM assessments are designed to be *dynamic indicators* of academic skills, meaning that they assess change over time, are highly sensitive to short-term effects of instruction or intervention, and are correlated with key behaviors that suggest success in an academic domain. This suggests that CBM assessments have high utility in screening and progress monitoring.

CBM was originally designed as a method of curriculum sampling to create equivalent measures for use in screening and progress monitoring. Now, however, most educators use materials that are readily available through assessment companies. These assessments often have established difficulty levels, technical adequacy, and normative information. Reliability and validity of such assessments can be demonstrated more readily with the availability of commercial CBM materials, which give further evidence for the use of such assessments as dynamic indicators of student performance within an academic domain.

Of the four types of CBA, CBM has the most robust research base. First, CBM can be used as an indicator of basic skills development because there is significant reliability and validity evidence. This suggests, as described earlier, that CBM can be used to track progress in skill development. CBM has also been shown to be a reliable and valid method for differentiation between higher and lower performing students in terms of reading achievement, which, again, suggests that CBM reading assessments are useful for screening. To that end, there is significant research to support the use of CBM for academic screening. However, there is also evidence to support the use of CBM for progress monitoring, in terms of both growth in response to instruction and growth in response to intervention activities.

CBA-ID

CBA-ID uses student performance and responsiveness to instruction as a guide for instructional planning, with the goal of delivering instruction that is as efficient and effective as possible. It is used to determine whether instruction is compatible with student skills. Broadly, CBA-ID involves assessing student proficiency in a given content domain and using that information to assess instructional match or tailor instruction to individual student needs. CBA-ID is based on the rationale that student skill deficits are maintained and potentially caused by a mismatch between incoming student skill level and classroom instructional level. The purpose of CBA-ID is to address this mismatch.

CBA-ID makes use of an SMM assessment paradigm to sample student performance within discrete academic objectives. This assessment paradigm is used to measure curriculum mastery for each student, such that instruction is tailored to individuals. CBA-ID operates under the assumption that students learn best when instructional material is neither too difficult nor too easy. Instruction that is at the *instructional level* of a student is challenging enough that students have potential to learn and show clear progress. This is the target level of instruction for CBA-ID.

There is a fairly limited research basis for CBA-ID. Research that has been done, however, supports the idea that instruction that optimizes the ratio of unknown and known information in a given lesson improves engagement, learning rates, and retention.

CBA-ID operates under four basic principles. First, the purpose of CBA-ID is to match assessment with instruction. The rationale behind this principle is that assessment material that is most informative for teachers is material that most closely aligns with material used in the classroom. The classroom is a natural context for assessment in terms of both student learning and teacher instructional practices. If assessment and instruction are matched, valuable information about the effectiveness of both can be inferred.

Second, CBA-ID uses the student's level of knowledge to determine the student's specific areas of weakness. This principle emphasizes the idea that CBA-ID is thoroughly student centered; instruction is focused on filling in each student's knowledge gaps individually. Third, as discussed earlier, CBA-ID places a high focus on correcting the mismatch between instruction and student skill level. This is accomplished through a determination of appropriate instructional match through both level of challenge and rate of instruction.

Finally, CBA-ID operates under the assumption that students benefit from instruction that is appropriately matched to individual skill level, which is assessed through mastery learning. This model of instruction and assessment is in direct opposition to most models of instruction in which the same curricular material is presented to all students at the same rate.

There are four steps to implement CBA-ID. First, the examiner must identify the materials used in the classroom that will be the focus of the assessment. Second, the examiner must determine the student's specific skill deficit using the identified assessment materials. Third. the examiner must determine the necessary modifications to instruction or additional strategies, thereby creating a match between current student skills and instruction. Finally, these changes are implemented and appropriate instructional material is chosen. Student skill development is monitored for progress in mastery of objectives.

CR-CBA

CR-CBA is used to discover the curriculum materials and instructional procedures to optimize the educational program for each individual student. In contrast to CBA-ID, CR-CBA makes use of a mastery criterion. CR-CBA uses sequentially ordered SMM assessments developed from curriculum objectives to determine student skill level and instructional needs by comparing student performance to local normative information. Measures used vary widely, as they are generally created by classroom teachers, but the distinguishing feature of CR-CBA is the comparison of student performance to a normative reference group for interpreting student performance on the skills measured. This criterion is often considered the level of performance necessary for mastery of a given skill.

Implementation of CR-CBA is based on a hierarchical arrangement of skills drawn from a given curriculum and the sequential assessment of these skills. First, examiners must list the desired skills in order, ensuring that all the relevant skills are included and that the order makes intuitive sense. Next, the examiner should create discrete objectives within each skill, writing test items for each objective. The resulting assessment can be used as a pretest, posttest, or assessment of retention before or after instruction takes place.

After taking the assessment, results should be examined to determine the skills or objectives students have mastered and the skills or objectives that represent specific areas of difficulty. In this way, examiners are able to determine whether instruction is appropriately matched to student skill level and whether students have the necessary prerequisite skills for future instruction.

CBE

CBE employs a problem-solving framework in conjunction with repeated SMM assessments to evaluate student skill level and mastery of curricular objectives. It is a hypothesis-testing framework in which information about a student is repeatedly gathered and analyzed to determine where the student's instructional level is relative to the entire curriculum.

CBE can be described as a task analysis model in which a series of interconnected tasks represented curricular objectives. According to this model, a student's instructional placement is determined by the student's position in the maze of tasks; it is the examiner's goal to determine the placement. That is, the examiner must determine the tasks the students mastered and tasks they are ready to learn based on their current performance level.

There are four key steps in CBE. The first step is problem identification, whose purpose is to determine whether the students exhibit a skill deficit according to their current performance level and the expectation according to the curriculum. Problem identification is accomplished through both assessment activities and examination of existing data, such as school-wide screening scores. Both formal and informal assessments can be included in assessment activities.

Second, information gathered through problem identification is analyzed to develop hypotheses about the problem and explain the cause of the observed skill deficit. These hypotheses provide direction for assessment of specific skills. Third, hypotheses are tested with SMM assessments to determine whether they can explain observed skill deficits. If the hypotheses are incorrect, new hypotheses are created and tested in the same way. Finally, hypotheses that are found to be correct are used to inform instructional changes, with the goal of improving the student's observed skill deficit. Student progress is monitored to determine whether instructional changes improve student skill deficits.

Theodore J. Christ and Danielle M. Becker

See also Curriculum; Curriculum-Based Measurement; Formative Assessment; Progress Monitoring; Screening Tests

Further Readings

Christ, T., Keller-Margulis, M., & Marcotte, A., (2014). The basics of curriculum-based assessment. In S. G. Little & A. Akin-Little (Eds.), Academic assessment and intervention. New York, NY: Routledge.

Deno, S. (1985). Curriculum-based measurement: The emerging alternative.

Exceptional Children, 52, 219–232.

Fuchs, L., & Deno, S. (1991). Paradigmatic distinctions between instructionally relevant measurement models. Exceptional Children, 57, 488–500.

Gickling, E., & Thompson, V. (1985). A personal view of curriculum-based assessment. Exceptional Children, 52, 205–218.

Hintze, J., Christ, T., & Methe, S. (2006). Curriculum-based assessment. Psychology in the Schools, 43, 45–56.

Theodore J. Christ Theodore J. Christ Christ, Theodore J.

Kirsten Newell Kirsten Newell Newell, Kirsten

Curriculum-Based Measurement Curriculum-based measurement

Curriculum-Based Measurement

Curriculum-based measurement (CBM) has emerged as the most prominent, researched, and influential among a number of curriculum-based assessment methodologies. CBM was developed to index the level and rate of student achievement in the basic skills of reading, mathematics, spelling, and written expression. CBM comprises a standardized set of procedures to administer and score student performance. This entry further describes CBM and then discusses its application in schools, its perceived benefits, and concerns raised about the timed conditions of CBM.

As the name indicates, the measurement content was originally sampled from the local curriculum used for instruction. Students would read, write, or perform mathematical calculations for short durations (1–3 minutes). Performance-based responses contrasted with multichoice-type response formats and curriculum-based content contrasted with the generic content of many published tests. These aspects of CBM were designed to establish more authentic and relevant measures of student performance in the specific learning environment. That information was thought to be useful for teachers to monitor student performance in the annual curriculum and, especially, the performance of students who received special education services.

CBM procedures were designed by Stanley Deno and colleagues to be reliable and valid, simple and efficient, easily understood, and inexpensive. They were also intended to be repeatable. Together, these features provided a measurement procedure for use by teachers to routinely evaluate the effects of an instructional program.

Routine and ongoing evaluations are useful to guide when and whether to change a student’s instructional program. That interpretation and use of a data is a type of formative evaluation, and CBM is often described as a formative assessment because it is intended for use to inform instruction. The use of CBM to monitor student progress and evaluate program effects gained substantial popularity and facilitated the emergence of response to intervention, which is a contemporary approach to diagnose learning disabilities. This often entails the use of CBM to develop local norms or compare student performance against benchmark standards to identify those who are at risk or who need supplemental or intensive academic supports.

CBM scores indicate the rate of accuracy performance per unit of time, which is typically in 1-minute intervals. For example, CBM oral reading is a 1-minute oral reading from a grade-level passage. The administrator listens and marks errors and calculates the words read correctly per minute. CBM maze is a 2-to 10-minute silent reading with every fifth or seventh word replaced with a multiple-choice list of three to four alternatives, with the correct word circled by the student. The items completed within the interval are scored to calculate correct selections per minute, which is sometimes replaced by more sophisticated scoring procedures to account for guessing. CBM computation is a 2-to 6-minute interval in which the examinee completes grade-level computation problems. The written work is scored to calculate the digits correct per minute. CBM written expression and CBM spelling are similar with outcomes of correct word sequences and correct letter sequences per minute, respectively.

The CBM rate-based scoring emphasizes the fluency and automaticity of basic skills. This has been widely influential and somewhat controversial. CBM drew attention and focus to promoting accurate and rapid performance of basic skills, such as word identification, computation, writing, and spelling. Fluent performance of the targeted basic skills functions as robust indicators of academic progress and well-being. Benchmark levels of performance and progress were adopted by many educators as key indicators of high-quality instructional programs and student achievement within the early grades.

CBM has been especially popular within the special education community, which serves students with disabilities. Certain disabilities and academic deficits are attributed, in part, by some, to fluency deficits. Many also have attributed the increased sensitivity of CBM procedures to the fluency feature. That is, many have advocated for and used CBM because it was believed to be more sensitive to smaller increments of student achievement and instructional effects than other to smaller increments of student achievement and instructional effects than other untimed measurement procedures.

Some educators and researchers reject the concept of fluency for a variety of reasons. One prominent objection is that it is stressful for children to be encouraged to perform quickly and assessed within timed conditions. Although such objections persist, CBM scores predict performance on many untimed and less efficient measures, which include state and national measures of academic accountability along with other national normed tests.

Theodore J. Christ and Kirsten Newell

See also Curriculum; Curriculum Mapping; Curriculum-Based Assessment; Formative Assessment; Progress Monitoring; Response to Intervention

Further Readings

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research curriculum-based measurement of math computation. Assessment for Effective Intervention, 33(4), 198–205.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. Exceptional Children, 52, 219–232.

Deno, S. L. (2003). Developments in curriculum-based measurement. The Journal of Special Education, 37(3), 184–192.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. School Psychology Review, 33(2), 188–192.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. Exceptional Children, 57, 488–500.

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. Exceptional Children, 78(4), 423–445.

Hintze, J., Christ, T., & Methe, S. (2006). Curriculum-based assessment. Psychology in the Schools, 43, 45–56.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. Journal of Special Education, 41(2), 68–84.

Marianne Perie Marianne Perie Perie, Marianne

Cut Scores

Cut scores

451

456

Cut Scores

The setting of cut scores is a specific and precise way of establishing a standard of performance for a test. Standard setting is defined as a process of determining the point on a test’s score scale used to establish whether a particular test score is sufficient for some purpose, but it’s also the primary test development activity where psychometrics, content, and policy intertwine. Determining the point at which student performance is “good enough” involves a policy directive, understanding of the content, and an evaluation of the reliability and psychometric difficulty of the individual items or of the test as a whole.

Although some have referred to standard setting as “alchemy” given that it relies on human judgment, it is a process that allows policy to interact with content and psychometric considerations. Decisions about performance-level descriptions, panels, and methods should be made by people who understand each of these three areas and inform the workshop facilitation, aligning it with the intended use of the outcomes. Complete standardization is an impossible goal but should be attempted to the degree possible when working with human judgments. This entry first looks at the increased importance placed on standard setting and determination of cut scores in recent years, then discusses the influence of the facilitator, panelists, and method in setting cut scores.

In the movement to standards-based testing in the 1990s and 2000s, standard setting started relying more heavily on written descriptions of the performance levels to determine the cut scores that defined them. As K–12 school systems began using standardized tests for purposes of accountability, multiple cut scores were needed, contrasting with certification and licensure tests that typically only set a pass/fail line. In K–12, policy makers and content experts determined the set a pass/fail line. In K–12, policy makers and content experts determined the rigor of each performance level through a written description of the expected knowledge and skill at each level. Later, these descriptions were also used to drive item writing, ensuring that sufficient items were developed that distinguished clearly among the levels.

Using performance-level descriptions in the standard-setting process also standardized the process further, as panelists could focus on those descriptions of knowledge and skills rather than relying solely on their own opinion of sufficient knowledge and skills to meet a performance level. This addition to the process resulted in greater interrater reliability in the cut score recommendations.

Yet, cut scores can still be influenced by three main factors: the facilitator, the panelists, and the method used for standard setting. There are multiple chapters, books, and articles about how to conduct a standard setting workshop, and on the many methods for determining cut scores. Here, the focus is on the parts of standard setting that can have the most effect on cut scores and how they can be standardized to the extent possible.

Facilitator

The facilitator can have an immense effect on the standard setting process. There are obvious ways a facilitator can have a negative impact—for instance, by not being prepared or not explaining the process clearly. However, facilitators can have much more nuanced influence as well. For instance, a facilitator's focus and tone of voice can convey that a particular factor is more or less important.

An example of the effect of the way a facilitator presents information would be in the introduction of impact, or consequence, data. Typically, these data regarding the actual or projected number of students to reach each cut score are provided at the end of Round 2. Panelists will have already made two judgments on where the cut score(s) should be placed. Now they see the results of those placements. A facilitator can place a lot of emphasis on the impact data, discussing the use and any consequences for test takers or others, and caution the panelists to consider the data carefully. Conversely, a facilitator can remind the panelists that they have spent two rounds making reasoned, criterion-based judgments, and these data are shown simply to provide supplemental information. Depending on how the data are presented, the amount of change made to the recommended cut scores in Round 3 can vary considerably.

Before any standard setting workshop, those running the workshop should work closely with policy makers to determine where the emphasis should be placed, how impact data should be presented, and how questions about the standards, test questions, and use of the assessment should be answered. Once an agreement is made on all issues that could influence the cut scores, a script should be prepared that introduces each topic, provides a list of frequently asked questions and answers, and specifies important policy context. The script should be approved by affected policy makers. Then, the script should be used by everyone facilitating that standard setting workshop. For instance, if a workshop is intended to set cut scores on science assessments in three different grade levels, there may be three facilitators. They should all present the information in a common manner.

In addition to a script, having the facilitators practice facilitating with that script is essential. The standard setting manager should listen to how each facilitator explains the task and introduces each task to ensure they are as similar as possible and match the policy makers’ requirements for emphasizing certain areas or responding to policy questions.

The script and rehearsal will greatly reduce the variance caused by a single facilitator, as will have a manager observing all rooms where standard setting is occurring. Final evaluation forms should ask panelists the degree to which they felt free to provide their own opinions versus feeling coerced as well as the primary factors influencing their cut score recommendation.

Panelists

The people brought in to give their judgments on the cut scores can also be hugely influential. Two groups of panelists, given the same instructions, may generate different cut scores. Because of their influence, panelists should be selected carefully. Prior to recruiting, the standard setting manager should determine the target panel. Depending on the assessment, the target panel may be composed primarily of content experts, teachers, or employees in the field of the assessment. Stakeholders may also be desirable, but the primary composition of the panel should be people who understand the construct being assessed and the requirements for the population of test takers.

After qualifications have been determined, the next consideration is demographics. For many assessments, it is important that the standard setting panelists are representative of the general population or at least the pool of qualified candidates. For instance, if the target panel is composed solely of seventh-grade mathematics teachers within a state, the demographic makeup of the panel should be comparable to the demographics of the pool of seventh-grade mathematics teachers in the state.

Within this determination is also the question of how many panelists are needed. Large panels are desirable for ensuring sufficient stakeholders are included, but they can make thoughtful discussion difficult, particularly for individuals who are less outspoken. Thus, large panels are typically divided into smaller tables for discussion purposes. Recruiting more people and separating them into tables also provides some measure of within-facilitator, cross-panelist variance. Ideal table sizes are no smaller than five people and no larger than eight to encourage participation of all panelists and sufficient numbers for a reliability analysis. Depending on the number of judgments needed, more or fewer panelists may be needed. For example, in a modified Angoff approach, the number of judgments equals the number of items times the number of cut scores. Therefore, a 50-item test with two cut scores will require 100 judgments.

A larger number of panelists are needed to achieve an appropriate level of reliability in the final cut score recommendations. A typical panel would include 30–35 panelists divided into five tables. In a bookmark approach, the number of judgments equals the number of cut scores. Standard setting workshops have been held with as few as eight panelists for a bookmark method but are better with multiple tables. In this case, three tables with six to eight panelists per table is preferred. For a holistic approach, such as with the body of work method, fewer judgments are required, but greater agreement is often desired. In those cases, fewer panelists may be preferred. This method is often conducted with six to eight panelists, although, again, having at least two tables allows for measurement of cross-panel variance.

A final consideration in creating a standard setting panel is the stakes associated with their recommendations and the finality of them. Typically, panels only make a recommendation and then a policy body adopts the final cut scores. Understanding what factors are important to that policy body is important in selecting an appropriate panel. If the workshop is discussed in the news, what reassurances about the people involved in the decision should be given? And, more panelists sounds more rigorous. In a high-stakes arena, the importance of the constitution of the panel cannot be understated.

Method

Finally, even with well-trained facilitators and qualified, representative panels, the cut scores could still differ by method. The most important rule in selecting a method is to ensure that the cognitive task required of the panelists matches the assessment design. For instance, if the assessment was designed as a portfolio of student work or consists primarily of a research paper or essay, a more holistic approach is needed to set cut scores. Conversely, if the assessment requires the students to respond to a large number of multiple-choice items, an item-based approach is more appropriate. For the purpose of this argument, we will focus on an item-based approach.

Even within one class of methods, there are multiple methods to choose from and many modifications or enhancements to that method. Consideration of the test design, relevant features, and both student and panelist cognitive tasks are important in selecting and modifying an assessment.

For example, consider a fourth-grade reading comprehension test comprising 60 dichotomous items. Bookmark is the most common standard setting method used in K–12 testing today. Given the importance of the passage in determining the difficulty of the test question, a traditional bookmark method can obfuscate that connection by separating the passages from the items, requiring the panelist to go back and forth from questions to passages and not see the full set of questions associated with a passage. For instance, a test question that asks the student about the author’s purpose could vary in difficulty for a passage where the author clearly states his purpose compared to a passage where the student must infer it from the information the author presents.

In order to keep the focus on the passage, a modification of the bookmark method groups the ordering of test questions within a passage and orders the passages themselves based on overall difficulty of the question set. This format allows panelists to first focus on the difficulty of the passage, discussing various components that contribute to its complexity. Next, they can analyze the interaction of the test question associated with that passage. Only as a later step, panelists would be asked to examine questions across passages in a fully ordered booklet.

In another scenario, consider another test comprising 60 dichotomous items, but now it’s a career pathway test given to high school students in a career and technical education program. Given that students take the test at different times of the year and in different years depending on their program and personal goals, it might take an entire year of testing to gain a representative sample of test takers for use in standard setting. Yet, cut scores are needed after the first administration to provide results to that first set of students.

Policy makers often worry about the validity of the impact data with a skewed sample, but it also affects the method chosen. Because the bookmark method requires test questions to be ordered by their psychometric difficulty, using a skewed or nonrepresentative sample to do so can greatly affect the results. In some cases, a modified Angoff method may be a preferred approach as it is not dependent on item difficulty. Some feedback on how the early sample did may be given, but it should be given in the context of the characteristics of those in that sample.

Another possibility is an ordered-item yes/no Angoff method that does provide panelists some information on how the items are ordered based on psychometric difficulty but, unlike the bookmark method, allows for some judgments to be out of line with that ordering. Panelists review each item as compared to the target definition for that cut score and say "yes, two thirds of students meeting this definition would answer this question correctly" or "no, they wouldn't." They record a yes or a no for each item. Typically, they have a pattern of yeses followed by a block of no's, but they can also choose to go out of order for certain items that they feel are misrepresented in difficulty based on the characteristics of the population that initially took the assessment.

Marianne Perie

See also Achievement Tests; Angoff Method; Classification; Common Core Standards; Ebel Method; High-Stakes Tests; Psychometrics; Standard Setting; State Standards; Tests

Further Readings

Cizek, G. (Ed.). (2011). Setting performance standards: Foundations, methods, and innovations (2nd ed.). New York, NY: Routledge.

Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. Educational Measurement: Issues and Practice, 32(3), 27–35.

Perie, M. (2008). A guide to understanding and developing performance level descriptors. Educational Measurement: Issues and Practice, 27(4), 15–29.

Perie, M., & Thurlow, M. (2011). Setting achievement standards on assessments for students with disabilities. In G. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed.). New York, NY: Routledge.

Skaggs, G., Hein, S., & Awuor, R. (2007). Setting passing scores on passage-based tests: A comparison of traditional and single-passage bookmark methods. Applied Measurement in Education, 20(4), 405–426.

Smith, R. W., Davis-Becker, S. L., & O'Leary, L. S. (2014). Combining the best of two standard setting methods: The ordered item booklet Angoff. Journal of Applied Testing Technology, 15(1), 18–26.

Zieky, M., Perie, M., & Livingston, S. (2008). Cutscores: A manual for setting performance standards on educational and occupational tests. Princeton, NJ: Educational Testing Service.

D

Danielson Framework

Danielson framework

457

457

Danielson Framework

See Framework for Teaching

Brandon LeBeau Brandon LeBeau LeBeau, Brandon

Data

Data

457

458

Data

Data are the fundamental building blocks upon which all educational research is built. The term *data* is the plural form of *datum*, referring to a single piece of information, whereas *data* refer to pieces of information. This distinction is most prominent in scientific or academic writing, whereas in other forms of writing, *data* can be singular or plural. Technology enhancements, particularly computer storage capacity, processor speed, and computer portability, have significantly increased the amount of data available to researchers. This entry explores in more detail common forms data can take in educational research and how they are used.

Forms of Data in Educational Research

Data can take on many different forms and have varying amounts of information attached to them. They can be qualitative or quantitative. *Qualitative data* is generally the term used for information that is not in the form of numbers or quantities. It might be written or spoken words or descriptions of observed behavior, or in a variety of other forms that cannot be easily summarized as amounts of something.

Quantitative data are in the form of numbers. These numbers, usually in the form of scores, can represent different levels of measurement and can be nominal, ordinal, interval, or ratio data. Moving from nominal to ratio levels of measurement will naturally carry more information with the data. Nominal data simply represent categories, whereas ratio data are quantitative with a meaningful zero.

Many variables in education are nominal, ordinal, or interval of nature. Understanding the level of measurement the data take is an important step because many inferential modeling procedures such as *t* tests or regression assume that the dependent variable is at least an interval scale of measurement.

One specific example commonly used in education is the collection of survey data through a rating scale (e.g., *strongly disagree* to *strongly agree*). It is common for educational researchers to represent these categories as numeric values; for example, if there are five categories, these are often represented as numeric integers from 1 (*strongly disagree*) to 5 (*strongly agree*). Sometimes responses to survey questions are summed to create a scale measuring some phenomenon; however, in other instances, the responses for a single item are used. Some caution should be exercised in assuming these types of data for a single survey question are interval in nature, as the gaps are likely not consistent across the entire range of the scale.

Using Data in Educational Research

Data in education can be collected from experiments, observational studies, or even computer simulations. These data are often used in descriptive or inferential analyses to help aid in making decisions regarding the effectiveness of educational programs. As computers have become more powerful and more portable, data can now be collected daily, hourly, or even every second. This has led to a significant increase in the amount of data available to educational researchers. In addition, more complex models are being used to explore this information that may make more assumptions regarding the data. Therefore, particular attention and time need to be made to thoroughly understand the type of data being used and whether assumptions are being met for the data analysis.

Brandon LeBeau

See also Descriptive Statistics; Inferential Statistics; Levels of Measurement; Qualitative Research Methods

Further Readings

Feldman, J., & Tung, R. (2001). Using data-based inquiry and decision making to improve instruction. ERS Spectrum, 19(3), 10–19.

Johnson, B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), Handbook of mixed methods in social and behavioral research (pp. 297–319). Thousand Oaks, CA: Sage.

Stiegelbauer, S. M. (1982). Through the eye of the beholder: On the use of qualitative methods in data analysis. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Walter L. Leite Walter L. Leite Leite, Walter L.

Zachary K. Collier Zachary K. Collier Collier, Zachary K.

Data Mining

Data mining

458

461

Data Mining

Data mining is a series of methods that aim to discover knowledge from data by applying algorithms. The algorithms for data mining are very diverse, depending on their intended objectives and the computational demand of the problem. Data mining methods have been developed at the intersection of the academic areas of statistics and computer science. Data mining methods can also be classified broadly into supervised and unsupervised learning. In this entry, methods for supervised learning used for prediction are reviewed first, followed by methods for unsupervised learning.

Supervised learning consists of methods applicable to data in which there is an outcome that can be used to determine whether the learning process was successful. The outcome is also commonly referred to as a dependent variable or response variable. Supervised learning methods can be used for prediction and learning about relationships between predictors and the outcome. Examples of methods of supervised learning include generalized linear models, classification and regression trees, random forests, and neural networks (NNs). Methods of supervised learning have found several applications in educational research, such as identifying students at risk of failing to reach achievement milestones or identifying the effects of educational interventions. Some methods for supervised learning allow inference about the general form of the relationship between predictors and the outcome or provide measures of predictor importance, whereas other supervised learning methods provide predictions but do not allow any inference about the functional form of the relationship between predictors and the outcome.

predictors and the outcome.

Unsupervised learning consists of methods in which there is no outcome, and therefore, their goal is to summarize data by finding similarities or associations between individuals or variables. Unsupervised learning methods include methods for clustering, association rule mining, principal components analysis, and exploratory factor analysis (EFA). At the student level, unsupervised learning has been applied in educational research to identify groups of students who respond similarly to measurement scales, have comparable growth trajectories, or benefit equally from educational interventions. At the instrument level, unsupervised learning has been used to group survey and scale items with respect to relationships with constructs and/or content areas measured.

Prediction Methods

The goal of prediction methods is to forecast an outcome using a set of predictors also known as independent variables or features. Prediction problems are commonly classified into regression or classification problems, depending on whether the outcome is continuous or categorical, respectively. In data mining for prediction, it is customary to use a training data set, which is a subset of the available data, to build the predictor, then a different subset as a test data set to evaluate the predictor. The quality of prediction can be evaluated with the mean squared error and the error rate in regression and classification problems, respectively.

There is a large number of parametric and nonparametric methods for prediction. Parametric methods make assumptions about the form of functional relationship between the outcome and predictors and the distribution of residuals, whereas nonparametric methods do not make these assumptions. One common type of parametric model is the generalized linear model, which uses different link functions (e.g., identity, log, and logit) and distributional families (e.g., Gaussian, Poisson, and binomial) to establish a linear relationship between a set of predictors and a function of the outcome. Examples of generalized linear models are linear regression, Poisson regression, and logistic regression.

For situations in which the form of the relationship between predictors and the outcome is very complex and cannot be adequately described by a generalized linear model, prediction can be obtained with generalized additive models, which use a smoothing function to approach the complex relationship. One example of generalized additive model is local linear regression, which consists example of generalized additive model is local linear regression, which consists of splitting the sample into sections based on predefined bandwidth and applying a linear regression within each section. There is a bias-variance trade-off in selecting the length of bandwidth: Having a small bandwidth achieves better fit to the data at a cost of larger confidence intervals, whereas large values of bandwidth produce smaller confidence intervals but may introduce bias.

Recursive partitioning methods are also used for prediction and rely on splitting the data into groups with similar values of the outcome. Recursive partitioning methods include classification and regression trees, bagging, and random forests. Recursive partitioning methods have the advantage of automatically identifying interactions between variables and can be applied to data with missing values. The simplest recursive partitioning methods are classification and regression trees, where classification trees predict categorical outcomes and regression trees predict continuous outcomes. These are iterative methods that split the sample into two nodes that are more homogenous with respect to a predictor than the entire sample. The process starts with identifying the variable whose split provides the most homogenous nodes, then continuing the process with other variables or other nodes of the same variable. Several measures can be used to quantify the extent of homogeneity within nodes, such as Bayes error, entropy, and the Gini index. Classification and regression trees have been implemented in a variety of algorithms, such as CART, C4.5, CHAID, and GUIDE. Bagging or bootstrapped aggregation consists of running multiple classification or regression trees on bootstrapped samples of the data with all available predictors and combining the results. Random forests are also based on bootstrapping samples, but it is different from bagging in that only a subset of predictors is used in each iteration, which increases the variability in the results and prevents any predictor from dominating the process. Generalized boosted modeling can be used to improve any predictor, such as classification trees, regression trees, and logistic regression, by iteratively applying the predictor with an adjustment at each step.

Artificial NNs are a prediction method built to mimic the way the brain processes information. NNs are divided into three layers: (1) input layer, (2) hidden layer, and (3) output layer. Similar to the way the brain learns through experiences, an NN learns by way of training the NN’s algorithm. Many training algorithms have been developed for NNs, but the most popular is back propagation. When training an NN, the goal is to partially adapt the model to the data before maximum likelihood estimation. However, a Bayesian framework has been recently used for NN instead of maximum-likelihood estimation.

Clustering Methods

The goal of clustering methods is to place subjects (e.g., students, teachers, and schools) into groups, such that subjects in a group are similar to each other and dissimilar to subjects in other groups. K-means clustering and latent class analysis are popular clustering methods in educational research. With the K-means clustering method, clusters are determined by grouping subjects into clusters with the nearest mean. The K-means clustering algorithm can be summarized in four steps: (1) K points are placed to represent the initial group centroids, (2) subjects are assigned to the nearest group, (3) K points are recalculated, and (4) Steps 2 and 3 are looped until centroids no longer move. Although a number of clusters within a data set are guaranteed to be selected using K-means, a major drawback is that the number of groups is based on the initial placing of the K points. A second clustering method, latent class analysis, is a mixture modeling technique and can be used to find unobservable groups that are different based on observed characteristics, identify and accurately enumerate the number of groups, identify characteristics that indicate groups well, estimate the prevalence of the groups, and classify individuals into classes. Latent class analysis is designed with the assumption that there are an unknown number of groups that can account for observed probabilities. As a result, latent class analysis models are selected by fitting models with a different number of classes and comparing them with respect to likelihood ratio tests and information indices.

Association Rule Mining

Association rule mining is a useful starting point for identifying relationships that require further analysis. Association rules contain antecedents and consequences within a given data set. Many queries pertaining to educational data sets concern themselves with choosing curriculum for students and professional developments for teachers. In such cases, researchers may be interested in finding pairs of curriculum (antecedent) and teacher professional developments (consequence) that school districts are likely to choose. The relationship between student curriculum and teacher professional development is represented as student curriculum → teacher professional development.

Once these if → then patterns are found, association rule mining determines the support, confidence, and interest of the rule. Support is based on the frequency that the items in question appear in the data set, whereas confidence indicates the that the items in question appear in the data set, whereas confidence indicates the percentage of times the if → then patterns are found true. Interest is the negative or positive correlation between the consequence and the antecedent. High values for support, confidence, and interest do not indicate a causal rule that student curriculum causes school districts to determine teacher professional development. The association rule indicates only that the antecedent and consequence co-occur. It is also possible that there could be more than one antecedent in the rule, such as multiple student curricula and teacher professional development that have high values for support, confidence, and interest:

> student curriculum 1, student curriculum 2, student curriculum 3 → teacher professional development.

Principal Component Analysis (PCA) and EFA

Both PCA and EFA aim to produce a reduced number of dimensions that account for the correlation matrix of responses to a larger set of variables. Examples of variables in educational research used in PCA and EFA are items from cognitive or noncognitive assessments, survey questions, and behavioral observations. The difference between PCA and EFA is that while PCA aims to create components that account for all the variance in the observed variables, EFA creates factors that only account for the shared variance between variables. Underlying EFA, there is a measurement model that specifies that the variance of each variable is partly due to a latent construct that is common to all variables and partly due to measurement error that is unique to each variable. For example, in an EFA of items of a mathematics test, part of the variance of the items would be due to mathematics achievement, whereas the remaining variance would be due to measurement error. Therefore, principal components are composites of the original variables, whereas factors are linear combinations of the original variables minus measurement error. EFA follows these steps: (a) extraction of factors, (b) determination of the number of factors to retain, (c) rotation of solution, and (d) factor interpretation. There are several methods for factor extraction, such as principal axis factoring and maximum likelihood estimation. The determination of the number of factors can be done by multiple criteria, such as eigenvalues larger than one, examination of a scree plot, and parallel analysis. There are also several rotation methods, with some allowing factors to correlate (e.g., oblimin rotation) and others forcing orthogonality of factors (e.g., varimax rotation). The objective of rotation is to increase the interpretability of the factors by approaching simple structure, where each item is most strongly related to a single factor. Factor interpretation consists of attributing meaning to the factors based on the nature of the observed variables found to relate to the factor most strongly.

Walter L. Leite and Zachary K. Collier

See also Cluster Analysis; Data; Exploratory Factor Analysis

Further Readings

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. New York, NY: Springer. doi:10.1007/978-1-4614-7138-7

McArdle, J. J., & Ritschard, G. (Eds.). (2014). Contemporary issues in exploratory data mining in the behavioral sciences. New York, NY: Routledge.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. D. (Eds.). (2011). Handbook of educational data mining. Boca Raton, FL: CRC Press.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington, DC: American Psychological Association.

Alexandru C. Telea Alexandru C. Telea Telea, Alexandru C.

Data Visualization Methods Data visualization methods

461

464

Data Visualization Methods

The term *data visualization* refers to the process of transforming data, which is generated by means of measurements of various processes taking place in the physical world or created by computer applications such as simulations, to pictures. The purpose of this transformation is 2-fold: (1) to help users understand their data better and easier and (2) to let them discover unknown facts about the underlying phenomena from which data are derived. To be able to do this, several classes of visualization methods have been developed, each of them focusing on a specific type of data, users, or application domain. This entry describes the landscape of data visualization methods and their relationships with science and engineering. Particular attention is devoted to the increasing role of visualization in education. The entry concludes with an overview of the challenges of teaching data visualization as well as a list of resources on data visualization methods.

Aims and Scope of Data Visualization

The core aim of data visualization is to provide ways to handle information by means of the study of graphical representations thereof. This serves several goals, as follows. Large amounts of data can be compactly presented, therefore easing the user’s burden of separating important aspects from details and also reducing the time and effort required to study a given process (*scalability*). Nontechnical users can be shielded from complex aspects related to data acquisition and processing, so they can focus on those high-level aspects that the data capture (*simplicity*). Different types of users having different backgrounds are enabled to communicate and learn about a data-intensive problem based on the same (visual) medium (*communication*). Finally, the visual depiction of data enables finding complex patterns that one is not aware of and which are hard to find when using solely traditional data analysis methods (*discovery*).

Data visualization serves two main purposes. First, data can be *presented* to interested audiences, such as scientists, professionals, or students. In this context, the presenter uses visualization as an enabling instrument to facilitate communication, by reducing the complexity of the exposition and focusing on the essential details. This usage of visualization is the most widespread and covers presentation means as diverse as infographics, PowerPoint presentations, narrated videos, and web-based dashboards. Central to this use-case is that the presenter already knows the information to be communicated and chooses the visual instruments and presentation techniques that best suit an efficient and effective presentation. Consumers of such visualizations range from professionals and specialists to students at all education levels and, more generally, the grand public. This usage of visualization is also known as "presenting the known."

A second role of data visualization addresses data *exploration*. The aims of the audience for this role of visualization differ from the previous one: Both presenter and public are now interested in discovering previously unknown aspects embedded in a given data collection. As such, there is a less clear difference between presenters and public, so one typically speaks about visualization *users*. This usage of visualization relies upon more specialized instruments, such as advanced computer programs that allow users to interactively drill down in large data collections, select specific data subsets of interest, and depict them visually by a wide range of techniques, each of which focuses on the discovery of a different kind of pattern present in the data. Consumers of such visualizations are professionals who are intimately familiar with the application domain from which the data at hand emerge and also with the aforementioned specialized visualization instruments. This usage of visualization is also known as "discovering the unknown."

Types of Data Visualization Methods

Before the advent of modern computing technology, collecting and visually depicting data has been done manually, leading to a variety of maps, graphs, and charts. The study of these depictions has led to the formulation of several key design principles for the creation of effective visualizations. As described by

Jacques Bertin and Edward Tufte, these involve the choice of suitable *visual variables* (such as position, shape, size, color, brightness, texture, and animation speed) to encode variables in the data to be displayed, the maximization of ratio of displayed data to ink being used in the display, the avoidance of visual clutter, and the consistent use of legends and annotations that explain how data have been visually encoded. Data visualization can be seen as a blend between graphics design, visual perception, cognitive science, and data science.

The increasing computing power available to generate, collect, and analyze data and the increase in resolution and color quality of computer displays have shifted the focus from hand-crafted visualizations to computer-generated ones. Following these developments, data visualization has evolved and diversified since the 1970s into a number of subfields. These are outlined in the following sections.

Scientific Visualization

A first main use-case for computer-generated visualizations has been to explore the increasing amounts of data generated by numerical simulations of physical phenomena. The resulting discipline has been called *scientific visualization* due to its original presence in scientific, engineering, and research laboratories. Since its appearance in the 1970s, scientific visualization has evolved to cover the visual analysis of data collections from geosciences, weather and climate science, and medical science. The main characteristic of scientific visualizations is that they target data that have a natural spatial embedding, such as surfaces and shapes that exist in two or three dimensions, and whose measured attributes are described on a continuous scale, such as length, density, or temperature. Key methods developed in the context of scientific visualization include techniques for the display of scalar data, such as temperature or pressure; vector data, such as force or velocity; and volume data, such as three-dimensional computer tomography scans.

Information Visualization

A subsequent main use-case for data visualization has been created by the explosion of data generated by developments in information technology, sensor devices, and the Internet with its various data-intensive facets such as social networks, online commerce, and e-governance. The rapid increase in volume, velocity of change, and variability of such data is currently known as the “3V” characteristics of so-called *big data*. To cope with this, new visualization methods have been developed. In contrast to scientific visualization, these aim at displaying data that do not have a natural embedding in two or three spatial dimensions and whose measured attributes are not necessarily continuous quantities. Examples of such data collections include tables stored in large data warehouses, document archives, Twitter feeds, and communication networks. Key methods developed in information visualization address the display of networks and hierarchies, tables having hundreds of columns and hundreds of thousands of rows, and trends and similarity relations in large time series such as stock exchange data. Used originally in business intelligence and homeland security contexts, information visualization has spread to many other application domains in which large collections of nonspatial data need to be explored, such as software maintenance, medicine, and bioinformatics.

Visual Analytics

Following the establishment of both scientific and information visualization, both practitioners and researchers have observed that the tools and techniques offered by these two subdisciplines are necessary, but not sufficient, for the end-to-end process of getting insight in complex data-intensive phenomena. As increasingly sophisticated tools have been provided by both scientific and information visualization, the complexity of the problems that such tools have to address has also increased. Getting insight into such problems has become more intricate than simply using one or a few suitable techniques to display the available data. For a given use-case, many exploration paths are possible, many hypotheses on the causes behind the observed data have to be investigated in parallel, and each such investigation can be done using multiple techniques.

Visual analytics has emerged as the answer to the explosion of the “search space” and “tool space.” Rather than imagining novel visualization techniques, visual analytics focuses on ways to combine techniques of different kinds, such as data mining and searching, statistics, machine learning, and visualization, into realizing end-to-end pipelines that help users answer complex questions on the data at hand. Typical to visual analytics is the iterative refinement of knowledge and insight into the problem under study, which is done by the use of interactive visualization techniques. As such, and in contrast to earlier scientific and information visualization solutions, visual analytics focuses on the needs of an analyst to make sense of data, atop of the needs of a user interested solely in analyst to make sense of data, atop of the needs of a user interested solely in seeing the data. Following this analogy, visual analytics focuses on supporting the process followed by the user to extract knowledge, of which the images that depict data are just one component.

Data Visualization in Education

Visualizing data is a crucial part of technical and scientific communication, starting from the undergraduate level and ranging up to the senior researcher and practitioner level. Its tools of the trade range from relatively simple infographics like bar, line, and pie charts to sophisticated tools for the visual exploration of dynamic networks and data tables of millions of elements.

As such, visual communication has become an increasingly important element of technical and scientific education. Typical elements covered by this educational process are the visual design of slide presentations, infographics, and illustrations for scientific articles and of the associated narratives. Typical instruments supporting this process are presentation tools such as PowerPoint and Keynote, but also more advanced data visualization tools such as Matlab and R.

While constructing scientific and information visualizations has become increasingly easy due to the diversification of available software tools, important challenges still exist in educating the upcoming generation of scientists and practitioners to create effective visualizations. One such key challenge relates to the scope, or focus, of visualization education forms, which can be roughly divided into two types. The first type focuses on visual design, perceptual factors, and presentation techniques that are required by an effective visualization—or in other words how to design (but not how to implement) visualizations, so as to avoid “chart junk,” or the creation of graphically rich, but information-scarce, visualizations. Such knowledge is provided in the context of many studies, ranging from exact sciences to cognitive and communication sciences and graphics design. This education type focuses heavily on the use of existing ready-to-use tools of low to moderate sophistication, such as PowerPoint, Keynote, and Tableau. As such, it serves a wide spectrum of users but cannot deliver customized visualizations for certain problems, data types, and user questions.

Conversely, the second type of visualization education focuses mainly on technical issues of all types of visualizations (scientific, information, and visual analytics)—or, in other words, how to implement a given visual design. This approach to visualization education covers aspects such as the choice and combination of data representation; data storage, processing, and data mining algorithms; and computer graphics and interaction techniques involved in the realization of an end-to-end visualization software application. This education type focuses on extending and adapting visualization software coming in the form of libraries and frameworks and creating novel research-grade visualization methods. As such, it addresses a narrower spectrum of users than the first type, as a solid background in mathematics, statistics, and software engineering is required. The drawback of this approach is its relatively specialized and narrow focus in terms of addressing only a subset of visualization problems and requiring advanced technical skills from its students. An optimal formula for visualization education combines an entry-level course in visualization design general principles, followed by an in-depth, more specialized course covering the implementation of visualization techniques for a selected application area.

Many resources are available to study data visualization. On the design side, books by Edward Tufte and Jacques Bertin give an excellent overview of the requirements and guidelines for creating effective, compelling infographics. The books by Tamara Munzner and Colin Ware cover information visualization design. Andy Kirk offered a communication-sciences view on data visualization. Stephen Few’s book focused treatment of information visualization with the accent on tables and graphs. Finally, Alexandru Telea’s book covered the technical implementation aspects of data visualization, with a focus on scientific visualization, and also provides a survey of popular visualization software tools and applications.

Alexandru C. Telea

See also Active Learning; Data-Driven Decision Making; Data Mining; Graphical Modeling; Multidimensional Scaling; R

Further Readings

Bertin, J. (2010). Semiology of graphics: Diagrams, networks, maps. Esri Press.

Dill, J., Earnshaw, R., Kasik, D., Vince, J., & Wong, P. C. (2012). Expanding the frontiers of visual analytics and visualization. Springer.

Few, S. (2012). Show me the numbers—Designing tables and graphs. Analytic Press.

Kirk, A. (2012). Data visualization: A successful design process. Packt Publishing.

Munzner, T. M. M. (2015). Visualization analysis and design. CRC Press.

Telea, A. C. (2014). Data visualization—Principles and practice (2nd ed.). CRC Press.

Tufte, E. (1983). The visual display of quantitative information. Cheshire, CT: Graphics Press.

Ware, C. (2012). Information visualization: Perception for design. Morgan Kaufman.

Amy S. Gaumer Erickson Amy S. Gaumer Erickson Erickson, Amy S. Gaumer

Patricia M. Noonan Patricia M. Noonan Noonan, Patricia M.

Data-Driven Decision Making Data-driven decision making

464

466

Data-Driven Decision Making

Data-driven decision making involves using information to inform decisions and enact change. Although seemingly simple, the process involves determining pertinent data, collecting and analyzing data, discussing results and drawing conclusions, enacting change, and monitoring progress. This process is often depicted as a cycle of continuous improvement. This entry describes the information used in data-driven decision making, the data-driven decision-making cycle, and the elements that must be in place for the process to be effective.

In educational evaluation, data can include any information about the student, classroom, school, or community that can be gathered, reviewed, and analyzed. Quantitative data might include student assessments administered to an individual, class, or grade as well as indicators such as attendance, office disciplinary referrals, or graduation rates. Quantitative data can be used to pinpoint trends and monitor progress. Qualitative data are often helpful in understanding processes, perceptions, and experiences that quantitative data cannot fully represent. Qualitative research, such as student and family interviews or classroom observations, is less likely to be conducted at the school-wide level but can offer depth of insight.

Effective data-driven decision making is not a single event; it is a complex process. Although numerous researchers have depicted data-driven decision-making cycles ranging from four to eight steps, common elements include (1) determining purposes, (2) collecting data, (3) analyzing data, (4) interpreting data, (5) enacting change, and (6) monitoring and assessing impact. This creates a continuous cycle of data-driven decision making.

Determining purpose

Determining the purpose requires educators to identify evaluation questions that can lead to meaningful change. These questions might relate to systems, curricula, instruction, or individual students.

Collecting data

Educators must then identify existing data sources or collect data to answer their driving questions. Many common data sources are readily available to educators, including graduation and dropout data, attendance, state assessment scores, progress reports, transcripts, and discipline data. Less common sources of data might include post–school outcomes data, student reflective writing samples, and parent feedback.

The data collection process includes discerning which data will be most useful for answering the driving questions. Preferably, multiple sources of data are used for decision making. This helps ensure that incorrect decisions aren't made. For instance, one poor test grade may mean that a student was tired rather than meaning that the student needs intervention to master the content. Conversely, a poor test grade, incomplete assignments, and a rating of emerging on a performance-based assessment might indicate that an intervention is needed.

It's important to recognize that educators don't always have complete or accurate data. For instance, when collecting post–school outcomes data, it is sometimes difficult to reach all students. This can distort the data and limit interpretation. Educators may need to strategize ways to obtain accurate and reliable data.

Analyzing data

Once collected, raw data still have to be organized in a way that yields useful information. Information doesn't become actionable knowledge until individuals or teams synthesize the data, apply judgments to prioritize it, and consider the relative merits of possible actions. Relaying information to the people who need it, in a format that they can use it, is critical. Visual displays can guide data interpretation. Graphics that present data over time can identify trends or patterns and assist individuals in making connections between various data. For instance, looking at displays of least restrictive environment data next to academic achievement data can illustrate how inclusion relates to learning for students with and without disabilities.

Interpreting data

Discussions can lead to more nuanced interpretations of data. Asking individuals with diverse views to consider data and pose questions or hypotheses can result in insights not seen by an individual. Using guided data questions directs discussions by focusing on specific indicators that inform priorities and can result in actionable change. These questions might include what do you see in these data? What you don’t see in these data? What questions arise? What do the data tell us? What can we learn from the data? How can we change our practice in light of the data? How do these data relate to other data? What other data do we need to collect? What are the contextual or influencing factors? Is there a root cause for the identified issues or problems? This interpretation process transforms data from information to actionable knowledge that can inform decisions.

Enacting change

In data-driven decision making, the analysis must result in action. When targeted improvements are identified, educators can prioritize their resources and professional development efforts to make programs and practices more efficient and productive. Enacting change may require high-quality professional development and the application of the new knowledge and skills associated with the chosen evidence-based strategy.

Monitoring progress and assessing impact

Once change is enacted, additional data should be collected to assess the effectiveness of those actions. The implementation of a new practice requires communication, collaboration, practice, feedback, and coaching; therefore, teacher adoption and implementation should be monitored along with progress in student outcomes. This creates a continuous cycle of collection, organization, and synthesis of data to support decision making.

Data-driven decision making is a systematic and ongoing process of data collection, interpretation, planning, implementation, and progress monitoring. In education, data-driven decision making has many purposes, including inquiry, evaluation, instructional or student feedback, and student monitoring. Although a variety of quantitative and qualitative data are available, in order to inform decision making, educators must ensure that the data are accurate and accessible. Therefore, educational professionals require the skills to analyze, interpret, and display data and must share these data with their colleagues. “Admiring the problem” is not enough; educators must enact changes and continually monitor progress. When all of these elements are in place, data have the capacity to drive powerful, positive transformations for students, educators, and schools.

Amy S. Gaumer Erickson and Patricia M. Noonan

See also Causal Inference; CIPP Evaluation Model; Classroom Observations; Problem Solving; Progress Monitoring

Further Readings

Corrigan, M. W., Grove, D., & Vincent, P. F. (2011). Multi-dimensional education: A common sense approach to data-driven thinking. Thousand Oaks, CA: Corwin.

James, E. A., Milenkiewicz, M. T., & Bucknam, A. J. (2008). Participatory action research for educational leadership: Using data-driven decision making to improve schools. Thousand Oaks, CA: Sage.

Kowalski, T. J., & Lasley, T. J. (2009). Handbook of data-based decision making in education. New York, NY: Routledge.

Kowalski, T. J., Lasley, T. J., & Mahoney, J. W. (2008). Data-driven decisions and school leadership: Best practices for school improvement. New York, NY: Pearson.

Mandinach, E. B., & Honey, M. (2008). Data-driven school improvement: Linking data and learning. New York, NY: Teachers College Press.

Mandinach, E. B., & Jackson, S. S. (2012). Transforming teaching and learning through data-driven decision making. Thousand Oaks, CA: Corwin.

Schildkamp, K., Lai, M. K., & Earl, L. (2013). Data-based decision making in education: Challenges and opportunities. New York, NY: Springer.

Sharratt, L., & Fullan, M. (2012). Putting FACES on the data: What great leaders do! Thousand Oaks, CA: Corwin.

Robert D. Ridge Robert D. Ridge Ridge, Robert D.

Debriefing

Debriefing

466

468

Debriefing

Debriefing is the process of explaining to research participants the general purpose of the research in which they have participated and answering questions they may have. It involves providing participants with the opportunity to obtain information about the nature of the study and the manipulations that were employed and correcting any misconceptions they may have. This entry looks at the reasons for debriefing and the steps taken during debriefing.

Both 45 CFR part 46, which are federal regulations dealing with the protection of human subjects, and the American Psychological Association require a debriefing when deception, either in the form of providing misleading information or failing to provide complete information, has been employed in a study. Although a debriefing is typically undertaken in a face-to-face interaction between the researcher and the participant, written debriefings may be employed in research that is conducted online. A thorough debriefing allows the investigator to identify any suspicions a participant may have about the research, including correctly guessing the hypotheses under investigation, as well as to explain any deception that was employed and ensure that the participant exits the research feeling as well as she or he did when entering.

When deception has been employed in a study, the careful researcher will want to determine whether participants have “figured out” the true purpose of the research and correctly guessed the researcher’s hypotheses. The debriefing allows the researcher to begin by asking participants if they have heard about the study from others or if they have suspicions about how the study was represented to them. The researcher may also query participants about specific manipulations that were critical parts of the methodology to determine whether manipulations that were critical parts of the methodology to determine whether they were understood as intended.

Participants may be invited to ask questions about the research and to assess their feelings about and reactions to their participation. This portion of the debriefing provides the researcher with important information that may dictate whether the data provided by the participant are useful. If the participant has not deduced the true purpose of the research and did not fail to be misled by the deception, then the data are most likely free of subject reactivity and may be included in the data set for subsequent analysis.

After assessing participants' perceptions of the study, the debriefing then discloses the true purpose of the research, including the hypotheses under investigation and any deception that was employed, and explains why deception was necessary. Deception is typically employed when the researcher is concerned that participants will not respond truthfully in the research situation. For example, if the researcher is interested in behavioral manifestations of prejudice and discloses this to the participant, it is likely that the participant will behave in a way so as to engender a favorable impression, that is, in a way to appear nonprejudiced. The researcher must mislead the participant about the true purpose of the study so that the participant will not try to manage an impression. This may be the only way to obtain valid information about prejudicial behavior.

When deception is used in research, the researcher is obligated to disclose any deceptive information that was provided to the participant or to provide any information that was withheld. Deliberately deceptive information would include being told that the investigator was measuring one behavior, when in fact he or she was measuring another. Withheld or incomplete information would be when the investigator fails to tell the participant that the purpose of the study is to investigate prejudice.

A critical aspect of a debriefing is to ensure that participants feel good about their participation and exit the study feeling as well as they did when they entered. This may be challenging when the study has investigated socially undesirable behaviors, such as prejudice or aggression. A careful debriefing respects the participant by explaining the necessity of using deception to circumvent people's tendency to want to look good for an experimenter.

During the debriefing, the researcher assures the participant that the study was carefully designed and executed to obscure the true purpose of the research and that the participant is not extraordinary or gullible for having believed the cover that the participant is not extraordinary or gullible for having believed the cover story. The researcher assures the participant that the typical response of participants is to believe what they are told and to act accordingly. In addition, the researcher reminds participants that their data are confidential and will be aggregated with others’ data so that they will not be individually identified. Finally, the researcher allows participants to withdraw their data if they feel that they would not have consented to participating in the study had they known what it was actually about.

There are rare exceptions to the requirement that a debriefing discloses all deceptive or withheld information from participants. If the debriefing would pose more risk to the participant than not making a full disclosure, then some information may be excluded. For example, if an investigator selected a person to participate in a study on the basis of others’ judgments that the person is dull, such information may cause more distress to the participant than not knowing the inclusion criteria. A risk–benefit analysis that is required to conduct research with human participants would most likely conclude that the benefit of disclosing this information would not exceed the risk to the participant’s self-esteem, so the ethical decision would be to exclude this information from the debriefing.

A careful and thorough debriefing fulfills the first principle, respect for persons, articulated in the Belmont Report, a document that explains the ethical principles that govern research with human participants. Participants in research must provide informed consent before participating. When deception is employed, true informed consent is not possible. A debriefing respects participants by giving them autonomy over their data and allowing them to decide how their data will be used. It demonstrates concern for them and their welfare, minimizes harm, and reflects professional and ethical behavior by the researcher.

Robert D. Ridge

See also Belmont Report; Deception in Human Subjects Research; Ethical Issues in Educational Research; 45 CFR Part 46; Informed Consent; Institutional Review Boards

Further Readings

American Psychological Association. (2010). Ethical principles of psychologists and code of conduct, including 2010 amendments. Retrieved July 11, 2016, from http://www.apa.org/ethics/code/index.aspx

Code of Federal Regulations, Title 45, Part 46: Protection of Human Subjects. Retrieved July 11, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April). Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Retrieved July 11, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/

Aldert Vrij Aldert Vrij Vrij, Aldert

Deception

Deception

468

471

Deception

Lying, a deliberate attempt to mislead others, is a two-pronged activity: It can be a social lubricant or a selfish act. Educators and educational researchers are interested in the practice and detection of social deception in both children and adults. To identify lies, it is important to examine how liars respond and how they can be detected. After a further examination of the basics of deception and its frequency, this entry reviews the four ways the lie detection tools aim to distinguish truth tellers from liars—by evaluating their nonverbal behavior, speech content, physiological activity, and brain activity. The entry concludes by looking at three techniques interviewers can use to elicit cues of deceit.

Detecting lies by examining nonverbal behavior is popular among practitioners, but there is no evidence that it actually works. Speech content can differentiate truth tellers from liars as long as the correct veracity tool is used. Physiological responses, measured with a polygraph, distinguish truth tellers from liars, at least in laboratory settings. How they fare in real life is unknown due to a lack of reliable data. Depending on the interview protocol used, there is a tendency to elicit false-positive or false-negative errors. Measuring brain activity is intrusive, expensive, and time-consuming and therefore not well suited for lie detection in applied settings. However, such measurements give valuable insight into what happens in the liar’s brain. If cues to deceit are faint and unreliable, perhaps investigators can elicit cues that liars spontaneously do not seem to show. Two verbal veracity assessment tools are based on this assumption.

Frequency of Lying and Types of Lies

Lying is a frequently occurring event. On average, people lie in one out of every four of their social interactions that lasts longer than 10 minutes. The lies people tell are outright (i.e., the information conveyed is completely different from what the liar believes to be the truth), exaggerations (i.e., over-or understated facts), and concealments (i.e., omitting relevant details). People lie to gain material advantage or to avoid materialistic loss or punishment. Such lies are often selfish, disruptive of social life, and hurtful to the targets. People also lie for psychological reasons, often to protect themselves from embarrassment (people don't want to reveal all their inadequacies, errors, and indecent and immoral thoughts), to avoid tension and conflict in social interactions, and to minimize hurt feelings and ill will (people don't want to tell the bold truth, thereby deliberately hurting the feelings of a good friend). Psychological lies act as a social lubricant and improve social relationships. The nature of lying is therefore two-pronged: it can be a selfish act or a social lubricant. Being able to detect selfish lies would benefit individuals or the society as a whole. To detect selfish lies, it is relevant to know how liars respond and how they possibly could be detected.

Lie Detection

Lie detection tools aim to distinguish truth tellers and liars based on their nonverbal behavior (e.g., gaze patterns, emotional expressions, posture), speech content (e.g., amount of detail, type of detail, plausibility, inconsistency), physiological activity (e.g., heart rate, blood pressure, galvanic skin response), and brain activity (e.g., P300 brain waves, cortex activity). All four approaches share a common element: A response uniquely associated to deception—a response always present during lying and never during truth telling akin to Pinocchio's growing nose—does not exist. In addition, the cues that distinguish truth tellers from liars are faint and unreliable. The diagnostic value of the most revealing verbal or nonverbal cue to deceit is the equivalent of the difference in height between 15-and 16-year-old girls.

Nonverbal Behavior

Analyzing nonverbal behavior to detect deceit is popular among practitioners. The assumption is that liars are more nervous than truth tellers and therefore will display more nervous behaviors. Nonverbal lie detection tools have not proven to be successful. There is no empirical evidence available showing that observing facial expressions and involuntarily body language (the method used by the fictitious character Dr. Cal Lightman in the TV series *Lie to Me*) actually works; neither is there evidence that the Behavior Analysis Interview can successfully distinguish truth tellers from liars. Police interview manuals tend to focus more attention on nonverbal behavior than on speech content when assessing credibility. They justify this by claiming that most of a message communicated between persons occurs at a nonverbal level. This claim is taken out of context and misleading. It is based on research where interviewees were requested to express their emotions in single words ("dear" or "terrible") and cannot be applied to police interviews or other interviews, where interviewees say considerably more than a single word.

Speech Content

Two verbal veracity assessment tools are frequently used: Statement Validity Assessment (SVA) and Scientific Content Analysis (SCAN). SVA was developed in Germany and Sweden to assess the credibility of child witnesses' testimonies in trials for sexual offenses. It is often difficult to determine the facts in an allegation of sexual abuse since often there is no medical or physical evidence, hence the desire to design a verbal veracity tool. SVA assessments are accepted as evidence in some North American courts and in criminal courts in several Western European countries, including Austria, Germany, Sweden, Switzerland, and the Netherlands. The key part of SVA is the criteria-based content analysis in which experts assess transcripts of the interviews with children and examine the presence of 19 verbal criteria. Criteria-based content analysis experts believe that each of the 19 criteria is more likely to occur in truthful than in deceptive statements because liars may lack the imagination to make up detailed stories or because they leave out certain details because they fear these details look suspicious. Twenty-five laboratory studies, typically with adult participants, revealed an accuracy rate of 70% in making truth/lie decisions based on criteria-based content analysis scores. The accuracy rate in real cases is unknown because in such cases the ground truth (is the interviewee actually telling the truth or lying?) is often unknown.

SCAN is very popular among practitioners and is used worldwide. In the SCAN procedure, examinees are requested to write down in their own words their activities during a certain period of time. SCAN experts assess the statement and examine the presence of around 12 criteria; some of them are supposed to indicate truthfulness and some of them deception. No theoretical rationale is indicate truthfulness and some of them deception. No theoretical rationale is given as to why truth tellers and liars would differ on the SCAN criteria. There is no empirical evidence available that SCAN actually works.

Physiological Responses

Physiological responses are measured with a machine called the polygraph. A polygraph is also referred to as a lie detector, but that term is misleading, as the machine cannot detect lies. It can only detect physiological arousal. Several polygraph interview protocols have been developed, and the most popular technique is the control question test, sometimes referred to as the comparison question test. In the test, levels of arousal in response to control questions and relevant questions are compared. The assumption is that only liars will display stronger responses to the relevant than the control questions. Laboratory studies show not only high accuracy rates (up to 80%) but also a bias to classify truth tellers as liars (false-positive error). Accuracy rates in the field are unknown due to a lack of ground truth in such cases (i.e., uncertainty of whether the examinee is actually telling the truth or lying). The empirical evidence suggests that false-positive errors are more prominent in the field than in the laboratory.

A second polygraph test is the Concealed Information Test. The test is used when examinees deny specific knowledge about the crime. During the test, multiple-choice questions are asked with at least four answer alternatives, of which one alternative is the correct answer. Examinees who actually know the answer are expected to show an orienting response (picked up by the polygraph sensors) when the correct alternative is mentioned; innocent examinees are expected to display the same responses for all answer alternatives. In laboratory settings, the Concealed Information Test is very accurate in classifying truth tellers (over 90% accuracy) but somewhat weaker in classifying liars (around 80% accuracy). Field studies suggest a relative weakness in identifying liars, the so-called false-negative errors. Such errors occur if the guilty examinee does not recognize the correct answer. This could happen because the guilty examinee has forgotten the correct answer or it may be due to poor questioning whereby questions are asked about minor details to which examinees typically do not pay much attention and therefore do not recognize.

Brain Activity

Brain activity is measured via EEGs and fMRI scans. Such measurements are intrusive, expensive, and time-consuming, and they therefore cannot be easily applied in field settings. The benefit of such brain measurements is that they analyze deception more directly because lying is a mental (brain) activity. As such, brain analyses enhance our understanding of what actually happens when people lie. fMRI research has, for example, shown that lying is cognitively more difficult than truth telling and that lying is associated with increased brain activity, particularly in areas that indicate inhibition, monitoring, and executive processes.

Eliciting Cues to Deceit

It has been argued that, if cues to deceit are faint and unreliable, interviewers perhaps could elicit such cues through the employment of specific interview protocols. Two (both verbal) veracity assessment tools have been developed based on this premise. The Strategic Use of Evidence considers inconsistencies between pieces of evidence that are known to the interviewer and the interviewee’s statement. The Strategic Use of Evidence rationale is that in interviews truth tellers are forthcoming, whereas liars do not wish to be linked to incriminating evidence and thereby use an “avoid and escape” strategy. A key of the approach is to initially withhold presenting the evidence to the interviewee (e.g., closed-circuit television footage showing the examinee’s car close to the crime scene around the time of the crime) while trying to make the interviewee talk about the evidence (e.g., use of the car by the examinee or others on that particular day). Liars are then more likely than truth tellers to provide a statement that contradicts the evidence.

The rationale of the cognitive lie detection technique is that certain instructions can be more difficult to follow for liars than truth tellers. The technique comprises three key elements. First, interviewers can exploit the fact that in interviews lying is typically more difficult than truth telling by making the interview setting more difficult through “imposing cognitive load” requests. Liars will have fewer cognitive resources left over to deal with such additional requests. For example, lie detection improves when interviewees are asked to recall their story in reverse order—a difficult task—than when they are asked to recall their stories in chronological order.

Second, interviewers can encourage interviewees to say more, for example, by using a model statement of a detailed response. A model statement gives the interviewee a good idea of how much detail is required. Methods that encourage interviewees to say more result in truth tellers providing more detail than liars, as liars may lack the necessary imagination or creativity to add the same amount of detail as truth tellers or may be reluctant to say more out of fear that any additional information will expose their lies to investigators.

Third, interviewers can ask unexpected questions. Liars prepare themselves for interviews by thinking of answers to possible questions but face difficulty when they are posed with questions they did not know would be asked. Investigators can exploit this by asking a mixture of questions that liars have expected and questions that they have not expected but that make perfect sense in the given context, such as spatial questions (e.g., “Where did you and your friend sit in the restaurant?”) and questions about the planning of activities. Typically, truth tellers and liars provide the same amount of detail when answering expected questions, but liars are less detailed than truth tellers when answering unexpected questions. A meta-analysis of the cognitive lie detection approach, including 14 studies, revealed a superior accuracy rate in cognitive load interviews (71%) than in standard interviews (56%).

Aldert Vrij

See also Interviews

Further Readings

Granhag, P. A., Vrij, A., & Verschuere, B. (Eds.). (2015) Detection deception: Current challenges and cognitive approaches. Chichester, UK: Wiley.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. Psychological Bulletin, 137, 643–659. doi:10.1037/a0023589

Kleiner, M. (Ed.). (2003). Handbook of polygraph testing (pp. 251–264). San Diego, CA: Academic Press.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (Eds.). (2014). Credibility assessment: Scientific research and applications. Oxford, UK: Academic Press.

Vrij, A., Leal, S., Mann, S., Vernham, Z., & Brankaert, F. (2015). Translating theory into practice: Evaluating a cognitive lie detection training workshop. Journal of Applied Research in Memory and Cognition, 4, 110–120. doi:10.1016/j.jarmac.2015.02.002

Stephanie Dyson Elms Stephanie Dyson Elms Elms, Stephanie Dyson

Deception in Human Subjects Research Deception in human subjects research

471

474

Deception in Human Subjects Research

This entry examines the use of deception in human subjects research, including the ethical framework in which it is used, regulations and standards governing the use of deception in research, and criticism over the use of deception in two well-known studies. Research studies commonly use deception to control the environment or to minimize potential bias in individuals participating in human research. Generally, deception is defined as providing false or incomplete information to a participant.

When used ethically, deception can protect the integrity of the research while respecting the participant. The practice of using deception in research remains controversial; in fact, some of the most notorious studies in human subjects research history have led directly to increased regulatory requirements because of the questionable ethics of the researchers’ deception.

Although some researchers may view deception as a fully codified procedure, in actuality, there are ethical codes, regulations, and local policies that combine to create a somewhat fragmented standard. The regulation for human subjects research in the United States, called the Common Rule, doesn’t mention deception. This leaves institutional review boards (IRBs), regulators, and investigators to puzzle out the standard practices for conducting an ethical deception study.

The impact of ethical issues must be considered not only because they can affect the validity of a study but also to minimize any potential harms to participants. Respecting the autonomy of participants is a vital ethical consideration. Guidelines in place for conducting studies involving deception make clear the importance of assessing and minimizing risk to participants and having a clear justification for using deception in the research.

justification for using deception in the research.

Defining Deception in Research

There is not a single regulatory definition for deception, but there is a general consensus that it falls into two broad categories: deception and incomplete disclosure. Deception occurs when a participant is deliberately given false information about the research, which includes false feedback, use of confederates, or other fabricated information about the procedures to be followed. Social psychology experiments use deception to study a host of different aspects of human behavior, such as participants taking intelligence tests and being told they did poorly.

Incomplete disclosure is referred to as omission or passive deception and may be more ethically permissible than outright deception or false feedback. Research studies that do not give complete information about the nature of the research to the participants are studies of incomplete disclosure—often some aspect of the research design is not revealed. An example of this may give participants vague information about a study and not specifying the true purpose—participants may be told the study is about social interactions of Americans when the true purpose is studying racism.

The Ethical Framework

The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, published in 1979, provides the ethical underpinnings for the federal Common Rule. It includes three ethical concepts that are meant to be building blocks for conducting human subjects research. The Belmont principles are beneficence, respect for persons, and justice, with the idea of respect for persons highlighted in deception research.

Respect for persons says that researchers should acknowledge the autonomy of individuals, or their ability to understand risks and benefits and make decisions for themselves, according to their own values. Under this concept, participants must be free from any coercion or undue influence and be fully informed about the research they are participating in, so that they can make a decision with complete information. The *Belmont Report* also states that information should not be withheld from participants "when there are no compelling reasons to do so" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979, n.p.). These ideas require a balance between protecting a person's autonomy and being able to conduct scientifically valid research, which is the heart of the deception dilemma.

Regulations and Standards for Deception Research

The regulations at 45 CFR part 46, which include the Common Rule, do not explicitly define or reference deception, but concepts of risk and informed consent, which the regulations stipulate, create a structure on which deception research methods are based. Calculating risk, from the context of the regulations, is to analyze the balance of probability and magnitude of harm. Probability is the likelihood that an adverse event may occur, while magnitude is the degree of the risk or harm.

When evaluating the potential risk, a matrix of probability to magnitude assists in determining the overall potential risk to participants. An example of high probability and low magnitude risk may be that most people participating in a deception research study would be a little embarrassed because they believed the cover story. Conversely, a study with a high magnitude of potential risk might be a study that asks students about past sexual trauma(s). There may be only a few participants who have an adverse reaction to this line of questioning, but for some individuals who have experienced sexual trauma, the magnitude of the risk may be very high. The overall potential for risk, in conjunction with deception, for a small group of participants who may not have been fully informed about the nature of the study or the types of questions increases greatly.

45 CFR 46.111(a)1 states that: "Risks to subjects are minimized by using procedures which are consistent with sound research design and which do not unnecessarily expose subjects to risk." These regulatory criteria introduce the idea of minimizing potential risks by considering the research design of a study while not exposing participants to potential risks needlessly. Researchers and regulators must consider the overall amount of risk to participants and work to ensure that those risks are justified and that the study will produce meaningful results. A justification is required for deception research to ensure that all aspects of the study, including alternatives, have been thoughtfully designed to ensure that participants encounter the lowest level of risk while maintaining scientific viability.

The ethical principle of respect for persons, or autonomy, supports the requirements for informed consent, which is the second part of the deception equation. Most regulatory bodies require informed consent for research, including the OHRP and the Federal Drug Administration, as well as ethics statements such as the Declaration of Helsinki and the Nuremberg Code. The full requirements for informed consent are stipulated in the Common Rule and include details such as fully explaining the purpose and procedures, that participation is voluntary, and potential risks to the subject. These conditions make it difficult to conduct deception research at any level; however, the regulations do include language that allows a full or partial waiver of informed consent in certain circumstances. This waiver allows researchers to conduct deception studies, despite the requirement for informed consent.

The conditions for the waiver of informed consent require an explanation of how the research involves no more than minimal risk, which is the risk a person may experience in everyday life; that the rights and welfare of the participant are not adversely affected; and that the research could not be practicably carried out—meaning that it is more than a matter of convenience. The last requirement is that participants are provided with more information about the study afterward, which is generally the debriefing procedure.

Every study must satisfy these requirements in order for the IRB to grant the waiver of informed consent and add additional requirements appropriate when waiving informed consent. It is important to note that a majority of the time, this waiver will not waive the entire informed consent requirement but only allow enough deviation from the true purpose or procedure to accommodate the deception—participants must still be informed as much as possible.

The debriefing procedure does not have any regulatory guidelines for content but should explain how the study deceived participants, the true purpose of the research, and why the deception was necessary. Many IRBs will have their own guidance or requirements for the debriefing, often incorporating the Common Rule requirements as well as standards from the American Psychological Association (APA).

The APA sets standards for the psychological field and other fields and provides additional, limited guidance for deception and debriefing. The APA requires that the study have “significant value”; that the deception cannot cause severe emotional stress or physical pain; and that researchers explain any deception as early as possible. The standard also stipulates that participants must be allowed early as possible. The standard also stipulates that participants must be allowed to withdraw their consent after they have been informed of the deception. The APA has criteria that require a debriefing for all research, including deception research, with additional provisions to minimize harm to subjects. IRBs at different types of institutions, including universities and hospitals, often impose additional restrictions or requirements on deception research, so it is important to verify requirements with specific institutions before conducting deception research.

Issues in Deception Research

There has always been some controversy about the ethics of conducting deception research, as well as its effectiveness. Many early studies pushed ethical boundaries and incurred public disapproval, which eventually led to the creation of regulations and standards on the use of deception in research. Nonetheless, deception research continues to capture public attention on a semiregular basis.

One of the main concerns with deception research is that of the debriefing procedure. While a significant amount of this research occurs across university campuses in a laboratory setting, deception research is also conducted outside of labs, and with increasing frequency through the Internet, which can negatively impact the ability to debrief a subject. As discussed earlier, an analysis of the regulations requires that the deception research be no greater than minimal risk; however, it is important to provide correct information or clear up any misunderstandings participants may have had about the research.

Although students can simply be debriefed before leaving the laboratory, debriefing anonymous subjects who participate in a study through the Internet can pose significant challenges. It is also important to consider the population involved in the research—will children, the elderly, or those with developmental delays have the same ability to be debriefed as the average social psychology research participant?

The research community continues to struggle with deception research and the ethics of conducting such research. Some investigators will label their research procedures as “mild” deception in an attempt to underscore the minimal risk nature of the study; however, using subjective descriptions of procedures can be less useful to a reviewer than a proper justification of the deception in a protocol. The lack of universal standards makes it difficult to apply objective terms to this

The lack of universal standards makes it difficult to apply objective terms to this research, which serves to underscore the importance of an effective justification for deception.

Use of Deception in Milgram and Facebook Studies

A seminal case of research involving the use of deception is the 1963 Stanley Milgram obedience study, which contained many different facets of deception, including a cover story, false feedback, and a confederate to further the deception. The study was presented as a teacher–learner experiment, while actually looking at participants' obedience to authority, where participants acted as "teachers" to unidentified "learners" and believed they were employing electric shocks to the point of death on these confederates. This study was conducted in the 1960s, before the implementation of the OHRP regulations, but plainly illustrates the need for ethical considerations in the pursuit of science.

More recently, the 2012 Facebook emotional contagion study, which manipulated the newsfeed of nearly 700,000 members without their knowledge, similarly received so much criticism in regard to the ethical considerations that the scientific merit of the study was called into question.

Stephanie Dyson Elms

See also Belmont Report; Debriefing; Deception; Declaration of Helsinki; 45 CFR Part 46; Informed Consent; Nuremberg Code

Further Readings

American Psychological Association. (1973). Ethical principles in the conduct of research with human participants. American Psychological Association.

Bankert, E. A., & Amdur, R. J. (2006). Institutional review board: Management and function. Jones … Bartlett Learning.

Criteria for IRB Approval of Research, 45 C.F.R. § 46.111 (2009). Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788–8790.

Milgram, S. (1963). Behavioral study of obedience. The Journal of Abnormal and Social Psychology, 67(4), 371.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Retrieved May 15, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Pascual-Leone, A., Singh, T., & Scoboria, A. (2010). Using deception ethically: Practical research guidelines for researchers and reviewers. Canadian Psychology/Psychologie Canadienne, 51(4), 241.

S. Earl Irving S. Earl Irving Irving, S. Earl

Decile

Decile

474

475

Decile

As with all quantiles, a decile is a rank-order statistic. A decile divides a distribution of data into 10 equally sized groups determined after ranking the data according to some measure or combination of measures. After offering two meanings for the term *decile*, this entry provides an explanation of how decile values can be utilized in educational settings, including an example of its use in New Zealand.

The most common quantiles are percentiles, quartiles, deciles, and sometimes quintiles. Percentiles offer too many groups to visualize and to make sense of each group, but deciles provide a fairly large number of groups that can still be treated at one time while still allowing the analyst to deal with each of them as separate entities.

The term decile can have two meanings—a point on the distribution that divides two of the 10 groups and the group to which a member of the distribution belongs. There are nine decile values that separate the data into the 10 groups. The ninth decile value is the point in the distribution where nine tenths of the data lies below that value. The second decile is the point below which two tenths of the distribution lies. If a point on the distribution lies between the second and third decile, then it is a member of Decile 3. If data are normally distributed, then *areas under the normal curve* tables can be used to identify deciles. For example, normally the mean is the fifth decile and the sixth decile is around a *Z* of .25 or .25 standard deviations above the mean.

When an analyst breaks data into quantile groups, the purpose is to simplify the way in which the analyst can visualize the data to look for patterns and trends or to compare and contrast high-performing and low-performing groups. Deciles are used in the real-world decision making in a number of ways. For example, they have been used for equity funding of state and state-integrated schools in compulsory education in New Zealand. Unlike other jurisdictions overseas, schools in New Zealand are not permitted to collect family demographic data such as parental qualifications or income in order to assess the socioeconomic background of students attending the school.

For each school, the home address of students is extracted and compared with mesh block data from the most recent 5-year census on five indicators: household income, occupation of employed adults, household crowding, educational qualifications, and income support (i.e., receipt of social welfare support). These indicators are combined, each school is ranked, and 10 groups with approximately equal numbers of schools are obtained.

Decile 1 schools have the highest proportion of students from low-income communities, and Decile 10 schools have the lowest proportion of these students. Although there are some funding components that are equally applied to all schools and students (e.g., the basic per capita grant), differential funding on other specific targeted equity components is applied to enable schools to address the challenges that are faced by schools in low-income communities.

S. Earl Irving

See also Box Plot; Descriptive Statistics; Percentile Rank; Quartile

Further Readings

Hald, A. (1998). A history of mathematical statistics from 1750 to 1930. New York, NY: Wiley.

Walker, H. M. (1929). Studies in the history of statistical method. Baltimore, MD: Williams and Wilkins.

Sara Tomek Sara Tomek Tomek, Sara

Decision Consistency

Decision consistency

475

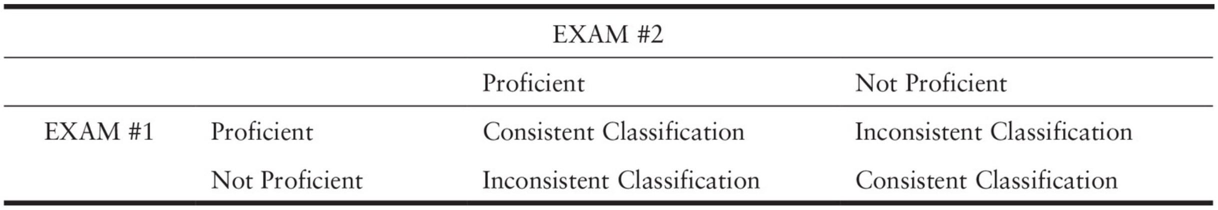
477

Decision Consistency

The term *decision consistency* refers to the measure of reliability of a test decision across either multiple forms of a single test or repeated administrations of identical tests. This measurement is similar to that of decision accuracy, though their purposes are different. Although decision consistency measures the consistency of multiple decisions for an examinee on an exam, decision accuracy refers to the correct classification of an examinee using the exam based on the examinee's true classification status. That is, decision consistency measures the reliability of an exam across multiple occurrences, whereas decision accuracy measures reliability within a single occurrence. For example, decision accuracy would be focused on measuring whether an end-of-year exam would be accurate in determining a proficient rating given that the student taking the exam has gained sufficient knowledge during the school year. In contrast, decision consistency would measure whether the end-of-year exam would consistently rate the student as proficient across multiple administrations of the exam. This entry defines decision consistency in a general form, discusses factors that may affect decision consistency, as well as discusses measures that have been developed to calculate decision consistency.

What Is Decision Consistency?

By definition, decision consistency refers to the accuracy in the classification of individual examinees on multiple forms or administrations of an exam. If we look at a simple example, end-of-year exams required by students in elementary and secondary education, the exam is focused on measuring whether the student is proficient or not proficient in different subject areas. Over each form or administration of the exam, two options are possible: either the examinee will be classified as proficient or the examinee will be classified as not proficient. This will create a 2 × 2 decision table of possible outcomes across the two exams (see Table 1).



		EXAM #2	
		Proficient	Not Proficient
EXAM #1	Proficient	Consistent Classification	Inconsistent Classification
	Not Proficient	Inconsistent Classification	Consistent Classification

Of interest in decision consistency is to measure the probability of a consistent classification. The distinction between decision consistency and other reliability measures for an exam (e.g., test–retest reliability, internal consistency) is that decision consistency assumes that the reliability of the scores themselves are not as important as the final decision. That is, the consistency in the items and the resulting test score is not being measured; rather, the final decision of the model to classify examinees into their categories (e.g., proficient vs. not proficient) is. Although there may be variation in the final scores, the probability of consistency of the final decision is being calculated.

The relationship between decision consistency and decision accuracy can be seen by replacing the second exam outcomes by the expected decision of the examinee of the exam based on the true score of the examinee. By looking at the relationship between the expected and observed outcomes, the probability of a consistent classification will estimate the decision accuracy of the exam.

Factors Affecting Decision Consistency

Multiple test characteristics may affect measures of decision consistency. For example, exams with more items may lead to better discrimination, which will increase decision consistency. The location of the cut point for examinee decision categories will also greatly impact the consistency of decisions. Cut points that are located closer to the center of the examinee distribution tend to have lower decision consistency. The greater the generalizability of the test scores, the greater decision consistency the exam will generate. And, as a last example, the greater the similarity between the examinee distributions of the two exams that are being compared, the greater the measure of decision consistency.

Measuring Decision Consistency

In the 1970s, H. Swaminathan, Ronald K. Hambleton, and James Algina introduced the concept of decision consistency within the context of classical test theory. Since that time, the measurement of decision consistency has been developed to span multiple different types of tests as well as multiple different modeling frameworks.

The simplest measure of decision consistency is that of the index *p*, which is the probability, or proportion, of examinees that are consistently classified into the same category across multiple administrations of an identical exam. However, differing indices have been developed that take into account different aspects of the exam and the decision measure. For example, the Livingston coefficient and Brennan–Kane index take into account the severity of certain misclassifications, whereas the index *p* assumes all misclassifications are equal in their severity. In addition, Quinn Lathrop and Ying Cheng outline a nonparametric method for measuring decision consistency, Ying Cui, Mark J. Gierl, and Hua-Hua Chang developed a decision consistency index for cognitive diagnosis model, and Tim Moses and Sooyeon Kim outline methods for calculating decision consistency when exams are of a mixed format. These are examples of a subset of measures used for decision consistency.

Some types of exams do not have an inherent ability to utilize one of these measures. For example, it is difficult to administer parallel forms of a criterion-referenced test; therefore, testing decision consistency may be limited. In response, methods have been developed to obtain estimates of decision consistency within a single administration. For example, Nina Deng and Ronald K. Hambleton outline this approach in the context of classical test theory and item response theory.

In practice, if the purpose of the test is to classify individuals, indices of decision consistency should be reported.

Sara Tomek

See also Classical Test Theory; Cognitive Diagnosis; Cut Scores; Item Response Theory; Proficiency Levels in Language; Reliability

Further Readings

Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (2006). Introduction to classical and modern test theory. Mason, OH: Cengage Learning.

Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. Journal of Educational Measurement, 49, 19–38. doi:10.1111/j.1745-3984.2011.00158.x

Deng, N., & Hambleton, R. K. (2014). Evaluating CTT-and IRT-based single-administration estimates of classification consistency and accuracy. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), New developments in quantitative psychology (pp. 235–250). New York, NY: Springer.

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. Journal of Educational Measurement, 51, 318–334. doi:10.1111/jedm.12048

Livingston, S. A., & Lewis, C. (2005). Estimating the consistency and accuracy of classification based on test scores. Journal of Educational Measurement, 32, 179–197. doi:10.1111/j.1745-3984.1995.tb00462.x

Moses, T., & Kim, S. (2014). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed-format licensure tests. Applied Psychological Measurement, 39, 314–329. https://doi.org/10.1177/0146621614563067

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 11, 263–267. doi:10.1111/j.1745-3984.1974.tb00998.x

Søren Holm Søren Holm Holm, Søren

Declaration of Helsinki Declaration of helsinki

477

478

Declaration of Helsinki

The Declaration of Helsinki was issued by the World Medical Association in 1964 and has since been revised several times. It is a foundational document in medical research ethics, but its requirements for informed consent and review of research projects by an independent committee have profoundly influenced research ethics and regulation outside the biomedical field, including in many fields of social science research. This entry looks at the development of the Declaration of Helsinki, its main principles, objections to its requirements, and the incompatibility of the declaration with certain practices used in education research.

The declaration came into being because of a number of papers and books appearing in the early 1960s, showing that medical researchers were often using vulnerable patients in risky research projects without consent. An American military tribunal in Nuremberg issued the Nuremberg Code in 1947 as part of the judgment in the so-called Doctors’ Trial where Nazi doctors and administrators who participated in war crimes and crimes against humanity in the Nazi concentration camps were prosecuted. Some of the crimes committed involved medical research on concentration camp prisoners.

The code required consent from research participants as a necessary condition for research to proceed, but in the early 1960s, it became evident that it was not generally followed and the World Medical Association decided to issue the Helsinki Declaration to improve the ethical standard of medical research. The 1964 declaration and its subsequent revisions have had significant impact on the development of research ethics because it was the first widely accepted international set of rules and principles for medical research involving human subjects. Principles first enunciated in the declaration have since been implemented in national laws and in binding international guidelines implemented in national laws and in binding international guidelines.

The following are the four main principles of the declaration:

1. The interests of society and science can never outweigh the interests of the participants.
2. Research participants can only participate in research if they have given their voluntary fully informed consent. If research participants are unable to consent because of lack of legal competence, someone else must provide proxy consent.
3. Research participants have an unconditional right to withdraw from research at any time.
4. Research projects need to be reviewed by an independent committee and can only begin if the committee approves the project.

The requirement for informed consent entails that the prospective participants must be fully informed about the purpose of the research, what their involvement will entail, and any risks involved as well as information about funding and possible conflicts of interest on the part of the researchers. After having received this information and having had a chance to ask questions and have them answered, the persons must then give their voluntary consent to participation.

The declaration puts the decision-making capacity and responsibility solely with the individual participant, except where the participant is incompetent. It does not allow consent from community leaders or family members on behalf of competent potential participants. In relation to minors, it recognizes the need to consult them about participation, if possible, and to respect any refusal to participate, but it places no weight on the assent of a minor.

Despite the declaration being a World Medical Association document and therefore morally binding on medical doctors only, the principles and rules in the declaration have influenced research ethics in many other fields outside of medicine where researchers recruit human participants for their research. Many countries and institutions have established research ethics committees (this is the term most often used outside of the United States), research ethics boards (the term used in Canada) or institutional review boards (the term used in the United States) for research involving human subjects, and informed consent has become a near universal requirement in relation to human research participation and data collection. Some see this as an unwarranted extension of a biomedical model of research governance to other areas with other research traditions and other standard types of research design.

standard types of research design.

The requirement for prior review of research projects has been seen as controversial and has been argued to be a bureaucratic hurdle. It is argued that fully trained researchers, such as social scientists with PhDs who are guided by their profession's ethics guidelines, will be fully able to design and conduct ethical research and respond to any ethical challenges that may occur during the research and that they do not need a review board to approve their research or provide them with advice. On the other hand, it is argued that researchers are not always free from conflicts of interest related to their desire to conduct research and make progress in their careers and that oversight is therefore necessary.

A particular issue of incompatibility between standard research designs in other fields and the declaration arises when researchers in education, psychology, or the social sciences employ deception in the information given to participants. This most often involves explaining the research procedures in full, but not revealing the research purpose in detail in cases where knowledge of the purpose would bias the participants and thereby invalidate the research. Any deception is incompatible with the declaration's focus on full informed consent, but mild deception of this kind is deemed warranted and unproblematic by many outside of medical research if research participants are properly debriefed after participation.

Another issue of potential incompatibility is created by the implicit assumption inherent in the declaration that a researcher approaches potential research participants one by one to seek their informed consent. This does not fit well with, for instance, observational research designs that study groups, as is often the case in educational and anthropological research. In such circumstances, it can be difficult or impossible to gain informed consent.

Søren Holm

See also Assent; Ethical Issues in Educational Research; Ethical Issues in Evaluation; Human Subjects Protections; Informed Consent; Institutional Review Boards; Nuremberg Code

Further Readings

Goodyear, M., Krleza-Jeric, K., & Lemmens, T. (2007). The Declaration of Helsinki. British Medical Journal, 335, 736.

Høyer, K., Dahlager, L., & Lynöe, N. (2005). Conflicting notions of research ethics: The mutually challenging traditions of social scientists and medical researchers. Social Science … Medicine, 61(8), 1741–1749.

Shuster, E. (1997). Fifty years later: The significance of the Nuremberg Code. New England Journal of Medicine, 337(20), 1436–1440.

World Medical Association. WMA Declaration of Helsinki—Ethical principles for medical research involving human subjects. Retrieved from http://www.wma.net/en/30publications/10policies/b3/

Ana Puig Ana Puig, Ana

Christopher M. Adams Christopher M. Adams Adams, Christopher M.

Delphi Technique

Delphi technique

479

480

Delphi Technique

In the 1950s, Norman Dalkey and Olaf Helmer developed the Delphi technique at the Rand Corporation as a way to forecast technological trends. The technique, named after the ancient oracle of divination, has evolved as a way to generate ideas and facilitate agreement among experts in a particular field through a series of questionnaires or surveys in which they anonymously and iteratively express opinions based on emerging agreement and consensus. Delphi methods can include qualitative (open-ended questions) and quantitative components (Likert-type survey items) and have been used in educational research, business, and health care, and are increasingly being utilized in counseling, psychotherapy, and psychology research. This entry explores the basic principles and procedures of the Delphi technique and examines other applications and limitations.

Basic Principles and Procedures

The basic premise of the Delphi technique is to develop consensus among expert opinions on a particular subject. Although there may be variations in design, the data gathering process in general includes four phases: (1) expert panel members are selected to respond to an open-ended questionnaire, informed by extensive literature review, to gather their opinions about a specific topic or area of focus, (2) the input from each content expert is recorded to grasp group perceptions about the topic, (3) researchers further investigate expert views via a follow-up survey, and (4) researchers review all information after the experts have analyzed preliminary data and provided input. The time span between waves can range from 2 weeks to 1 month, depending on the number of statements provided in the initial review document.

The Delphi technique is most useful to gather opinions of experts who are regionally or geographically apart. The panel can range from a handful to over 100 people, depending on how many people are deemed to have expertise on the subject. Delphi studies usually comprise three to five waves of data collection; however, the number may change based on the subject matter being studied. The data analyses for each wave can vary depending on the researchers' design and aims. Preferred methods of analyses include mean, median, and mode (measures of central tendency); standard deviation; and interquartile range. After the first wave or round, researchers begin to rank the answers provided and develop a list of items that experts review and provide additional input about. Researchers prioritize these items and clearly state their rationale for doing so. A survey is developed that allows experts to continue to provide input, modify their judgment, and indicate importance of each aspect of the topic. Subsequent waves provide strengthening of consensus about the subject matter.

The Delphi technique has been criticized for lacking rigor. Adaptations of the method are common, and its design has varied since the technique's inception. To address this problem, in 2000, Felicity Hasson, Sinead Keeney, and Hugh Patrick McKenna published a set of guidelines for researchers aiming to use this method. They provide a detailed checklist researchers can use to guide their process and contend that the technique can be useful when attempting to gather opinions and develop consensus among a large group of individual experts.

Other Applications

The Delphi technique has been extensively applied in the field of technological forecasting across the globe and as such has an established, strong base in that field. Many other disciplines have adopted the Delphi technique. In 2004, Chitu Okoli and Suzanne D. Pawlowski reported how the method can be used for theory building, citing benefits from generalizability of the theory (based on expert consensus) to strengthening of construct validity. In 2006, Jon Landeta explored the method's current validity in the social sciences due in part to its proliferation. He provides a comprehensive survey and evaluation of published studies from 1995 to 2004 and reports finding several variations in the technique. Namely, it has been used to explore expert opinions about economic and statistical models and for analyses of complex social realities with the aim to develop policy decisions. In the 1980s, the Delphi technique was at its height of popularity; however, in the early 2000s, Landeta reports a resurgence of the method as evidenced, in part, by an increase in application and adaptation by doctoral students.

Limitations

The Delphi technique is not without limitations. One of the most notable problems in using Delphi methods relates to the expert panel itself. Specifically, there is a lack of clear guidelines and agreement in the literature, regarding the size of an expert panel as well as how to select panel members (inclusion criteria). In addition, panel members have the potential for bias, as they may be limited in their perspective on a topic, given their high level of expertise. Another problem related to the panel is attrition. Delphi methods can be time-consuming and may require even more than the suggested minimum of three rounds of data collection. Because of this, panel members may withdraw from a study, resulting in a low response rate across rounds. A final issue related to the panel is the matter of consensus. There is no agreement in the literature on what constitutes a consensus by panel members.

Delphi results can also be affected by how broad or narrow the initial questionnaire items are. Similarly, results may be misleading. For example, measures of central tendency tend to be reported in Delphi studies and could be skewed by factors such as sample size. Finally, the results of Delphi studies cannot be generalized or used to make statements regarding causality, and follow-up studies are often needed to verify the findings.

Ana Puig and Christopher M. Adams

See also Educational Research, History of

Further Readings

Dalkey, N. (1972). The Delphi method: An experimental study of group opinion. In N. Dalkey, D. Rourke, R. Lewis, & D. Synder (Eds.), Studies in the quality of life: Delphi and decision-making (pp. 13–54). Lexington, MA: Lexington Books.

Hasson, F., Keeney, S., & McKenna H. P. (2000). Research guidelines for the Delphi Survey Technique. Journal of Advanced Nursing, 32, 1008–1015.

Hsu, C., & Sandford, B. (2007). The Delphi technique: Making sense of consensus. Practical Assessment, Research … Evaluation, 12, 1–8. Retrieved from http://pareonline.net/getvn.asp?v=12…n=10

Linstone, H. A., & Murray, T. (Eds.). (2002). The Delphi method: Technique and applications. Boston, MA: Addison-Wesley Publishing. (Original work published 1975) Murry, J. W., & Hammons, J. O. (1995). Delphi: A versatile methodology for conducting qualitative research. The Review of Higher Education, 18, 423–436.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design consideration, and applications. Information … Management, 42, 15–29.

Jeffrey N. Rouder Jeffrey N. Rouder Rouder, Jeffrey N.

Delta Plot

Delta plot

480

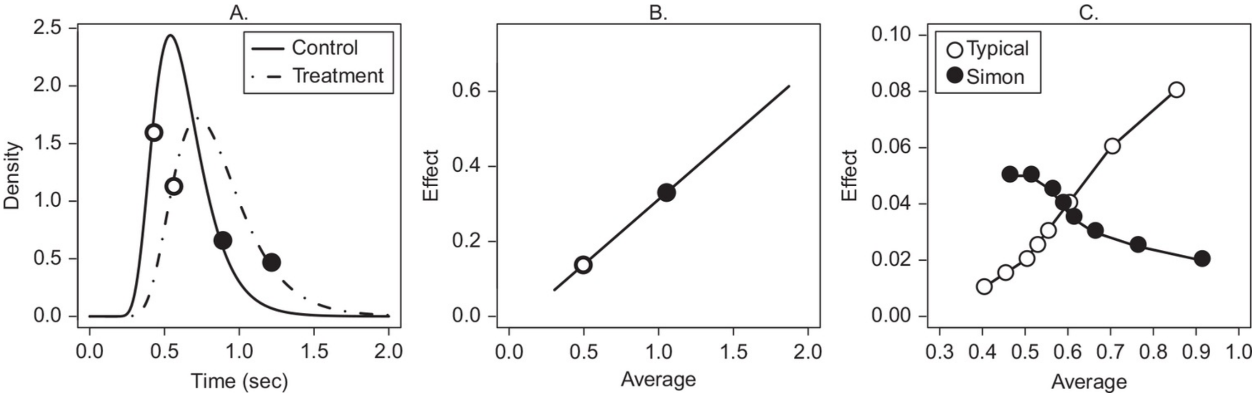
481

Delta Plot

A delta plot is a graphical display for comparing two distributions that is closely related to quantile-quantile (QQ) plots. A common application in education is to inspect whether there is differential item functioning by plotting distributions of item discriminability in one population versus that of another. The setup is one with two populations (or conditions), which are generically called *Population A* and *Population B*. Let *X(p)* and *Y(p)* denote the *p*th percentile for the dependent measure for Populations A and B, respectively; let *d(p)* = *Y(p)* − *X(p)* be the difference, or the effect, at the *p*th percentile; and let *m(p)* be the average effect at the *p*th percentile. The delta plot is a plot of the effect, *d(p),* as a function of the average, *m(p)*.

Figure 1 provides an example for item discriminability parameters. The distributions are in Panel A; the corresponding delta plot is the solid line in Panel B. The open circles show the 10th percentile for each distribution, and the values are respectively .68 and .74 for the control and treatment conditions. The difference at the 10th percentile is .06, and the average is .71, which is the open circle in Panel B. The filled circles show the same for the 90th percentile. The line is the plot for the remaining percentiles, and it has positive slope.

Figure 1 Delta Plots: (A) Two distributions of IRT discriminability parameters that are to be compared. (B) Corresponding delta plot shows that these distributions are not equal. (C) Delta plot patterns are sometimes used to document the time course of processing.



The key question in education is often whether two distributions differ. If they are the same—that is, if items are functioning without differentials across populations—then the delta plot line should be a horizontal line at the value of 0, which in the figure is indicated with the dashed line labeled “equality.” The delta plot line is not on this equality line, and the deviation indicates differential item functioning. Items tend to have a larger standard deviation in Population B than in Population A.

Delta plots are rotated versions of quantile-quantile plots, and the rotation allows the analyst to choose a smaller range for the *y*-axis. With this smaller range, deviations from horizontal lines at 0 are visually emphasized.

In experimental psychology, delta plots are used with response time distributions to interpret the time course of effects. Panel C shows two delta plots that have opposite patterns: The one labeled “typical” shows a rising pattern (i.e., the effect is a slowing of the slower responses) and the other plot, labeled “alternative,” starts high and decreases. The effect is a slowing of the faster responses. The typical plot is indeed typical in many applications; for example, it describes the slowing from cognitive aging. The fastest of the elderly are about as fast as the fastest college-age individuals, but the slowest of the elderly are quite a bit slower than the slowest of the college-age individuals. The alternative pattern occurs for just a handful of phenomena, and it indicates a fast effect that decays quickly in time.

Jeffrey N. Rouder

See also Distributions; Histograms

Further Readings

De Jong, R., Liang, C. C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response concordance. Journal of Experimental Psychology: Human Perception and Performance, 20, 731–750.

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. Attention, Perception … Psychophysics, 72, 2013–2025.

David E. Meens David E. Meens Meens, David E.

Kenneth R. Howe Kenneth R. Howe Howe, Kenneth R.

Democratic Evaluation Democratic evaluation

481

486

Democratic Evaluation

The term *democratic evaluation* (DE) refers to theoretical frameworks that conceptualize the assessment of public programs or initiatives (in education, health care, etc.) in terms of its role in and contribution to democratic politics and culture. Although many, perhaps most, theories of program evaluation include democratic elements (e.g., emphasis on stakeholder participation and evaluator responsiveness, a commitment to the empowerment, and even “liberation” of disadvantaged individuals or communities), DE goes further, explicitly linking the narrower and more immediate goals specified for any particular evaluation to an overarching goal of creating a more just and democratic society. The main task of evaluators working within such a framework involves identifying (or developing) methods of engagement, analysis, and dissemination adequate to this aspiration.

The focus of this entry is the description and appraisal of the two major democratic theories of evaluation: Barry MacDonald’s DE and Ernest House and Kenneth Howe’s deliberative democratic evaluation (DDE). The entry first discusses the conception of evaluation that these theories seek to correct or supplant. Then the basic outlines of DE and DDE are described with an eye toward their respective rationales and implications. Finally, the entry examines the reception of democratic theories within the field, including key criticisms, and concludes with a brief appraisal of their impact and ongoing relevance.

Technocratic–Managerial Evaluation

Both MacDonald and House and Howe developed their theories in response to what can be termed the *technocratic–managerial* theory of program evaluation, which has shaped the field from its earliest days. This conception is technocratic in that it seeks to position evaluation as a value-neutral and apolitical activity, formally eschewing judgments about program values and goals. It is managerial in that it aligns itself with the interests and perspectives of program managers, fiscal patrons, and other parties with a vested interest.

As House argues in an influential history of professional evaluation, traditional institutions have declined in importance in modern capitalist societies. This is not to say that such institutions are necessarily less important in the lives of particular individuals and communities; rather, they have ceased to provide the generally accepted justification for social practice. This presents a challenge for modern states whose governments must appear responsive to the demands of diverse constituencies whose goals and interests often conflict. A “solution” that emerged from the mid-1960s on, most notably in the United States and the United Kingdom, seeks to ground social action in appeals to the authority of “reason” (i.e., scientific rationality) as applied to social programs.

Prior to 1965, formal program evaluation had been a marginal activity. With the passage of the Elementary and Secondary Education Act, however, it became a federal mandate. This expanded role for evaluation occurred at a time of social upheaval, during which marginalized groups and their allies vigorously pressed claims for the redress of long-standing injustice, and many members of historically privileged groups mobilized to preserve the status quo. Absent a shared cultural basis for determining political priorities, policy makers promoted evaluation as a value-neutral method for determining the merit of contentious social programs, many of which involved significant public expenditures and controversial expansions of government bureaucracy.

The notion of scientific reason underlying the technocratic–managerial conception depends upon a sharp distinction between judgments of fact and judgments of value. This is apparent in the work of Donald Campbell, a towering figure in the development of evaluation theory. Campbell accepted the axiom, inherited from the logical positivists, that value claims are epistemologically different than claims about facts. The latter are subject to rational determination, the former are not. On this view, evaluation of whether a particular program has worked, or worked better than others, involves only judgments of fact. Judgments of value, however—concerning, for example, the justice of goals set by managers, funders, or policy makers—cannot be determined rationally and so necessarily lie beyond the scientific evaluator’s determined rationally and so necessarily lie beyond the scientific evaluator's purview.

In keeping with this basically positivist orientation, the prevalence of the technocratic–managerial conception was reflected in the privileging quantitative/statistical research methods. Through the application of such methods, it was thought, evaluation could identify causal mechanisms that would enable effective technocratic control of social phenomena. Such aspirations align closely with the interests and perspectives of program managers, and it is unsurprising, therefore, that in this early phase there was little emphasis on stakeholder participation, nor much attention to the diversity of legitimate (and competing) aims present within the context of social programs.

The Political Turn: MacDonald’s *DE*

The view of program evaluation as value neutral and apolitical almost immediately gave rise to critiques within the emerging community of evaluation scholars. By the mid-1970s, a transatlantic group of the “next generation” evaluators, including Robert Stake, Michael Scriven, and Lee Cronbach, to name a few, challenged the general character of evaluation studies, rejecting the premise that legitimate evaluative findings must be generalizable in order to be valid. This more skeptical attitude toward generalization turned evaluators’ attention toward the specificity and complexity of a given program context, and authorized a methodological shift toward the use of case studies and qualitative rather than quantitative methods.

MacDonald was a key participant in these developments. The author went further than many of his contemporaries in arguing that evaluators inevitably engage, wittingly or not, in adjudicating political conflict. This is because evaluation inevitably confronts the distribution and exercise of power, and evaluators must make decisions that amount to inveighing on behalf of some constituency or another. In light of this principle, MacDonald contrasted three types of evaluation, each defined in terms of constituents whose interests it privileges.

The first, *bureaucratic*, provides unconditional service to government agencies and seeks to maximize efficiency as a consultant to management. The second, *autocratic*, is a modification of the first in which the evaluator remains independent of the government agencies and maintains a degree of professional autonomy. This autonomy as an “outside expert,” however, remains in the service of government agencies and program managers. Professional autonomy serves to enable evaluators to more effectively legitimize existing policy directions. The third type, *DE*, serves not only government agencies but also the broader public and its interests.

In MacDonald’s view, a democratic commitment to the public good includes but also constrains evaluators’ bureaucratic/autocratic responsibilities. It foregrounds the issue of who determines the focus and scope of evaluation research, who participates in the process, and who “owns” evaluative findings. In modern democratic societies, decision making about social policy often involves the public as a whole. Therefore, the public’s right to know requires dissemination of evaluative knowledge beyond program managers and official decision makers. MacDonald recommended that such considerations be formally addressed in evaluation contracts agreed upon by all major stakeholders.

As MacDonald and his followers conducted DE in the United Kingdom and beyond, the approach was soon subject to critique. Robin McTaggart argued that MacDonald’s democratic approach in one case actually served to advance the interests of already powerful stakeholders at the expense of others. In the Australian educational program she studied, teachers, administrators, and program evaluators agreed to a set of principles and procedures at the start of the study. One key informant withdrew her consent, however, when it became clear to her that having her criticisms of the program published as part of the findings could have negative consequences, perhaps even costing her job. The evaluators thus faced a dilemma between their valuing of the “public’s right to know” and the individual’s right to “own facts about their own lives.”

Although DE represented a major democratic advance for evaluation theory, MacDonald leaves relatively open the question of what, if any, particular values should be advanced by evaluators within a given situation or program. A stated commitment to democratic values of transparency and publicity provides little guidance for how conflict over substantive values is to be resolved or how power differentials among participants are to be mitigated. Most problematic in this case, as McTaggart argues, was the trappings of democratic processes and language served to mask rather than correct the power disparities at play. MacDonald’s focus on processes that “give voice” to diverse interests provides guidance on how to identify, much less correct, power imbalances in the process.

House and Howe’s DDE

Parallel to MacDonald’s work in England and around the same time, House argued against the technocratic conception in the U.S. context. Like MacDonald, House argued convincingly against pretensions to value neutrality on the part of evaluators. House went further than recognizing values pluralism, suggesting that evaluators have a special responsibility to advance a particular set of values, namely, those associated with “social justice.” Too often, according to House, evaluation has not paid meaningful attention to the interests of the least advantaged, those whose needs on social programs are nominally designed to meet. A more ethical and democratic conception of evaluation, on this view, is centrally concerned with addressing power imbalances and redressing issues of inequality.

Initially, House drew upon the work of political philosopher John Rawls in conceptualizing the requirements of justice within the context of evaluation. Rawls’ seminal *Theory of Justice*, first published in 1971, put forward an egalitarian conception of justice focused especially on the consequences of institutional arrangements for the “least advantaged.” In the Rawlsian perspective, justice is essentially concerned with the fair distribution of social benefits and burdens, within a context defined by equal liberties for all persons and “fair equality of opportunity.” Defining evaluation as an institution that ought to be committed to justice, House provided a theoretical basis for including substantive values in evaluation—in particular, advocacy for the interests of the poorest and most vulnerable. A Rawlsian focus on distributive social justice does not, however, address all the practical and theoretical issues discussed vis-à-vis evaluation as practiced in democratic societies—in particular, the question of whether (and to what extent and how) the “least advantaged” should have a say in defining their own interests and needs. Like MacDonald, House’s earlier work suggests a representative rather than participatory conception of democracy.

In the late 1990s and early 2000s, House shifted toward a more participatory conception and, in collaboration with philosopher of education Howe, developed a distinct theory—DDE. DDE combines a procedural concern of giving voice to the values and interests of diverse stakeholders through democratic procedures and process (*à la* MacDonald) with House’s long-standing advocacy for evaluation as an institution that advances an egalitarian conception of justice.

MacDonald challenged bureaucratic/autocratic/technocratic evaluation on

MacDonald challenged bureaucratic/autocratic/technocratic evaluation on political and ethical grounds. House and Howe add to this a critique of the epistemological basis of the technocratic–managerial conception, which they identify as the "received view" of values in evaluation. Drawing upon philosophical epistemology and philosophy of science, they argue that a hard-and-fast fact/value distinction is untenable. They provide numerous examples of statements in which facts and values intertwine and "shade into one another."

If statements of fact and statements of value are not necessarily distinct kinds, then two related pillars of the received view collapse. The first is what House and Howe term the "radical undecidability thesis," which stipulates that values are not amenable to rational determination—in Campbell's phrase, values are "chosen" but not "justified." The second is an emotive/preferential conception of democracy, in which stakeholders put forward value statements that must be accepted at face value, as there is no rational basis for critique. Politics, on this view, becomes a competition over values that are not subject to reasoned critique and so cannot be rationally adjudicated.

DDE rests upon the premise that values are, like facts, subject to reasoned critique and stand in need of justification based upon evidence and argumentation. The question becomes what sorts of evaluative procedures can provide a reliable warrant for value claims. Drawing upon the deliberative conception of democracy—influential in political theory in the late 20th century—House and Howe position DDE as a "mid-range" theory that 1) addresses the question of how judgments of value within program contexts can best be justified and 2) connects the practice of evaluation to its broader political and institutional contexts in ways that not only reproduce but improve upon the status quo. To the question of how value claims are best justified in evaluative contexts, House and Howe propose three related procedural requirements: inclusion, dialogue, and deliberation.

Inclusion refers to the involvement of diverse (relevant) stakeholders in the evaluation process. House and Howe distinguish between two types of inclusion, formal and substantive. Formal inclusion is "thin," in that stakeholder representatives may be present but still lack the opportunities or resources necessary to influence the process. Substantive inclusion, by contrast, means that all participants are enabled to contribute on equal terms.

Dialogue requires that once included, participants have the opportunity to represent their own interests in conversation with others. This conversation can be elucidating, in which case the goal is to generate understanding of stakeholder views in their own terms or it can be critical, which requires that views not only be understood but also are thoughtfully questioned. It is through critical dialogue that stakeholders come to a more thorough understanding of their interests and possibly modify this understanding in light of the interests expressed by others.

Deliberation refers to the purposeful discussion about how to resolve the value conflicts that emerge in dialogue. In contrast to the emotive conception of political discourse, democratic deliberation is a cognitive process, grounded in reasoning, consideration of evidence, and principles of valid argument. House and Howe assert that substantive inclusion shades into critical dialogue, which in turn shades into deliberation. Taken together, these principles provide a regulative ideal for justifying judgments of value as essential evaluative findings.

DDE conceives program evaluation as a fundamentally participatory process of collective inquiry. This does not mean that the role of the evaluator is diminished. On the contrary, effective democratic evaluators provide skillful guidance on how to make reasoned judgments about values. Gathering quality information to be fed into deliberations and skills and knowledge related to the design and facilitation of meaningful dialogue are crucial factors that increase the likelihood of DDE’s success in any given case. If anything, DDE seems to require more, not less, of evaluators in their professional role than does the received view.

Reception and Prospects

Over the years, DD and DDE have been received sympathetically by theorists from a variety of traditions (e.g., Helen Simons, Jennifer Greene, and Cheryl MacNeil). These models have not, however, had the more thoroughgoing transformative effect on the field that DDE, at least, suggests is desirable. We conclude this entry by briefly summarizing some critical responses offered within the academic evaluation community and with an appraisal of continuing influence and prospects of DE.

To reiterate, the basic criticism leveled at MacDonald’s DE is that it fails to deal adequately with power imbalances inherent in institutional contexts and provides little guidance for crucial decisions about which values and voices ought to be included. DDE explicitly addressed these issues and in turn gave rise to distinctive criticisms in its own right. One of these asserts that DDE's embrace of a set of substantive democratic values (inclusion, dialogue, and deliberation) goes too far.

A response of this type is offered by Robert Stake. Renowned for his role in the turn to the case study method and contributions to responsiveness in evaluation, Stake rejects the overarching value commitment of DDE to advancing democracy in general. He faults House and Howe for advocacy (albeit a "literary" one), even "zealous rallying" on behalf of a particular and debatable conception of democracy. In order to maintain the faith of clients and the public in the profession, Stake suggests evaluators restrict their concerns to the immediate goals defined by the program and the particular interests of stakeholders in a given context. This argument is, to an extent, reminiscent of the original impetus to the expansion of formal program evaluation in the United States during the 1960s. In a context of values pluralism, on this view, the perception of evaluation as a narrowly technical, value-neutral activity is essential if it is to help legitimize potentially controversial social programs.

A modest commitment to the promotion of democracy is all that Stake is willing to countenance. DDE, in his view, goes much too far. Others have argued, on the contrary, that DDE does not go far enough. Stafford Hood offers criticism along this line, pointing out that the experience of democracy in the U.S. context (and undoubtedly in others) has been extremely varied and that substantive inclusion is more easily said than done. The continued salience of race provides a conspicuous example of a pervasive social phenomenon that undermines inclusion, dialogue, and deliberation in subtle and overt ways. DDE's focus on methodological requirements will not, on its own, neutralize such factors, absent additional remedies such as greater representation of historically disadvantaged racial groups within the program evaluation community.

There is reason to doubt the long-term viability of the technocratic, value-free, or value-minimalist conception of evaluation. A half-century after the emergence of program evaluation as a profession, faith in "scientific reason" and technical expertise as a neutral arbiter of social conflict has waned significantly. When pretensions to value neutrality are generally regarded skeptically, yet another type of justification for social practice is necessary. The ongoing relevance of DDE lies in the fact that it provides one defensible model for developing such justification, without appeal to discredited epistemic and political assumptions that have dominated the field of program evaluation during its first 5 decades.

Recognition that substantive inclusion, critical dialogue, and deliberation are difficult to achieve in practice does not, in the view of proponents, discredit DDE so much as set a course for its future development.

David E. Meens and Kenneth R. Howe

See also Social Justice; Stakeholders

Further Readings

Arens, S. A. (2000). Review of values in evaluation and social research, by E. R. House and K. R. Howe. Evaluation and Program Planning, 23(3), 331–333.

House, E. R. (1993). Professional evaluation: Social impact and political consequences. Thousand Oaks, CA: Sage.

Howe, K. R., & Ashcraft, C. (2005). Deliberative democratic evaluation: Successes and limitations of an evaluation of school choice. Teachers College Record, 107(10).

Kushner, S., & Adelman, C. (2009). Program evaluation: A democratic practice. In J. L. Green, G. Camilli, & P. Elmore (Eds.), Handbook of complementary methods in education research (pp. 711–726). New York, NY: Routledge.

MacDonald, B. (1976). Evaluation and the control of education. In D. Tawney (Ed.), Curriculum evaluation today: Trends and implications (pp. 125–136). London, UK: Macmillan Education.

MacDonald, B. (1978). Evaluation and democracy. Edmonton, Canada: Public Address at the University of Alberta Faculty of Education.

McTaggart, R. (1991). When democratic evaluation doesn’t seem democratic. Evaluation Practice, 12(1), 9–21.

Ryan, K. E., & DeStefano, L. (Eds.). (2000). Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation: New Directions for Evaluation, No. 85 (pp. 3–12). San Francisco, CA: Jossey-Bass.

Gregory J. Marchant Gregory J. Marchant Marchant, Gregory J.

Demographics

Demographics

486

488

Demographics

Demographic variables are characteristics inherent in individuals and groups. Common demographic variables include age, grade level, race, and sex. Indicators of socioeconomic status (SES) are common demographics considered in educational research. Income, occupation, education level, and possessions of the subjects or the subjects' parents/family are common SES factors. The factors related to SES may be considered on their own or combined into an index or score. Other common demographic variables include marital status, family structure, religion, and political affiliation. This entry explores the use of demographic variables in educational research and policy.

Demographics in Research

Demographics cannot be manipulated or randomly assigned; therefore, research efforts using these variables are considered "quasi-experimental" designs. The lack of control over these variables makes it impossible for research on them to be considered "experimental." Establishing causation can be difficult in these studies, and eliminating alternative hypotheses may be problematic. There are many factors known to be related to demographic variables. Consider the following example: Coming from a single-parent family is associated with lower achievement because these families usually have lower incomes, more stress, or less parent interaction. All of these factors are related to lower achievement, but which ones are involved?

Research questions often focus on the demographic variables, such as, do boys differ from girls in reading or is SES related to empathy? However, because demographics tend to be so strongly related to achievement, they are often considered in studies of other factors. For example, if a study were to be conducted on an innovative reading program, the researchers would want to consider sex because they know boys perform differently than girls. If achievement in charter schools were being compared to public schools, all the demographic variables related to achievement should be considered among the students of the schools before making value judgments regarding the schools' relative achievement.

Therefore, before the research question is addressed, demographics are crucial to any study. They define the population and therefore the sample. They impact validity of a study and limit the generalizability of the findings. Selection bias is one of the greatest threats to validity in educational research, and demographics are often at the heart of the problem. The nature of students when comparing teachers, grade levels, schools, districts, and states matters in both simple and complex ways. The interaction of demographic factors with each other and various outcome variables provide serious challenges that cannot be ignored.

Random assignment of demographic factors is impossible, and random selection of subjects in educational research is difficult. Often the sample in educational research is a convenience sample of available students, teachers, classrooms, or schools. If groups differ based on demographics, anything related to those demographics is likely to differ. This means the differences or relationships the researcher is looking for related to schools, teachers, or outcomes may simply be the demographics. To compensate for the lack of randomization, researchers sometimes match groups based on demographics or statistically control for the group demographics (e.g., covariants).

Demographics in Policy

In the United States, demographic categories came to the forefront in policy when the No Child Left Behind legislation required that schools meet an overall achievement criterion not only for their students but also for demographic subgroups. Each group based on race, poverty, special education, and English-language proficiency was required to meet the criterion. If one group did not pass, the school failed to pass. This disaggregation—breaking a large group into smaller subgroups—may have been well intended, so the subgroups would not be ignored, but it failed to account for interactions. In other words, some students were members of more than one subgroup, making comparisons unfair.

For example, there are achievement gaps for students who are Black and students in poverty. If one school has poor Black students, but another school has poor students who are not Black, or Black students who are not poor, the school with the poor Black students is going to be depressed in both categories (making it more difficult to pass either category).

One of the easiest ways to improve achievement for a school is to manipulate the demographics of those enrolled. If minorities, poor students, and special needs students are not recruited to a private or charter school, the school may exhibit higher achievement. If students who traditionally score lower on achievement tests are expelled or suspended during testing, a school may demonstrate artificial achievement gains. Demographics also play a role in teacher evaluations. Teachers having students from backgrounds known to demonstrate lower achievement are at a disadvantage. Demographic differences are related not only to achievement but also to changes in achievement. Even efforts to overcome this problem are flawed. Statistical efforts to look at growth or determine the “value added” by the teacher are still influenced by the nature of the students. This results in teachers being evaluated as successful one year but failing the next as their students change.

Whether the demographic factor is a major variable in the study or a characteristic of the sample, the nature of the participants is a necessary consideration. There is no perfect approach to dealing with the quasi-experimental nature of demographic variables. Due to interactions of these variables with each other and outcome variables, matching and statistical controls help but do not overcome the complex issues posed by demographics. Researchers and evaluators ignore demographic factors at their own peril.

Gregory J. Marchant

See also No Child Left Behind Act; Quasi-Experimental Designs; Random Assignment

Further Readings

Frey, W. H. (2014). Diversity explosion: How new racial demographics are remaking America. Washington, DC: Brookings Institution Press.

Kena, G., Hussar, W., McFarland, J., de Brey, C., Musu-Gillette, L., Wang, X.,

& Dunlop Velez, E. (2016). Population characteristics in conditions of education. National Center for Educational Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016144

Paulson, S. E., & Marchant, G. J. (2009). Background variables, levels of aggregation, and standardized test scores. Educational Policy Analysis Archives, 17(22). Retrieved from http://epaa.asu.edu/ojs/article/view/389/512

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

U.S. Census Bureau. (2014). Census explorer. Retrieved from https://www.census.gov/censusexplorer/censusexplorer.html

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

Descriptive Statistics Descriptive statistics

Descriptive Statistics

Statistical approaches are subdivided into two major divisions, descriptive and inferential statistics. As the name implies, descriptive statistics entail describing, organizing, and summarizing information. Data are described by graphical methods, numerical indices, and tables. Descriptive statistics also often include a commentary discussing the data structure and any emergent patterns. In contrast, inferential statistics make inferences or estimations about a population from a sample through hypothesis testing and confidence intervals. Inferential statistics are associated with probability theory in order to reach conclusions about a variable beyond the data collected and determine the relative certainty of those conclusions. Statistical methodology, therefore, encompasses descriptive statistics that summarize data and inferential statistics that generalize data from a small group to a larger group. This entry focuses on descriptive statistics, revealing their primary goal, describing the three main types of descriptive statistics—measures of central tendency, measures of dispersion, and measures of distribution shape—and reviewing how graphics can illustrate these statistics.

The primary goal of descriptive statistics is to maximize information and communication effectiveness while minimizing the loss of information. Through a few quantitative values and/or graphical summaries, descriptive statistics reduce large data sets into a simpler, more manageable form. The challenge then is to determine which statistics best summarize the major characteristics of the data set, yet avoid misleading results.

Selecting the proper descriptive statistic is largely dependent on the data characteristics and underlying research goals. The data measurement level (nominal, ordinal, interval, or ratio) determines the types of mathematical operations possible. Calculation methods are also altered depending on whether or not the data are grouped (weighted) or ungrouped (unweighted). In spatial or not the data are grouped (weighted) or ungrouped (unweighted). In spatial (geographic) data sets, the statistics employed and their interpretation are dependent on the study area boundaries, spatial resolution, and aggregation level (e.g., county, state). Regardless of the data properties, descriptive statistics cover three main types: (1) measures of central tendency, (2) measures of dispersion, or (3) measures of distribution shape.

Measures of central tendency indicate the middle or typical data value. The three most common measures of central tendency are the mean, median, and mode. The mean or average is the summation of all the values divided by the number of observations; hence, the mean can be strongly influenced by isolated values that are exceptionally large or small and are known as outliers. In contrast, the median is based on the middle position within a set of ranked values where the same number of data points lie above and below the middle value. The mode identifies the most frequent observation in a set of ungrouped data and is most appropriate for data sets with multiple tied observations. A less common measure of central tendency is the midrange, the average of the maximum and minimum values. The “best” measure of central tendency depends on the characteristics of the data distribution (e.g., relative symmetry, outliers) and inferential statistics requirements.

Measures of dispersion focus on the spread or variability in the data. The simplest measure is the range (not to be confused with midrange) or the difference between the maximum and minimum values. If outliers are present, then the range highlights only the data extremes, and other methods that showcase the amount of clustering or spread are needed. Quantiles divide the observations into equal amounts or percentages, usually in quartiles (quarters), quintiles (5th), or deciles (10th); thus, specific intervals within the distribution are examined. The most common dispersion measure is standard deviation, which integrates the least squares property of the mean (the difference between the data value and the mean) and accounts for variations in sample size. In general, relatively larger or smaller standard deviations indicate larger or smaller variability in the data set. The square of the standard deviation or variance is a frequent component in many inferential statistics applications; however, variance is not usually reported alone because the values can be extremely large and more difficult to interpret.

In addition to communicating the middle and spread of the data, descriptive statistics can express the data distribution shape. Measures describing the frequency distribution shape are often in reference to the normal distribution, an idealized bell-shaped (symmetric) curve from which the mean (and median and mode) is at the single center peak. Skewness describes the symmetry of the distribution by measuring whether the extent data are evenly or unevenly spread on either side of the mean. More positive or negative skewness indicates the distribution has some values much greater or lower than the mean, whereas a skewness value near 0 denotes the distribution is symmetric about the mean. Kurtosis describes the shape of the distribution peak. Negative kurtosis values refer to a relatively flat peak or platykurtic distribution, whereas positive kurtosis values describe a comparatively sharp peak or leptokurtic distribution. Peaks that are neither flat nor pointed are called mesokurtic and have kurtosis values around 0.

Descriptive statistics not only include quantitative indices for central tendency, variability, and shape but also graphical visualization of these indices. For instance, pie charts quickly highlight the highest and lowest frequency responses as a percentage or proportion of the whole. Dot plots, stem and leaf plots, and bar charts are useful in showing the distribution shape and how well the data conform to a normal distribution, an important data assumption in many statistical tests. Histograms, a type of bar graph, show the relative frequency of observations for grouped data. Boxplots are useful for emphasizing data variability using the data range and quartiles, whereby the box is formed from the boundaries of the first (25%) and third (75%) quartiles; lines (called whiskers) are then extended from the box to the maximum and minimum values. Scatterplots showcase the relationship between two quantitative variables, with dots in an *X*-*Y* grid representing how much and in what direction one variable influences another variable. Encompassing both quantitative measures and graphical displays, descriptive statistics summarize data into a practical format and serve as the basis for most quantitative analysis.

Jill S. M. Coleman

See also Inferential Statistics; Interquartile Range; Kurtosis; Levels of Measurement; Measures of Central Tendency; Measures of Variability; Quartile; Skewness; Standard Deviation; Variance

Further Readings

Coolidge, F. L. (2012). Statistics: A gentle introduction. Thousand Oaks, CA: Sage.

McGrew, J. C., Jr., Lembo, A. J., Jr., & Monroe, C. B. (2014). An introduction to statistical problem solving in geography. Long Grove, IL: Waveland Press.

Tokunaga, H. T. (2015). Fundamental statistics for the social and behavioral sciences. Thousand Oaks, CA: Sage.

Kent J. Crippen Kent J. Crippen Crippen, Kent J.

Julie C. Brown Julie C. Brown Brown, Julie C.

Design-Based Research

Design-based research

489

493

Design-Based Research

Design-based research (DBR) is a form of inquiry characterized by iterative cycles of development, testing, and refinement of an intervention that is developed in collaboration with stakeholders and then deployed and evaluated in the rich, real-world contexts. DBR is simultaneously committed to providing theoretical contributions and practical solutions to educational problems. In education, DBR has been used to study curriculum, instructional strategies, professional development, and technology-enhanced learning environments. Ann L. Brown and Allan Collins first introduced the idea of DBR in 1992 in response to the critique that laboratory studies lacked ecological validity or the ability to approximate real classroom situations. Although DBR has been appropriated in numerous ways and has evolved over time, there are several core features that define the approach, namely that it is interventionist, theory driven, context-specific, collaborative and contains a dual, concomitant focus on local impact and theory generation. This entry begins with an explanation of the common processes, then provides an illustration of the practice, and concludes with a discussion of the critiques of DBR.

The Common Processes of DBR

DBR begins with the exploration, analysis, and subsequent identification of a practical problem that is to be addressed by a designed intervention (i.e., intervention or design intervention). Thus, the problem is defined and subsequently addressed within the context of its occurrence. A discrepancy between the intended and actualized state of an educational system defines that problem. For example, consider the situation in which an implemented policy is to promote equity among students in mathematics, yet a curriculum has been adopted and is being used that prioritizes the interests and ways of knowing for one population over another. Such a situation is a prime example of a problem for DBR. Through the processes of exploration and analysis, the sociocultural context of the problem is detailed; the needs of people who are potentially impacted by the problem are assessed; and relevant, published, and authoritative reference material related to the issue and context is explored. The outcome of this process is a clearer understanding of the dimensions of the problem and their likely causes. This level of understanding is essential for designing a responsive intervention.

The problem is defined as an emergent phenomenon, emanating from the sociocultural context from which it is situated. The sociocultural context includes the people, ideas, tools, information, language, history, stories, and documents of a community in a certain place and time. To gain a better understanding of the problem, design researchers often collaborate with multiple stakeholders. Collaboration can take on many forms from more traditional roles for the researchers and participants to an intensive collaboration in which stakeholders are intimately involved in all aspects of the research and development process. The level of collaboration depends upon contextual demands, available resources, and ultimate aims. However, the main argument for use of collaboration in DBR is to ensure that the design intervention is constructed as a plausible solution to a legitimate problem identified by the stakeholders and supported through the published literature. Understanding and addressing the situated nature of the problem generates the ecological validity that is attributed to DBR. With its focus on empirical and theoretical grounding, the phase of problem definition is viewed as the first descriptive and theory-generating process of DBR.

Once identified, the dimensions of the problem and their likely causes are addressed through an equally detailed and complementary solution that is enacted as an intervention. Construction of an intervention begins by posing high-level conjectures about how learning is posited to happen in the context. Conjectures are also constructed that articulate the specific features of the intervention that are predicted to produce desired outcomes, including key mediating processes that facilitate these outcomes. Collectively, these predictive statements articulate the logic and theories (i.e., conceptual or design framework) that guide the study of the intervention and commonly take the form framework) that guide the study of the intervention and commonly take the form of a logic model or conjecture map. Logic models specify the relationships among the inputs (investments of the research team), activities (processes used by the project), outputs (creations of the project), outcomes (short-and medium-term results), and impacts (long-term results or change produced by the project). Unlike hypotheses, conjectures are intended to be flexible in response to findings from ongoing analyses. Through conjectures, design researchers identify specific features to test while increasing the argumentative grammar associated with this form of research.

After logic models are constructed and conjectures are clearly stated, the intervention is constructed with the collaboration of partners who have relevant forms of expertise (e.g., practitioners, learning scientists, educational researchers, and policy makers). This construction involves a balance of creative and analytical processes. The detailed construction of a design intervention is both a descriptive theory-generating step and a prescriptive, solution-generalization step in the DBR process. The theory-generating capacity during construction is based upon the comprehensive and grounded nature of the intervention, which involves the overlap and amalgamation of multiple theoretical frameworks with data and first-person perspectives from the context as well as the creativity of the design team for addressing the complexity of the problem. Because the intervention represents a best-case attempt to address the complexity of the problem, the details of the design serve as a prescription for how the same intervention could be used in a different but similar context. This is one way that design research aims to produce usable knowledge.

Once actualized, the design intervention is enacted and its capacity for addressing the problem is assessed. During enactment, data are collected and analyzed in an ongoing fashion to understand how, for whom, and under which conditions the intervention achieves success. Often, a combination of qualitative and quantitative methods is used to inform researchers about participants' interactions with(in) the intervention and how outcomes are facilitated. Results that stem from the ongoing analysis are used as building blocks for theory generation to test the conjectures and logic that were used to support the design and to refine the intervention in preparation for a new cycle of enactment. In this way, DBR is distinct from more traditional intervention research because it is characterized by these iterative cycles of intervention development, testing, and refinement. As an intervention matures through the iterative cycles of DBR, the research methods typically progress from an initial emphasis on a qualitative, single-case exploratory approach to that of a quantitative, evaluative, and even experimental approach. Early cycles of DBR focus on usability of the intervention and the correlational relationships among variables. These cycles build to studies of efficacy and effectiveness, which include causal relationships and estimates of effect on outcomes as well as evidence of impact.

DBR projects commonly result in two types of theoretical contributions—local theories and design frameworks. Local theories, also referred to as domain-specific theories, describe how learning occurs in specific settings. An example is describing how high school physics teachers learn to reason scientifically about core thermodynamics principles when engaged in a professional development program. Depending upon the research design, local theories can also offer evidence-based explanations as to why learning occurs in that context. Through local theories, design researchers develop theoretical understanding about learning *in situ*. Design frameworks, sometimes referred to as design principles, detail the characteristics that are required of the features of an intervention and the conditions under which they must exist, in order to affect the desired outcomes. To continue the example of a physics teacher learning, a design framework might state that, if high school physics teachers are to learn to reason scientifically about thermodynamics, they must have opportunities to engage in argumentation around related science concepts while in professional learning communities. This form of theoretical contribution is more prescriptive than descriptive in that it denotes the specific intervention features and conditions needed to achieve desired outcomes in a particular context. By advancing theoretical understanding about how to best facilitate learning in context, DBR aims to produce usable knowledge that can be flexibly adapted to new settings.

Seeking Balancing: The Practice of DBR

DBR is based on a system's perspective to problem solving, which defines an environment as a network of interconnected and interdependent individual elements. Thus, the practice of DBR implies a continuous balancing act, one that occurs along multiple dimensions. To be successful, research teams are tasked with negotiating an acceptable point of compromise along such dimensions between two seemingly opposed elements. Achieving the necessary collaboration for successful DBR is a good example of one dimension of the balancing act. Collaboration lies at the heart of DBR and implies the need for an intimate working relationship and melding of priorities among a diverse team of people. Research teams include students, consumers, practicing professionals, and policy makers, as well as researchers who individually may have diverse forms of expertise. All research team members have their own perspectives, agendas, needs, and methods for addressing issues. However, successful DBR involves the continuous negotiation of these perspectives and their individual needs in order to leverage the team's capacity for addressing a complex problem in a sustainable way while also affording the opportunity to develop theoretical understanding of the problem itself. The concept of balance in DBR can also be illustrated through the dual focus on theory generation and solution of a practical problem, the balance of creative and analytical processes, as well as the affordances and constraints of working in the rich context of a practical, pragmatic problem.

Balancing the dual focus on theory generation while addressing a practical educational problem—two distinct and opposing themes at the heart of DBR—is a difficult task. An overly prescriptive project focus, one that overemphasizes a solution to the problem, diminishes the capacity for building understanding for how the intervention works, for whom, and under which conditions. For example, to design and create a new educational program without including the capacity to evaluate the logic in light of the existing theories that ground its design is not consistent with DBR. DBR is equally concerned with describing and relating the mediating processes that lead to the outcomes and producing evidence-based claims to this effect. On the contrary, an overly descriptive project diminishes the effectiveness of the intervention for producing the intended outcomes and impact for the participants, limiting the generalizability of the solution beyond the context being addressed. To design and construct an intervention without the accountability for actually addressing the problem is a form of academic exercise, not DBR. DBR equally values the effort to improve the lives of people through its application. Strategies for maintaining this balance include a diverse design team, one consisting of an equal mix of researchers and practitioners, as well as regularly scheduled discussions of how a balance of perspectives is being achieved.

DBR also requires a mixture of creative and analytical processes. An overemphasis of one process over the other limits the potential of the intervention for addressing the problem or, alternatively, the capacity of the inquiry for explaining how the intervention functions. This balance of processes is inexplicably linked and pursuant to the balance between a practical solution and theory generation. Both processes can be synergistically fed by the collaboration of the design team, again drawing from their diverse backgrounds collaboration of the design team, again drawing from their diverse backgrounds and expertise.

The practical context in which DBR is practiced includes many elements that have the potential to either help (afford) or hinder (constrain) the efforts of a design team. These elements could include policies (e.g., scheduling, adopted materials, and protocols), cultural norms (e.g., ways of doing, ways of interacting, and perspectives), qualifications of individuals (e.g., administrator, teacher, and academic credentials), and types of physical resources (e.g., textbooks, computers, and laboratory equipment). The helping or hindering aspect is based upon the nature of each element (what it is), how it is used or implemented, and its interdependencies with other elements. Many elements are immutable and highly interdependent. The design team must decide which elements will have to be used, which need to be used (including those that must be added or eliminated), and, ultimately, how each is used. The team must consider the helping or hindering aspect of each element and its interdependent contribution to the whole system in order to tip the balance in favor of helping, thus improving the chances of ultimately addressing the problem.

Critiques of DBR

Despite its advantages, specific concerns have been raised about DBR. For instance, the contextually rich nature of DBR projects and the embedded role of design researchers possess have been criticized for offering limited generalizability of findings and replicability of a deeply contextualized intervention. Additionally, the dual aims of theory generation and intervention design and testing have occasionally been regarded as improbable, given that each can be considered a major undertaking. Finally, the iterative nature of DBR has been viewed as a weakness because of uncertainty around when intervention development and related research projects are complete. In response to these concerns, design researchers have raised several counterpoints to demonstrate how they establish methodological rigor and attend to the credibility of findings. Although not an exhaustive list, some of these approaches include collecting and analyzing a variety of data; triangulating multiple data sources to exhibit trustworthiness, reliability, and adequate external validity; collaborating in teams of experts to elicit multiple perspectives; working in the context for prolonged periods of time; making the research and design processes transparent and richly detailed; including clear conjectures about salient intervention design features and their theoretical underpinnings; and arguing for flexibly adaptive theories that can be applied to design as a desired aim as opposed to generalizability.

Kent J. Crippen and Julie C. Brown

See also Applied Research; Ecological Validity; Evaluation Versus Research; Formative Evaluation; Improvement Science Research; Responsive Evaluation

Further Readings

Anderson, T., & Shattuck, J. A. (2012). Design-based research: A decade of progress in education research? Educational Researcher, 41(1), 16–25.

Bang, M., Medin, D., Washinawatok, K., & Chapman, S. (2010). Innovations in culturally-based science education through partnerships and community. In M. Khine & I. Saleh (Eds.), New science of learning: Cognition, computers and collaboration in education (pp. 569–592). New York, NY: Springer.

Barab, S. (2014). Design-based research: A methodological toolkit for engineering change. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (2nd ed., pp. 151–170). New York, NY: Cambridge University Press.

Brown, J. C., & Crippen, K. J. (IN PRESS). Designing for culturally responsive science education through professional development. International Journal of Science Education. doi:10.1080/09500693.2015.1136756

Confrey, J. (2006). The evolution of design studies as methodology. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 135–152). Cambridge, NY: Cambridge University Press.

McKenney, S., Kali, Y., Markauskaite, L., & Voogt, J. (2015). Teacher design knowledge for technology enhanced learning: An ecological framework for investigating assets and needs. Instructional Science, 1–22. doi:10.1007/s11251-014-9337-2

McKenney, S., & Reeves, T. C. (2012). Conducting educational design research. New York, NY: Routledge.

Penuel, W. R., Fishman, B. J., Haugan, C. B., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. Educational Researcher, 40(7), 331–337. doi:10.3102/0013189x11421826

Stephen W. Loke Stephen W. Loke Loke, Stephen W.

Jamie C. McGovern Jamie C. McGovern McGovern, Jamie C.

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Developmental Disabilities Developmental disabilities

493

498

Developmental Disabilities

Developmental disabilities are a group of conditions that begin during conception or early development and result in physical, learning, language, and behavior impairments. This entry outlines the definition and history of developmental disabilities, the causes of developmental disabilities, psychopathology associated with developmental disabilities, the effects on individuals with developmental disabilities across multiple contexts, treatments and supports for individuals with developmental disabilities, and the need for early intervention and its associated benefits.

Definition of Developmental Disabilities

The Developmental Disabilities Assistance and Bill of Rights Act of 2000 (Public Law 106–42)defines developmental disability as a severe and chronic disability that is due to a mental and/or physical impairment, occurs before age 22, is likely to continue indefinitely, requires individualized services or treatments across the individual’s lifetime or for a prolonged duration, and results in functional limitations in at least three of the following areas: self-care, receptive and expressive language, learning, mobility, self-direction, capacity for independent living, and economic self-sufficiency. In addition, individuals who have substantial developmental delay or specific congenital or acquired conditions from birth through age 9 may be considered to have a developmental disability without meeting three of the aforementioned criteria if they have a high probability of meeting the criteria later in life if services are not provided to them.

History of Developmental Disabilities

Historically, individuals with developmental disabilities have been mistreated, abandoned, and persecuted. In the mid-1800s, institutions to house individuals with developmental disabilities emerged and were popularly thought to keep society safe from them. However, deinstitutionalization was not promoted until the exposure of inhumane conditions at these institutions in the mid-1960s. This heralded in the movement toward independent living and self-advocacy.

In 1963, the Developmental Disabilities Assistance and Bill of Rights Act was the first legislation enacted to protect and to provide opportunities to individuals with developmental disabilities. This act funded programs, such as the Protection and Advocacy System to protect the rights of individuals with developmental disabilities and the University Affiliated Facilities, which was later renamed the University Centers for Excellence in Developmental Disabilities Education, Research and Service, to train providers to serve individuals with developmental disabilities, conduct and disseminate research relating to developmental disabilities, and to conduct community outreach efforts. Other legislation, including the Individuals with Disabilities Education Act and the American with Disabilities Act, also sought to integrate individuals with developmental disabilities into mainstream society by providing individuals with developmental disabilities with access to the same public education afforded to typically developing peers and to address discrimination that individuals with developmental disabilities face in society. In addition, in 1987, to foster greater opportunities for individuals with developmental disabilities to lead fulfilling lives and to achieve their maximum potential, Ronald Reagan, the then president of the United States, proclaimed March as National Developmental Disabilities Awareness Month.

Developmental Disability Conditions and Prevalence

Numerous developmental disabilities have been identified. Some of these include intellectual disabilities, attention-deficit/hyperactivity disorder, cerebral palsy, fetal alcohol syndrome, Down syndrome, and Williams syndrome.

The U.S. Centers for Disease Control and Prevention reported that approximately 14% of children had been at some point diagnosed with approximately 14% of children and adolescents were diagnosed with a developmental disability between 1997 and 2008. In addition, C. A. Boyle and colleagues found that there has been an increase in the prevalence of developmental disabilities over time, with 12.8% of children and adolescents diagnosed as having a developmental disability in 1997–1999, whereas 15.04% of children and adolescents were diagnosed as having a developmental disability in 2006–2008. Among U.S. children and adolescents, there is also an increasing trend of children and adolescents being diagnosed with attention-deficit/hyperactivity disorder and autism over time. For example, there was a 289.5% increase in autism diagnosis from the period between 1997 and 1999 and the period between 2006 and 2008.

Gender differences among children diagnosed with a developmental disability have also been found. Although boys are generally more likely to be diagnosed with a developmental disability than girls, some developmental disabilities (e.g., Rett syndrome) are more likely to occur in girls than in boys. In addition, general ethnic differences have been found among children with developmental disabilities, with Hispanics being least likely to be diagnosed with a developmental disability.

Causes of Developmental Disabilities

There are many causes of developmental disabilities. Genetic or chromosomal causes have been identified for some developmental disabilities. For example, Down syndrome is caused by an extra chromosome 21, whereas fragile X syndrome is caused by a mutation in the *FMR1* gene. Environmental factors have also been found to contribute to developmental disabilities. These factors may include prenatal factors (e.g., maternal use of alcohol or drugs), birth complications (e.g., deprivation of oxygen at birth), and exposure to environmental toxins (e.g., exposure to lead or polychlorinated biphenyls). However, there are still some developmental disabilities that do not have a clear cause and may be caused by the complex interaction of both genetic and environmental factors.

Psychopathology Associated With Developmental Disabilities

Individuals with developmental disabilities are more likely to have mental health problems than those without developmental disabilities. Sometimes, practitioners find it more difficult to recognize symptoms, determine causes of symptoms, and accurately diagnose psychopathology among clients with developmental disabilities. This problem may be due to misattribution of symptoms and challenges presented by available psychodiagnostic tools. Misattribution of symptoms can occur when an individual with a developmental disability is experiencing a symptom of a mental health problem, but caregivers and clinicians attribute the symptom to the developmental disability rather than considering a potential coexisting mental health problem. It is also possible that clinicians and significant caregivers sometimes prefer to focus attention and energy on addressing the direct effects of a client's developmental disability rather than dividing resources between the developmental disability and a mental health concern. A second cause of difficulty with understanding psychopathology in clients with developmental disabilities is the small number of mental health diagnostic tools designed for this population. Additionally, diagnostic workups of internalizing psychopathology, such as depression and anxiety, often rely on a client's self-report of symptoms, which can be difficult or impossible to obtain from an individual with a developmental disability who experiences difficulties with speech, language, and/or cognition. As such, to gather the necessary information for a diagnosis, Rush and colleagues recommend in their 2004 journal article that a multimethod assessment approach (e.g., using record reviews, interviews, observations, and rating scales) be used.

Effects of Developmental Disabilities

Effects on Adaptive Functioning

The effects of developmental disabilities can be quite varied in scope and severity. One common area that can be affected is an individual's adaptive functioning. Adaptive skills are those skills that allow individuals to perform everyday tasks without assistance. Adaptive skills can be grouped into many different categories such as communication skills, social skills, domestic skills, personal care skills, and community living skills. Adaptive skills in the communication domain might include making appropriate requests for needed objects or demonstrating an understanding of instructions. An example of a domestic skill included under the umbrella of adaptive abilities could be keeping one's living area clean or caring for one's home in other ways. Personal care skills deal with independent completion of activities such as bathing, dressing, and toileting. Community living skills can include knowing how to get from place to place in one’s community, using money, and accessing needed community resources. When individuals with developmental disabilities have low adaptive functioning, they often need assistance in their daily lives. Assistance provided to support adaptive functioning could range from regular check-ins with a case manager to round-the-clock care from personal attendants or aides.

Effects on Health

Individuals with developmental disabilities can also have a higher incidence of certain medical and health problems than those without developmental disabilities. Some developmental disabilities are linked to specific medical problems. For example, individuals with Down syndrome are more likely to have congenital heart disease. Individuals with different developmental disabilities also may have a variety of other health or medical concerns. Feeding issues, such as sensitivities to certain textures, extreme pickiness, or swallowing problems, can occur in individuals with developmental disabilities. Digestive problems, such as stomachaches, constipation, or difficulties with toilet training, can also be challenges for individuals with developmental disabilities. Sleep problems, such as having difficulty falling asleep, not sleeping through the night, and having an early waking time can also be present among those with developmental disabilities. Finally, weight problems such as being overweight or obese can also occur in people with developmental disabilities. One possible cause for being overweight or obese may be more restricted options for physical activity among those with certain types of developmental disabilities.

Effects on Family Members

There is much variability in how siblings and parents of children with developmental disabilities are affected by their family member’s disability. Young siblings may have concerns about “catching” their brother or sister’s developmental disability. Typically functioning siblings may also perceive that attention and resources are centered on the sibling with a developmental disability. Parental stress due to the sibling’s developmental disability may also affect typically functioning siblings. In addition, typically functioning siblings may be asked to take on responsibilities that are not often experienced by their peers, including eventual care and decision making for a sibling with a developmental disability. More positively, siblings of an individual with a developmental disability may learn more about developmental disabilities and may feel more comfortable interacting with people who have them.

The effect of having a child with a developmental disability can also vary greatly for parents. Caring for a son or daughter with a developmental disability can place a strain on a family’s resources, depending on the child’s needs and the family’s resources. Initial adjustment to having a child with a developmental disability may be difficult or emotional for parents. Parents’ levels of stress may be higher for a variety of reasons. Some parents of children with developmental disabilities find that they take on the role of a case manager, nurse, and/or advocate in addition to that of a parent. Accessing appropriate care and services for a son or daughter with a developmental disability can also be difficult. Arranging for care and assistance for an individual with a developmental disability when parents will be unable to provide it can further be daunting and uncomfortable for some families.

Effects on School Functioning

Developmental disabilities can impact an individual’s ability to learn and function in the school environment. School systems can address the needs of students with developmental disabilities in several ways. Under Section 504 of the Rehabilitation Act of 1973, a student with a disability that has a significant effect on a major life activity is eligible to receive accommodations within the general education setting. These accommodations are meant to allow the child or adolescent to access the curriculum and the school environment. Students with developmental disabilities may also be eligible for special education services. Students with developmental disabilities may qualify for special education services under a number of categories, such as developmental delay, intellectual disability, other health impairment, or autism. In addition to meeting criteria for a disability category, there must be an adverse academic impact of the disability in order for the student to qualify for special education services. Early intervention services, otherwise known as infant-toddler services, are provided from birth up to 3 years of age for those who qualify under Part C of the Individuals With Disabilities Education Act. From age 3 through 21, children with developmental disabilities may qualify for special education services under Part B of the Individuals With Disabilities Education Act.

When a child qualifies for infant-toddler services, an individualized family service plan is provided. Similarly, when a student qualifies for preschool or school-aged special education services, an individualized education program is provided. These plans can include numerous services and supports, depending on the needs of the child or adolescent. These supports are described in the next section.

Treatments and Supports for Individuals With Developmental Disabilities

Although there is no cure for individuals with developmental disabilities, numerous therapeutic treatments that may improve the individual’s functioning are available. Speech/language therapy may improve the individual’s speech and language skills, whereas physical and occupational therapy may improve the individual’s gross motor and fine motor skills, respectively. In addition, children with an autism spectrum disorder may benefit from applied behavior analysis. These therapeutic treatments may also be provided in school as part of the student’s individualized education program.

Other individualized education program services to support children with developmental disabilities may include specialized academic instruction, assistance in the regular education classroom, counseling, behavioral support, and social skills instruction. In addition, a student’s plan can include assistive technology, transportation services to and from school, and transition services to help plan for the student’s postgraduation goals and activities. Furthermore, this plan can also enumerate accommodations that will be made to help students complete their work or participate at school as well as any needed modifications to the students’ schoolwork.

Medications may be used to treat symptoms related to the developmental disability or to a mental health condition that the individual with a developmental disability may be diagnosed with. Stimulants may be used to treat symptoms related to attention-deficit/hyperactivity disorder, whereas antidepressants may be prescribed to treat depressive symptoms in individuals with developmental disabilities. In some cases, individuals with developmental disabilities may undergo surgery. For example, some individuals with an intellectual disability who also have epilepsy may consider epilepsy surgery.

Benefits of Inclusion

William R. Henninger IV and Sarika S. Gupta have indicated that children with developmental disabilities benefit from an inclusive education, especially when it begins at an early age. An inclusive education is when children with and without disabilities are educated together in the same classroom. Both children with and without disabilities benefit from being schooled in the same classroom. For children with developmental disabilities, opportunities to interact with typically developing children increase, and through these interactions with same-age peers, these children's social skills may improve and friendships may develop. In addition, children may gain more confidence in their own abilities and skills through their interactions with typically developing peers. Another benefit of inclusion for children with developmental disabilities is in the academic domain. When children with developmental disabilities are educated in the same classroom as children without disabilities, they are more likely to stay in school and graduate and to achieve more academically. For those children without disabilities, understanding and acceptance of children with developmental disabilities may occur when children are given the opportunity to interact with each other at an early age.

Early Intervention

Developmental Milestones and Screenings

Developmental milestones are skills that children should acquire by a certain age and children with developmental disabilities show delays in social, emotional, communication, cognitive, and/or motor skills. Children with developmental disabilities may not respond to loud sounds, coo, roll over, sit up, wave one's hand, walk, or speak in sentences at expected ages. When problems are observed in children not reaching their developmental milestones at expected ages, these children should be referred for a developmental screening. Developmental screenings can be conducted by health-care professionals to determine whether children are learning basic skills when they are expected to acquire those skills. For those children who are showing delays, early intervention treatment services can be offered to improve these children's skills.

Outcomes for Early Intervention

Outcomes of early intervention services are important to document to determine whether these treatments are effective for children with developmental disabilities and their families. Geraldine Dawson and colleagues and Kathleen Hebbeler and colleagues have found that early intervention services for children with developmental disabilities have been shown to have a positive impact on children’s challenging behaviors and cognitive development, and their communication skills and socioemotional development, respectively. Earlier and more intensive interventions seem to be more effective than interventions administered with less intensity during the childhood years. Early intervention services may also be beneficial to families with a child with a developmental disability. Hebbeler and colleagues indicate that providing these services early to a child will not only help the child but may address the family’s concerns about the child as well as other family concerns and reduce the likelihood that special education services will be needed for the child in the future.

Stephen W. Loke, Jamie C. McGovern, and Patricia A. Lowe

See also Autism Spectrum Disorder; Individuals With Disabilities Education Act; Intellectual Disability and Postsecondary Education

Further Readings

Boyle, C. A., Boulet, S., Schieve, L. A., Cohen, R. A., Blumberg, S. J., Yeargin-Allsopp, M., & Kogan, M. D. (2011). Trends in the prevalence of developmental disabilities in US children, 1997–2008. Pediatrics, 127, 1034–1042. doi:10.1542/peds.2010-2989

Centers for Disease Control and Prevention. (2015). Developmental disabilities. Retrieved from http://www.cdc.gov/ncbddd/developmentaldisabilities/index.html

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., & Varley, J. (2010). Randomized controlled trial of the Early Start Denver Model, a developmental behavioral intervention for toddlers with autism: Effects on IQ, adaptive behavior, and autism diagnosis. Pediatrics, 125, e17–e23. doi:10.1542/peds.2009-0958

Hebbeler, K., Spiker, D., Bailey, D., Scarborough, A., Mallik, S., & Singer, M. (2007). Early intervention for infants … toddlers with disabilities and their families: Participants, services, and outcomes. Final report of the National Early Intervention Longitudinal Study (NEILS). Menlo Park, CA: SRI International.

Henninger, W. R., IV, & Gupta, S. S. (2014). How do children benefit from inclusion? In S. S. Gupta, W. R. Henninger IV, & M. E. Vinh (Eds.), First steps to preschool inclusion: How to jumpstart your programwide plan (pp. 33–57). Baltimore, MD: Brookes.

Rush, K. S., Bowman, L. G., Eidman, S. L., Toole, L. M., & Mortenson, B. P. (2004). Assessing psychopathology in individuals with developmental disabilities. Behavior Modification, 28, 621–637. doi:10.1177/0145445503259830

Michael Quinn Patton Michael Quinn Patton Patton, Michael Quinn

Developmental Evaluation Developmental evaluation

498

503

Developmental Evaluation

Developmental evaluation provides evaluative information and feedback to social innovators, and their funders and supporters, to inform adaptive development of change initiatives in complex dynamic environments. This entry defines evaluation, describes how developmental evaluation differs from other approaches to evaluation, and explains how developmental evaluation is informed by systems thinking and complexity theory. It then looks at the methods and guiding principles of developmental evaluation and trends influencing the future of developmental evaluation.

In 2014, an American Evaluation Association task force chaired by Michael Quinn Patton defined evaluation as “a systematic process to determine merit, worth, value, or significance.” The task force’s statement continued:

> Program evaluation answers questions like: To what extent does the program achieve its goals? How can it be improved? Should it continue? Are the results worth what the program costs? Program evaluators gather and analyze data about what programs are doing and accomplishing to answer these kinds of questions… . Because making judgments and decisions is involved in everything people do, evaluation is important in every discipline, field, profession and sector, including government, businesses, and not-for-profit organizations (American Evaluation Association task force, n.p.).

The Niche of Developmental Evaluation

The developmental evaluation niche focuses on evaluating innovations in complex dynamic environments because that’s the arena in which *social innovators* and *change agents* are working. Innovation as used here is a broad framing that includes creating new approaches to intractable problems, adapting programs to changing conditions, applying effective principles to new contexts (scaling innovation), catalyzing systems change, and improvising rapid responses in crisis conditions. Social innovation unfolds in social systems that are inherently dynamic and complex, and often turbulent. The implication for social innovators is that they typically find themselves having to adapt their interventions in the face of complexity and changing conditions. Funders of social innovation also need to be flexible and adaptive in alignment with the dynamic and uncertain nature of social innovation in complex systems.

Developmental evaluators track, document, and help interpret the nature and implications of innovations and adaptations as they unfold, both the processes and outcomes of innovation, and help extract lessons and insights to inform the ongoing adaptive innovation process. Developmental evaluation brings to innovation and adaptation the processes of asking evaluative questions, applying evaluation logic, and gathering and reporting evaluative data to inform and support the development of innovative projects, programs, initiatives, products, organizations, and/or systems change efforts with timely feedback.

At the same time, this provides accountability for funders and supporters of social innovations and helps them understand and refine their contributions to solutions as they evolve. Social innovators often find themselves dealing with problems, trying out strategies, and striving to achieve goals that emerge from their engagement in the change process, but which they could not have identified before that engagement, and that continue to evolve as a result of what they learn. The developmental evaluator helps identify and make sense of these emergent problems, strategies, and goals as the social innovation *develops*. The emergent/creative/adaptive interventions generated by social innovators for complex problems are significant enough to constitute *developments,* not just improvements, thus the need for *developmental* evaluation.

Traditional evaluation approaches advocate clear, specific, and measureable outcomes that are to be achieved through processes detailed in a linear logic model. Such traditional evaluation demand for upfront, preordained specificity doesn’t work under conditions of high innovation, exploration, uncertainty, turbulence, and emergence. Indeed, premature specificity can do harm and generate resistance from social innovators, as, indeed, it has, by constraining generate resistance from social innovators, as, indeed, it has, by constraining exploration, limiting adaptation, reducing experimental options, and forcing premature adoption of a rigid model, not because such a model is appropriate, but because evaluators, funders, or other stakeholders demand it in order to comply with what they understand to be good evaluation. Developmental evaluation emerged as a response to criticism of traditional evaluation by social innovators and their expressed need for an alternative way to engage in evaluation of their work.

Developmental evaluation involves evaluative thinking throughout. Judgments of merit, worth, significance, meaningfulness, innovativeness, and effectiveness (or such other criteria as are negotiated) inform ongoing adaptive innovation. Such evaluative judgments don’t just come at the end of some fixed period (for example, a 3-year grant); rather, they are ongoing and timely. Nor are evaluation conclusions reached and rendered by the evaluator independently. Developmental evaluation is a collaborative, interactive process. Because this process is utilization focused, and because it unfolds in complex dynamic systems where the particular meaning and significance of information may be difficult to predetermine, making sense together of emergent findings involves the developmental evaluators interpreting patterns in the data *collaboratively* with social innovators, their funders, advocates, change agents, and systems change supporters. Through this empirically focused interaction, developmental evaluation becomes an integral part of the innovative process.

How Is Developmental Evaluation Different From Other Approaches?

Because developmental evaluation claims a specific purpose and niche, questions about how it differs from other approaches are common.

Developmental Evaluation in Contrast to Formative and Summative Evaluation

Developmental evaluation offers an alternative to formative and summative evaluation, the classic distinctions that have dominated evaluation. A formative evaluation serves to improve a program. It answers questions about a program’s strengths and weaknesses. Summative evaluation, in contrast, renders an overall judgment of effectiveness at the end of some designated period of time. It answers questions about whether goals were attained and whether the program works well.

In its original conceptualization, the purpose of formative evaluation was to prepare a program for summative evaluation by identifying and correcting implementation problems, making adjustments based on feedback, providing an early assessment of whether desired outcomes were being achieved (or were likely to be achieved), and getting the program stabilized and standardized for summative assessment. It is not uncommon for a new program to go through 2–3 years of formative evaluation, working out startup difficulties and getting the program model stabilized, before a summative evaluation is conducted.

Over time, formative evaluation has come to designate any evaluative efforts to improve a program. Improvement means making it better. In contrast, developmental evaluation focuses on adaptive development, which means making the program different because, for example, (a) the context has changed (which comes with the territory in a complex dynamic environment); (b) the clientele have changed significantly; (c) learning leads to a significant change; or (d) a creative, innovative alternative to a persistent issue or challenge has emerged. Here are examples of such adaptive developments.

A program helping mainly White, low-income, high school dropouts complete their high school degrees adapts to demands to also serve a different population, for example, immigrants, people coming out of prison, or people with particular disabilities. This kind of adaptation goes beyond improvement. It requires developmental adaptation.

A workshop or course moves online from the classroom. Teaching effectively online requires major adaptation of both content and process, as well as criteria for interpreting success. Again, this goes well beyond ongoing improvement.

Drug abuse education adapts to include attention to new illegal drugs, such as synthetic marijuana. Innovation and adaptation of educational interventions become the order of the day as new drugs are created and wind up being peddled to children by drug dealers.

Development of child care options in high schools for teenage parents that can accommodate children from birth to age 5 years.

Development of a local food service that uses local food sources as a response to the failure of multinational food distribution to solve hunger and nutrition challenges.

The Relationship Between Developmental Evaluation and Development Evaluation

Developmental evaluation is easily and often confused with development evaluation. They are not the same, though developmental evaluation can be used in development evaluations. Development evaluation is a generic term for evaluations conducted in developing countries, usually focused on the effectiveness of international aid programs and initiatives. An evaluation focused on development assistance in developing countries could use a developmental evaluation approach, especially if such developmental assistance is viewed as occurring under conditions of complexity with a focus on adaptation to local context. But developmental evaluation is by no means limited to projects in developing countries. The "-al" in developmental is easily missed, but it is critical in distinguishing development evaluation from developmental evaluation.

How Systems Thinking and Complexity Theory Inform the Practice of Developmental Evaluation

Thinking systemically is fundamental to developmental evaluation. This means, at a minimum, understanding interrelationships, engaging with multiple perspectives, and reflecting deeply on the practical and ethical consequences of boundary choices. The shift in thinking required is from focusing on discrete components of a program to thinking in terms of relationships. Innovation involves changing an existing system at some level and in some way.

Findings from program evaluations show that projects and programs rarely lead to major change. Effective projects and programs are often isolated from larger systems, which allows them the autonomy to operate effectively but limits their larger impact. On the other hand, projects and programs often fail because they operate in dysfunctional systems. Thus, social innovators are interested in and motivated by changing systems—health-care systems, educational systems, food systems, criminal justice systems. In doing so, they engage in efforts and thinking that supersede traditional project and program logic models.

To evaluate systems change, developmental evaluators need to be able to engage in systems thinking and to treat the system or systems targeted for change as the *evaluand* (the thing being evaluated). This means inquiring into, tracking, documenting, and reporting on the development of interrelationships, changing boundaries, and emerging perspectives that provide windows into the processes, effects, and implications of systems change.

Thinking systemically comes into play even in small pilot projects. Systems and complexity concepts are helpful for understanding what makes a project innovative. Moreover, even small innovations eventually face the issue of what it will mean to expand the innovation if it is successful—which directly and inevitably will involve systems change. Developmental evaluation is attuned to both linear and nonlinear relationships, both intended and unintended interactions and outcomes, and both hypothesized and unpredicted results. Fundamental systems-oriented developmental evaluation questions include these: In what ways and how effectively does the system function for whose interests? Why so? How are the system's boundaries perceived? With what implications? To what extent and in what ways do the boundaries, interrelationships, and perspectives affect the way the innovative change process has been conceptualized and implemented? How has social innovation changed the system, through what processes, with what results and implications?

The Complexity Perspective

Viewing innovation through the lens of complexity adds another way of framing, studying, and evaluating social innovations. Innovations involve uncertain outcomes and unfold in situations where stakeholders typically disagree about the nature of the problem and what should be done to address it. These two dimensions, degree of uncertainty and degree of disagreement, define the zone of complexity. In essence, complexity theory directs our attention to characteristics and dimensions of dynamic systems change—which is precisely where innovation unfolds.

Core developmental evaluation questions driven by complexity theory include these: In what ways and how can the dynamics of complex systems be captured, illuminated, and understood as social innovation emerges? To what extent do the dynamics of uncertainty and disagreement shift and change during the unfolding of the innovation? How is innovation's development captured and understood, revealing new learning and knowledge that can be extrapolated or applied elsewhere?

The Methods Used in Developmental Evaluation

Developmental evaluation does not rely on or advocate any particular evaluation method, design, tool, or inquiry framework. A developmental evaluation can include any kind of data (quantitative, qualitative, and mixed), any kind of design (e.g., naturalistic and experimental), and any kind of focus (processes, outcomes, impacts, costs, and cost-benefit, among many possibilities)—depending on the nature and stage of an innovation, and on the priority questions that will support development of and decision making about the innovation. Methods and tools can include rapid turnaround randomized controlled trials, surveys, focus groups, interviews, observations, performance data, community indicators, and network analysis—whatever sheds light on key questions.

The process and quality of engagement between the primary intended users (social innovators) and the developmental evaluators is as much the method of developmental evaluation as any particular design, methods, and data collection tools are. Asking evaluation questions, examining and tracking the implications of adaptations, and providing timely feedback on an ongoing basis—these are the methods of developmental evaluation.

Whatever methods are used or data are collected, rapid feedback is essential. Speed matters. Dynamic complexities don’t slow down or wait for evaluators to write their reports, get them carefully edited, and then have them approved by higher authorities. Any method can be used, but it will have to be adapted to the necessities of speed, timely reporting, and just-in-time, in-the-moment decision making. This is a major reason why the developmental evaluators may be part of the innovation team: to be present in the real time as issues arise and decisions have to be made.

Methods can be emergent and flexible; designs can be dynamic. Contrary to the usual practice in evaluation of fixed designs that are implemented as planned, developmental evaluation designs can change as an innovation unfolds and changes. If surveys and interviews are used, the evaluators may change questions from one administration to the next, discarding items that have revealed little of value or are no longer relevant, and adding items that address new issues. The sample can be emergent as new participants or sites emerge, and others are abandoned. Both baselines and benchmarks can be revised and updated as new information emerges.

Developmental Evaluator Skills

Developmental evaluators need to be agile, open, interactive, flexible, observant, and highly tolerant of ambiguity. A developmental evaluator is, in part, an instrument. Because the evaluation is cocreated and the developmental evaluator is part of the innovation team, bringing an evaluation perspective and evaluative thinking to the team, an evaluator's capacity to be part of the team and facilitate the evaluation elements of the innovative process involves both essential "people skills" and is part of the method for developmental evaluation.

The Eight Guiding Principles of Developmental Evaluation

Developmental evaluation is not a set of methods, tools, or techniques. There isn't a set of steps to follow. There's no recipe, formula, or standardized procedures. Rather, developmental evaluation is a way of approaching the challenge of evaluating social innovation through guiding principles. There are eight guiding developmental evaluation principles: 1. *Developmental purpose principle*: Illuminate, inform, and support what is being developed, by identifying the nature and patterns of development (innovation, adaptation, and systems change), and the implications and consequences of those patterns.

2. *Evaluation rigor principle:* Ask probing evaluation questions, think and engage evaluatively, question assumptions, apply evaluation logic, use appropriate methods, and stay empirically grounded—that is, rigorously gather, interpret, and report data.

3. *Utilization focus principle*: Focus on intended use by intended users from beginning to end, facilitating the evaluation process to ensure utility and actual use.

4. *Innovation niche principle*: Elucidate how the change processes and results being evaluated involve innovation and adaptation, the niche of developmental evaluation.

5. *Complexity perspective principle:* Understand and interpret development through the lens of complexity and conduct the evaluation accordingly. This means using complexity premises and dynamics to make sense of the problems

being addressed; to guide innovation, adaptation, and systems change strategies; to interpret what is developed; to adapt the evaluation design as needed; and to analyze emergent findings.

6. *Systems thinking principle*: Think systemically throughout, being attentive to interrelationships, perspectives, boundaries, and other key aspects of the social system and context within which the innovation is being developed and the evaluation is being conducted.

7. *Cocreation principle*: Develop the innovation and evaluation together—interwoven, interdependent, iterative, and cocreated—such that the developmental evaluation becomes part of the change process.

8. *Timely feedback principle*: Time feedback to inform ongoing adaptation as needs, findings, and insights emerge, rather than only at predetermined times (e.g., quarterly or at midterm and end of project).

Social Change and the Future of Developmental Evaluation

The future of developmental evaluation depends on four intersecting social change trends, with developmental evaluation sitting at the point where these trends converge. First is the worldwide demand for innovation. The private sector, public sector, and nonprofit sector are all experiencing pressure to innovate. As the world's population grows, climate change threatens, and technology innovations expand horizons and possibilities exponentially (to mention just three forces for change), social innovation is recognized as essential to address global problems.

The second trend consists of systems change. Project-level evaluation doesn't translate directly into systems change evaluation. Treating a system as a unit of analysis—that is, as the evaluand (thing evaluated)—requires systems understandings and systems thinking. Developmental evaluation brings a systems orientation to evaluating systems change.

The third trend is complexity. Innovation and systems thinking point to complexity theory as the relevant framework for making sense of how the world is changed. Attention to and appreciation of complexity seem likely to become increasingly important in the context of global systems challenges.

The fourth trend is the acknowledgment of developmental evaluation as a legitimate and useful evaluation specialization with trained developmental evaluators available to meet the increasing demand for and conduct of developmental evaluations.

Michael Quinn Patton

Adapted with permission from Patton, M. Q., McKegg, K., … Wehipeihana, N. (2016). Preface. In M. Q. Patton, K. McKegg, … N. Wehipeihana (Eds.), *Developmental evaluation exemplars: Principles in practice* (pp. v–x). New York, NY: Guilford Press, and Patton, M. Q. State of the art and practice of developmental evaluation (pp. 1–24), and The developmental evaluation mindset: Eight guiding principles (pp. 289–312), in M. Q. Patton, K. McKegg, … N. Wehipeihana, eds., *Developmental evaluation exemplars: Principles in practice*. New York, NY: Guilford Press.

See also Empowerment Evaluation; Evaluation; Evaluation, History of; Evaluation Consultants; Evaluation Versus Research; External Evaluation; Formative Evaluation; Logic Models; Program Evaluation; Progress Monitoring; Summative Evaluation; Utilization-Focused Evaluation

Further Readings

American Evaluation Association Task Force (2014, January 28). What is evaluation? Retrieved from http://www.eval.org/p/bl/et/blogid=2…blogaid=4

Dickson, R., & Saunders, M. (2014). Developmental evaluation: Lessons for evaluative practice from the SEARCH Program. Evaluation, 20(2), 176–194.

Gamble, J. A. (2008). A developmental evaluation primer. Montreal, Canada: J. W. McConnell Family Foundation.

Lam, C. Y., & Shulha, L. M. (2014). Insights on using developmental evaluation for innovating: A case study on the cocreation of an innovative program. American Journal of Evaluation, 36(3), 358–374.

Patton, M. Q. (2011). Developmental evaluation: Applying complexity concepts to enhance innovation and use. New York, NY: Guilford Press.

Patton, M. Q. (2016). The developmental evaluation mindset: Eight guiding principles. In M. Q. Patton, K. McKegg, & N. Wehipeihana (Eds.), Developmental evaluation exemplars: Principles in practice (pp. 289–312). New York, NY: Guilford Press.

Patton, M. Q. (2016). State of the art and practice of developmental evaluation. In M. Q. Patton, K. McKegg, & N. Wehipeihana (Eds.), Developmental evaluation exemplars: Principles in practice (pp. 1–24). New York, NY: Guilford Press.

Patton, M. Q. (2016). What is essential in developmental evaluation? American Journal of Evaluation, 37(2), 250–265.

Patton, M. Q., McKegg, K., & Wehipeihana, N. (Eds.). (2016). Developmental evaluation exemplars: Principles in practice. New York, NY: Guilford Press.

Patton, M. Q., McKegg, K., & Wehipeihana, N. (2016). Preface. In M. Q. Patton, K. McKegg, & N. Wehipeihana (Eds.), Developmental evaluation exemplars: Principles in practice (pp. v–x). New York, NY: Guilford Press.

Diana Joyce-Beaulieu Diana Joyce-Beaulieu Joyce-Beaulieu, Diana

Diagnostic and Statistical Manual of Mental Disorders Diagnostic and statistical manual of mental disorders

503

507

Diagnostic and Statistical Manual of Mental Disorders

The *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), published by the American Psychiatric Association, is considered the authoritative source within the United States for mental health diagnoses. The manual offers detailed guidance on mental health concerns across the life span from early childhood neurodevelopmental disorders to adult personality disorders and later geriatric neurocognitive disorders. Clinicians and researchers utilized this resource across multiple disciplines, including counseling, education, medicine, psychology, psychiatry, rehabilitation, and social work fields. Therefore, the *DSM* offers a common theoretical framework for understanding mental health issues and a recognized nomenclature to facilitate cross-discipline collaboration. In addition, the *DSM* coded diagnoses data collected through hospitals and treatment providers, yielding important national information on diagnoses trends, which then informs policy decisions for service provision, research funding, and educational initiatives. This entry begins by reviewing the history of the editions of the *DSM* and how the fifth edition of *DSM* (*DSM-5*) is organized. Next, the importance of the *International Classification of Diseases* (*ICD*) is considered, followed by a look at how symptoms and measures in the *DSM* assist in diagnoses. Finally, the entry provides a warning about using the *DSM* without proper qualifications and considers changes that may be made to future revisions of the *DSM*.

History of the *DSM*

The first edition of the *DSM* (i.e., *DSM-I*) was published in 1952 and focused primarily on adult mental health needs across three classifications (i.e., organic brain disorders, functional disorders, and mental deficiency disorders). The manual also offered brief diagnostic descriptions of 106 subcategories from a psychobiological perspective often using the term “reactions” rather than symptoms. In 1968, the second edition of the *DSM* (i.e., *DSM-II*) shifted to a psychoanalytic approach to understanding mental health, and disorders were described in more detailed narratives. Although an important manual, the early editions of the *DSM* were very brief and yielded only moderate agreement among clinicians on diagnosis, given their lack of symptom specificity.

The third edition of the *DSM* (i.e., *DSM-III*), published in 1980, offered detailed lists of specific symptoms for disorders and was designed in a multiaxial format for diagnosis. The patient’s functioning was coded across five axes or domains in an effort to better document environmental, daily life functioning, and psychosocial impact of impairment. Clinicians coded most mental health diagnoses on Axis I (e.g., depression, anxiety); however, personality disorders and intellectual disability (denoted as *mental retardation*) were noted on Axis II. When more than one diagnosis was indicated, comorbidities were reported, noting the primary clinical syndrome. Axis III was reserved for documenting medical conditions relevant to the individual’s mental health functioning. The purpose of Axis III was to alert care providers to the reciprocal effects between physical illness and mental health, as each may negatively or positively impact the other domain. Psychosocial events and environmental factors that may strongly impact a patient’s functioning were listed in Axis IV. This axis acknowledged the influence that incidents may have on an individual’s presenting emotional state (e.g., bereavement). In Axis V, a measure of overall functioning provided in the *DSM-III*, known as the Global Assessment of Functioning Scale, was reported. This scale served to document the severity of negative impact and could be reassessed over time to note changes based on treatment. The *DSM-III* also added more child and adolescent diagnostic criteria. Additionally, it provided significant information on the etiology, prevalence, and associated features of diagnoses.

The fourth edition of the *DSM* (*DSM-IV*) was released in 1994 and followed in 2000 with a text revision that corrected some errors in the *DSM-IV*. These editions, like the *DSM-III*, followed the multiaxial format and included the Global Assessment of Functioning Scale. There were significant improvements in utilizing extensive expert work groups during development of the manual to enhance empirically supported decisions for diagnostic validity and reliability.

Specifiers and subtypes within syndromes also were added to improve specificity in diagnosis that could best inform treatment needs.

The fifth edition (i.e., *DSM-5*) was published in 2013 and was again preceded by multiple years of research, expert panel reviews as well as large field studies at medical centers to verify diagnostic criteria. The *DSM-5* framework moved away from a multiaxial format and the use of the Global Assessment of Functioning to a life span approach. Diagnoses that occur early in life (e.g., language disorders, autism spectrum disorder) are presented first in the manual from a neurodevelopmental viewpoint followed by diagnoses occurring later in life (e.g., neurocognitive disorder due to Alzheimer's) from a neurocognitive perspective.

Diagnostic and Statistical Manual of Mental Disorders

Organization of the *DSM-5*

Twenty-two chapters of diagnoses are organized by common etiology (e.g., internalizing, externalizing, and neurocognitive). Across syndromes, diagnostic symptoms, cultural issues, gender differences, specifiers, prevalence, development/course, risk/prognosis, comorbidities, and differential diagnosis factors are delineated in each diagnostic chapter. The *DSM-5* also introduces a section on cross-cutting measures of symptoms, cultural formation interviewing, and proposed criteria for possible future diagnoses (e.g., Internet gaming, caffeine use disorder) that may be included in the next version of the *DSM*. Lastly, the *DSM-5* appendices offer lists of codes for the 9th and 10th revisions of the *International Statistical Classification of Diseases and Related Health Problems* (*ICD-9* and *ICD-10*). The *ICD* manuals are published by the World Health Organization as a classification system for monitoring and analyzing human health problems (both physical and mental health). Codes also provide statistical data for cause of death worldwide. The *DSM-5* includes the *ICD* codes, as practitioners in medical settings will utilize these to cross code both *DSM* and *ICD* diagnoses. In fact, since 2014 the National Center for Health Statistics and Centers for Disease Control and Prevention have required *ICD-10* codes on health data including mental health hospitalizations.

ICD

As far back as 1853, the International Statistical Congress initiated efforts to devise a tracking system for understanding mortality, including deaths attributed to mental health concerns across the world. Even earlier versions of death records in London in the 1500s noted lunacy as a cause of death. There are a number of benefits to understanding patterns of mental health hospitalization across countries. For researchers, these data can facilitate cross-cultural studies of mental health syndromes to better understand cultural influences on behavior. For policy makers, these data can serve as an early warning system for emerging trends of mental health needs and to foster international collaborations in developing service models. In a global society, understanding *ICD* also offers a common classification system and nomenclature to facilitate communication between service providers.

Diagnosis Measurement and Evaluation

Historically, diagnosis depended on the expert knowledge of practitioners and their subjective assessment regarding whether the presence and severity of symptoms was consistent with the symptoms of a particular syndrome. Knowledge of an individual’s symptoms was based primarily on observations and interviews conducted in the clinician’s office. This process of diagnosis based on observation and interview was known as exercising clinical judgment and resulted in a yes or no (i.e., categorical) decision regarding whether an individual met criteria for a disorder. Although there were multiple assessment tools within the field of psychology to gather objective and dimensional data for symptoms, the inclusion of these data was not formally discussed in the *DSM* and thus was left to the individual preference of practitioners. The benefit of dimensional data is that they offer discrete measurement along a scale and may assist in identifying subclinical, at-risk individuals prior to meeting full diagnosis criteria, thus facilitating prevention and early intervention. Additionally, dimensional data facilitate the ability to measure even small change within a category. For example, the improvement of a patient in therapy may be measured even though the individual’s symptoms remain present as does the diagnosis.

DSM-5 Cross-Cutting Symptom Measures

For the first time, the *DSM-5* introduced a series of symptom measures acknowledging the limitations of strictly categorical diagnostic processes. These cross-cutting symptom measures are available in the manual and at the publisher's website and offer a set of Likert-type scale ratings that can structure and quantify the presence and severity of symptoms. There are two levels of cross-cutting symptom measures: the adult self-report or informant measures across 13 domains (e.g., depression, anxiety, and substance use) and the parent-reported measures for children across 12 domains (e.g., anger, sleep, anxiety, and depression). The cross-cutting measures can prompt practitioners to explore symptoms within a specific area that may require a diagnosis and/or be used to track progress of time for treatment outcomes. Additionally, the *DSM-5* website also offers several syndrome-specific rating scales, a disability measure of daily functioning from the World Health Organization, and personality inventories. The *DSM-5* symptom measures do have some limitations in that they are not norm referenced, and not all syndromes are covered.

Other Symptom Measures

In addition to the *DSM-5* cross-cutting symptom measures, there also are a plethora of other psychological instruments to assist clinicians in assessing the presence, frequency, and severity of mental health disorders. Many practitioners will supplement their clinical judgment with a variety of these objective and quantitative measures including norm-referenced data, so that symptoms are compared to their relative presence in both the general population and clinical populations. Omnibus rating scales offering a sampling of items across common mental health diagnoses (e.g., depression, attention-deficit/hyperactivity disorder) can aid practitioners in discerning the disorders most likely present. These measures can be especially helpful when patients report a wide range of symptoms or note symptoms that overlap across syndromes. Single-construct measures (e.g., Yale–Brown Obsessive Compulsive Scale) provide items for a specific syndrome with more thorough coverage of all symptoms in that syndrome. These may be most helpful when a clinician is clear of the specific syndrome and is most interested in symptom type, pervasiveness, or severity.

Some rating scale instruments offer multiple versions of assessments including parent, teacher, and self-report formats. Additionally, some assessment systems include brief screener versions to be administered to groups to identify individuals at risk and progress monitoring measures utilized to track small changes (e.g., Behavioral Assessment System for Children, 2nd edition). Progress monitoring instruments may be especially useful in assessing treatment outcomes. Additionally, many mental health measures will include scales to outcomes. Additionally, many mental health measures will include scales to detect lying, inconsistent answers, and bias reporting. These types of indicators are useful in helping clinicians detect responses that may not be valid for diagnoses. Lastly, a wide range of psychopathology instruments are designed to measure maladaptive behaviors or traits (e.g., Minnesota Multiphasic Personality Inventory; Adult Suicidal Ideation Questionnaire) that may cross-diagnose and warrant therapy.

Qualifications to Utilize the *DSM*

It should be noted that the *DSM* is a technical, complex, and scholarly guide to diagnosis, offering discrete criteria for disorders but also presumes significant training in the principles of psychology. An in-depth understanding of typical and atypical development, human behavior, cognitive processes, as well as psychopathology are needed to undergird the ability to provide sound clinical decisions in making mental health diagnoses. Professional credentials also are required to provide diagnoses, justify treatment recommendations, and provide mental health services.

Future Directions

With each edition of the *DSM*, validity and reliability of diagnoses are improved. It is likely this trend will continue as each additional version has access to new emerging research to better understand mental health functioning. Expected *DSM* trends include further debate regarding the categorical and dimensional approaches to establishing symptom severity and acknowledgment that there are still many overlapping symptoms across disorders that may need greater clarity. Additionally, the *DSM-5* notes several new disorders under consideration for *DSM-6*. Lastly, as editions of the *ICD* and *DSM* continue to align, the utilization of a dual diagnostic coding system will no doubt continue to be a point of discussion.

Diana Joyce-Beaulieu

See also Diagnostic Tests; Rating Scales

Further Readings

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychiatric Association. (2016). Guide to using DSM-5 in the transition to ICD-10. Retrieved from http://www.dsm5.org/Documents

First, M. B. (2013). DSM-5 handbook of differential diagnosis. Washington, DC: American Psychiatric Association.

Kupfer, D. J., Kuhl, E. A., & Regier, K. L. (2013). *DSM-5*: The future arrived. Journal of the American Medical Association, 309(16), 1691–1692.

Madden, R., Sykes, C., & Ustun, T. B. (2007). World health organization family of international classifications: Definition, scope, and purpose. Geneva, Switzerland: World Health Organization.

Sattler, J. M. (2014). Foundations of behavioral, social, and clinical assessment of children (6th ed.). La Mesa, CA: Author.

World Health Organization. (2004). International statistical classification of diseases and related health problems (10th ed.). Geneva, Switzerland: Author.

Laine Bradshaw Laine Bradshaw Bradshaw, Laine

Diagnostic Classification Models Diagnostic classification models

507

512

Diagnostic Classification Models

The term *diagnostic classification models* (DCMs) refers to a family of psychometric models that are used in education to provide statistically driven classification of examinees according to mastery levels of a predefined set of knowledge components, skills, or abilities. The knowledge components, skills, or abilities are typically called *attributes*. Attributes comprise the construct of interest for a diagnostic assessment—they are the latent variables that the assessment is designed to measure. The distinguishing feature of DCMs from other latent variable models—for example, item response theory or factor analysis—is that the latent attributes are assumed to have a categorical distribution instead of a continuous distribution. As a result, DCMs classify examinees into groups instead of scaling examinees along a continuum. This entry describes the theoretical underpinnings of DCMs and explains the statistical model form of the general family of DCMs. The entry concludes with a discussion of the utility of DCMs to support both educational assessments and educational research.

Statistical Foundations of DCMs

The categorical attributes in diagnostic assessments commonly are assumed to be binary and follow a Bernoulli distribution, though they could have more than two levels and follow a categorical distribution. This entry focuses on DCMs for binary attributes and uses mastery and nonmastery as the labels for the two levels of an attribute. In practice, the appropriate labels for attribute levels depend on the context and purpose of the assessment. Examples of other labels include on-track versus needs improvement and proficient versus emerging.

For a diagnostic assessment that measures A binary attributes, there are 2^A combinations of attribute mastery levels. Each combination represents a unique attribute pattern, or latent class, into which examinees can be classified. The attribute patterns, additionally known as *attribute profiles*, are denoted by α_c= [$\alpha_{c1}\alpha_{c2}\ldots\alpha_{cA}$], where $c \in \{1,2, \ldots,2^A\}$; $\alpha_{ca} = 1$ if attribute a is mastered in profile c and $\alpha_{ca} = 0$ if attribute a is not mastered in profile c. As an example, with three attributes, there are 2^3 or 8 possible attribute profiles: [000], [001], [010], [011], [100], [101], [110], and [111].

DCMs as Confirmatory Latent Class Models

The attributes are operationalized, or thoroughly defined, as part of designing a diagnostic assessment; thus, the latent classes into which examinees will be classified are defined prior to analyses of response data collected from the diagnostic assessment. This feature of DCMs make them a special case of a larger family of models known as latent class models: DCMs are confirmatory latent class models because the number and the nature of the latent classes are hypothesized and specified prior to analyses.

The general latent class model defines the probability of a scored item response vector (denoted $\boldsymbol{x}_e$) for a given examinee e as a function of the attribute profile c of the examinee ($\boldsymbol{\alpha}_e = \boldsymbol{\alpha}_c$) as:

$$P(\boldsymbol{X}_e = \boldsymbol{x}_e) = \sum_{c=1}^{2^A} \upsilon_c \prod_{i=1}^{I} \pi_{i|\alpha_e}^{x_{ei}} \left(1 - \pi_{i|\alpha_e}\right)^{1-x_{ei}} .$$

This equation has two main components: the *structural component*, which describes the relationships and distributions of the attributes, and the *measurement component,* which specifies the relationships between the attributes and items. The structural parameter υ_c represents the proportion of examinees who are members of latent class c. Because the classes, defined by attribute patterns, are exhaustive and mutually exclusive, these proportions sum to 1 . The structural model is commonly parameterized by using a log-linear model where attributes are predictors of the class proportions or by specifying a higher order structure where attributes are predictors of one or more higher order continuous factors. Using either method, marginal proportions, or *base rates*, of mastery for individual attributes, as well as correlations of attribute pairs, can be derived.

The measurement parameter represents the probability that examinee *e* provides the correct response for item *i* ($x_{ei} = 1$), given his or her attribute pattern ($\boldsymbol{\alpha}_e$). For items that measure all attributes on the assessment, there are as many unique ’s as there are classes. On diagnostic assessments, however, many items measure only one or two attributes. This creates equivalence classes in terms of the conditional item response probability: Attribute profiles that have the same attribute status for the subset of attributes measured by an item will have the same conditional item response probability.

As with the attribute definitions, the subset of attributes that are measured by each item are hypothesized as part of the diagnostic assessment development process. These hypotheses guide the specifications of which latent classes have equivalent item response probabilities for each item. Similar to a test blueprint, the alignment of items with attributes is often expressed in an item by attribute matrix known as a Q-matrix. Entries of “1” in the Q-matrix denote the item is hypothesized to measure the attribute, whereas entries of “0” denote the item is not hypothesized to measure the attribute. DCMs assume that variations in item responses can be completely accounted for by the attributes that the items measure. This assumption is known as *local independence*, a common assumption for psychometric models, and states that item responses are assumed to be independent conditional on the examinee’s attribute pattern. Thus, this assumption is closely tied to the correctness of the Q-matrix: If additional attributes exist that influence the item response and were not specified in the Q-matrix, the assumption is likely violated. Due to this assumption, the joint probability of the item responses across the assessment, expressed in the product portion of Equation 1, is the simple product of the independent, conditional item response probabilities.

Modeling Item–Attribute Relationships

A number of DCMs exist and differ in how they model item–attribute relationships, which can be seen in the differences in how they parameterize as a function of the attributes. More recently developed DCMs express the relationship of attributes as predictors of item responses in a generalized linear mixed model form that is familiar in statistics. Robert Henson, Jonathan

Templin, and John Willse (2009) unified a collection of earlier developed—and seemingly different—DCMs by demonstrating how to reparameterize these models as a generalized linear mixed model. They also introduced the saturated form of the generalized linear mixed model that includes interactions among the latent attributes, which yielded a new, general DCM that subsumed other commonly used DCMs. This entry focuses on explaining this saturated form, termed the *log-linear cognitive diagnosis model* (LCDM), as well as demonstrating the flexibility it provides to yield earlier developed DCMs by imposing parameter constraints.

Log-Linear Cognitive Diagnosis Model: A General DCM

For dichotomously scored response data, the LCDM models the item response function using a logistic regression form where the log odds of correct response are equal to a linear predictor *k*:

$$\log\left(\frac{\pi_{i|\alpha_e}}{1-\pi_{i|\alpha_e}}\right)=\log\left(\frac{P(x_{ei}=1\mid\boldsymbol{\alpha}_e)}{P(x_{ei}=0\mid\boldsymbol{\alpha}_e)}\right)=k.$$

The conditional probability of a correct response can be expressed by the inverse logit function:

$$P(x_{ei}=1\mid\boldsymbol{\alpha}_e)=\frac{\exp(k)}{1+\exp(k)}.$$

The linear predictor of the LCDM is similar to the linear predictor in an analysis of variance model in that it contains only categorical predictors, which are the examinees’ attribute levels. In the most general form of the LCDM:

$$k=\lambda_{i,0}+\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_i).$$

The intercept $\lambda_{i,0}$ is the log odds of a correct response for examinees who have not mastered any of the attributes measured by item *i*. The term is a condensed notation to express a sum of analysis of variance–like main and interaction effects. The row vector contains the main effects and interactions, where T represents the transpose. The term $\boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i)$ is a column vector of 0s and 1s that correspond to the terms in , where 1s indicate the parameter in is present in the linear predictor of a given examinee and item and 0s indicate the parameter is not. The entries of the Q-matrix for item i are given in vector $\boldsymbol{q}_i=[q_{i1}, q_{i2},\ldots, q_{iA}]^T$ and recall the vector $\boldsymbol{\alpha}_e$ is the attribute pattern for examinee e (i.e., $\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2},\ldots,\alpha_{eA}]$). Thus, an element of $\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_i)$ equals 1 when (a) the item measures the attribute(s) corresponding to the effect (q_{ia}'s=1), *and* (b) the examinee possesses the attribute(s) corresponding to the effect (α_{ea}'s=1). Otherwise the element equals 0. The terms are expanded as:

$$\lambda_i^T \boldsymbol{h}(\alpha_e, q_i) = \sum_{a=1}^{A} \lambda_{i,1(a)}(\alpha_{ea} q_{ia}) + \sum_{a=1}^{A-1} \sum_{b=a+1}^{A} \lambda_{i,2(ab)}(\alpha_{ea}\alpha_{eb} q_{ia} q_{ib}) + \ldots,$$

where $\lambda_{i,1(a)}$ is the main effect for attribute a on item i and $\lambda_{i,2(ab)}$ is the two-way interaction effect between attributes a and b for item i. Note the second subscript for these terms indicates the level of the effect, where the intercept level is 0, the main effect level is 1, and the two-way interaction level is 2. The ellipses indicate the equation continues for three-way through A-way interactions to allow for items that measure any combination of attributes.

An Example Item

To illustrate the LCDM, consider a diagnostic assessment that measures three attributes and an example item i that measures two of the attributes: Attribute 2 and Attribute 3. The linear predictor of the LCDM for examinee e is equal to:

$$\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1}q_{i1} + \lambda_{i,1,(2)}\alpha_{e2}q_{i2}$$
$$+ \lambda_{i,1,(3)}\alpha_{e3}q_{i3} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2}q_{i1}q_{i2}$$
$$+ \lambda_{i,2,(1,3)}\alpha_{e1}\alpha_{e3}q_{i1}q_{i3} + \lambda_{i,2,(2,3)}\alpha_{e2}\alpha_{e3}q_{i2}q_{i3}$$
$$+ \lambda_{i,3,(1,2,3)}\alpha_{e1}\alpha_{e2}\alpha_{e3}q_{i1}q_{i2}q_{i3}.$$

Because the item does not measure Attribute 1, $q_{i1} = 0$, reducing the equation to:

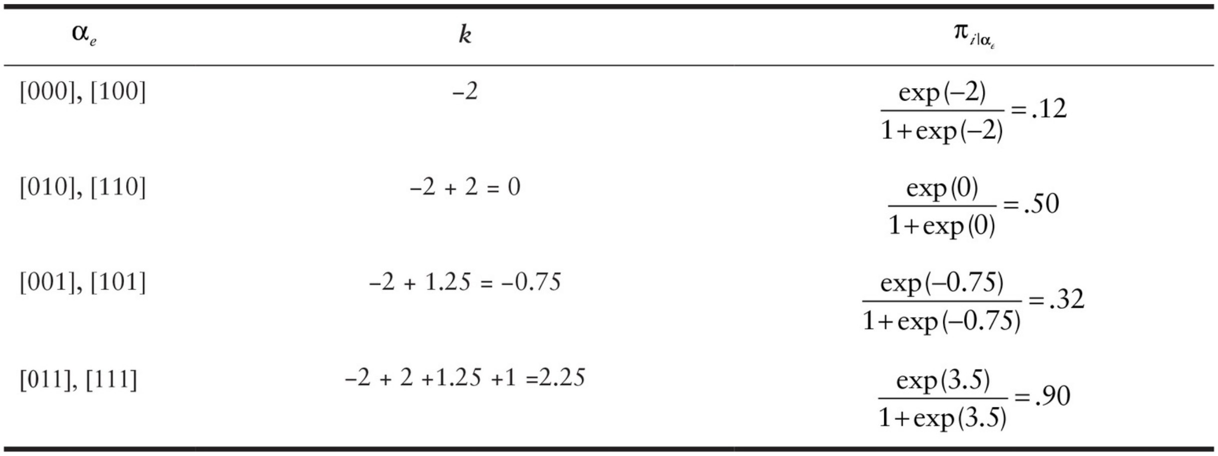
$$\lambda_{i,0} + \lambda_{i,1,(2)}\alpha_{e2}q_{i2} + \lambda_{i,1,(3)}\alpha_{e3}q_{i3}$$
$$+ \lambda_{i,2,(2,3)}\alpha_{e2}\alpha_{e3}q_{i2}q_{i3}.$$

Similarly substituting $q_{i2} = 1$ and $q_{i1} = 1$ because the item measures Attributes 2 and 3 yields,

$$\lambda_{i,0} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,1,(3)}\alpha_{e3} + \lambda_{i,2,(2,3)}\alpha_{e2}\alpha_{e3}$$

and looks like a familiar analysis of variance model with an intercept, two simple main effects, and an interaction effect that correspond to two binary predictors. For examinees who have mastered neither Attribute 2 nor Attribute 3 (i.e., profiles [000] or [100]), the linear predictor equals the intercept. For examinees who have mastered only Attribute 2 (i.e., profiles [010] or [110]), the linear predictor equals $\lambda_{i,0}+\lambda_{i,1,(2}$), where the simple main effect of Attribute 2 ($\lambda_{i,1,(2)}$) represents the increase in the log odds of a correct response for mastering Attribute 2, conditioning on not having mastered Attribute 3. Main effects are required to be greater than zero to reflect the hypothesis that the item measures the attribute; examinees who have mastered the attribute cannot be predicted to have a lower probability of correct response than examinees who have not. Similarly, the linear predictor of profiles [001] and [101] equals $\lambda_{i,0}+\lambda_{i,1,(3)}$. Finally, for examinees who have mastered both required attributes (i.e., profiles [011] and [111]), the linear predictor equals $\lambda_{i,0}+\lambda_{i,1,(2\ )}+\lambda_{i,1,(3)}+\lambda_{i,2,(2,3)}$, where the interaction term $\lambda_{i,2,(2,3)}$ may be positive, zero, or negative depending on whether there is an under-or overadditive effect of mastering both attributes instead of only one attribute. The interaction terms have constraints similar to main effects to ensure that mastering additional attributes that are hypothesized to be measured by the item does not result in a decreased probability of correct response.

To further illustrate, assume $\lambda_{i,0}$=−2, $\lambda_{i,1,(2)}$=2, $\lambda_{i,1,(3)}$=1.25, and $\lambda_{i,2,(2,3)}$=1 for the example item. For each attribute profile, Table 1 shows the linear predictor *k* and the corresponding probability of correct response .



α_e	k	$\pi_{i\mid\alpha_e}$
[000], [100]	−2	$\frac{\exp(-2)}{1+\exp(-2)}=.12$
[010], [110]	−2 + 2 = 0	$\frac{\exp(0)}{1+\exp(0)}=.50$
[001], [101]	−2 + 1.25 = −0.75	$\frac{\exp(-0.75)}{1+\exp(-0.75)}=.32$
[011], [111]	−2 + 2 +1.25 +1 =2.25	$\frac{\exp(3.5)}{1+\exp(3.5)}=.90$

Source: author created specifically for this entry

Submodels of the LCDM

Submodels of the LCDM can be formed by imposing parameter constraints on the saturated form of the LCDM. This entry presents three of the most commonly used submodels. The first is called the *compensatory reparameterized unified model*. The *compensatory reparameterized unified model* linear predictor contains only the intercept and the main effects; all interaction terms are constrained to be zero. The second, the *deterministic inputs noisy and gate model*, is a completely *noncompensatory* model, meaning that lacking the mastery of one or more required attributes cannot be compensated by mastering other required attributes. In other words, all attributes are required to yield an increase in the predicted probability of correct response. To reflect this assumption, the *deterministic inputs noisy and gate* model linear predictor contains only two parameters for each item: the intercept and the highest order interaction term. For example, if the item measures three attributes, the linear predictor will contain an intercept and a three-way interaction term. As a result, only examinees who have mastered all three measured attributes will have a linear predictor of $\lambda_{i,0}+\lambda_{i,3,(a,\ b,\ c)}$. All other examinees will have the same, lower predicted log odds equal to the intercept value. The third model is a completely *compensatory* model where mastering at least one of the measured attributes fully compensates for lacking mastery in other measured attributes. This model, known as the *deterministic inputs noisy or gate model*, yields the same item response probability for any examinees who have mastered one or more of the measured attributes, which is higher than the item response probability for examinees who have mastered none of the measured attributes. The *deterministic inputs noisy or gate* model also estimates two parameters for each item. To illustrate, consider the example item in Equation 7 again. The *deterministic inputs noisy or gate* model constrains the absolute value of both main effects and the interaction term to be equal and also constrains the interaction term to be negative:

$$\lambda_{i,0}+\lambda_{i,1}\alpha_{e2}+\lambda_{i,1}\alpha_{e3}-\lambda_{i,1}\alpha_{e2}\alpha_{e3}.$$

As a result, the linear predictor is equal to $\lambda_{i,0}+\lambda_{i,1}$ for profiles [001], [101], [010], [110], [011], and [111], resulting in equivalence classes for all profiles where either Attribute 2, Attribute 3, or both are mastered. The remaining profiles, [000] and [100], have a linear predictor equal to the intercept only, resulting in a lower predicted response probability in comparison.

These three models were developed prior to the LCDM and impose the previously described constraints for every item on an assessment. The LCDM provides flexibility to allow items on the same assessment to have different functional forms, including the three described above. Analogous to methods in GLMs when seeking to use a simpler model to achieve parsimony, the model–data fit of constrained items should be compared to their saturated form to provide statistical evidence to support the assumptions made by the simpler model forms.

Extensions of DCMs

This entry focused on DCMs for dichotomously scored item responses—for binary response data. DCMs can also be applied to ordinal, nominal, or continuous response data. The distribution of the item response will determine which link function to use with the LCDM.

This entry also focused on the LCDM for attributes that are assumed to have nonunit correlations. In education, some attributes may represent knowledge components that must be mastered before other knowledge components (i.e., attributes) can be mastered such that complete dependencies exists between the two attributes. Such hierarchical relationships among attributes are known as *attribute hierarchies*. These hierarchies can be specified in the LCDM framework by eliminating classes, and the corresponding item parameters, that do not exist under the hypothesized attribute hierarchy. By eliminating unnecessary classes, the presence of attribute hierarchies simplifies the DCM and also allows for a more accurate representation of the learning theories underlying the diagnostic assessment design.

DCMs: A Tool to Support Educational Assessment Goals

DCMs provide a classification of each examinee as a master or nonmaster of each attribute. The DCM estimates the probability that the examinee is a master of the attribute and then classifies the student as a master or nonmaster of the attribute, depending on which classification is more likely. For example, suppose a student completed an assessment that measures three attributes and her or his probabilities of mastery of the attributes, respectively, are .14, .84, and .32. The student would be classified as a nonmaster of Attribute 1 because the probability that she or he is a master equals .14 and that is less than the probability that she or he is a nonmaster, which is equal to 1 − .14, or .86. Analogously, the student would be classified as a master of Attribute 2 and a nonmaster of Attribute 3.

The multivariate profiles of attribute mastery provided by DCMs are aligned with needs in education to give feedback to students and teachers about where students’ strengths and weaknesses are in terms of understanding. This type of assessment feedback can be used by teachers to inform their decisions about which students need instructional support in which areas. Our example student mentioned in the previous paragraph consistently demonstrated understanding of Attribute 2 but may need instructional support to further develop the knowledge components, skills, or abilities defined by Attributes 1 and 3.

DCMs are also well suited to provide this type of feedback under practical constraints in educational settings, namely, that time for assessment is limited. Simulation studies have demonstrated that assessments with as few as 6 to 8 items per attribute can yield reliabilities above .80. In comparison to traditional assessments that provide scores representing abilities on a continuum, these test lengths are considerably shorter. The gain in efficiency in DCMs comes from the different purposes of the assessments; classification into two groups is an easier psychometric task than scaling along a continuum. Thus, when deciding whether to use a traditional assessment or a diagnostic assessment, it is important to consider the purpose of the assessment as well as practical constraints. If classification into mastery levels according to multiple latent characteristics serves the purpose of assessment, then DCMs are a useful model and shorter assessments can be used. In contrast, if rank ordering students along one or more continuum is the purpose of the assessment, other continuous latent variable item response models are useful and longer assessments can be used.

DCMs: A Tool to Support Educational Research Goals

In addition to being a psychometric tool for enabling assessments with reasonable lengths to provide diagnostic information with high utility in classrooms, DCMs are also a tool for supporting the advancement of multivariate learning theories that underlie the design of the assessment. Theories about the knowledge components, skills, or abilities that are the target of an assessment can include complex relationships among components. Due to the efficiencies described in the preceding section, DCMs help facilitate the estimation of more dimensions and their relationships, which in turn, enables the study of these relationships. The DCM framework can be used to specify hypotheses about relationships among attributes as well as the relationships between items and attributes. Those hypotheses can then be evaluated through the collection of data and empirical analyses in order to provide evidence to bolster, or inform changes to, the underlying theory. This evidence can be used to refine the theory and ultimately the assessment design, which improves to the knowledge base about the given construct as well as the quality of the assessment feedback.

Laine Bradshaw

See also Cognitive Diagnosis; Diagnostic Tests; Formative Assessment; Generalized Linear Mixed Models; Item Response Theory; Latent Class Analysis; Psychometrics; Summative Assessment

Further Readings

Bradshaw, L. (2016). Diagnostic classification models. In A. Rupp & J. Leighton (Eds.), Handbook of cognition and assessment. Wiley-Blackwell.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. Psychometrika, 74(2), 191–210. doi:10.1007/s11336-008-9089-5

Leighton, J., & Gierl, M. (Eds.). (2007). Cognitive diagnostic assessment for education: Theory and applications. Cambridge University Press.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). Cognitively diagnostic assessment. Hillsdale, NJ: Erlbaum.

Rupp, A. A., Templin, J., & Henson, R. (2010). Diagnostic measurement: Theory, methods, and applications. New York, NY: Guilford.

Templin, J. (2016). Diagnostic assessment. In F. Drasgow (Ed.), Technology and testing: Improving educational and psychological measurement (pp. 285–304). New York, NY: Routledge.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. Journal of Classification, 30(2), 251–275. doi:10.1007/s00357-013-9129-4

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. Psychometrika, 79(2), 317–339. doi:10.1007/s11336-013-9362-0

Jessica Hoth Jessica Hoth Hoth, Jessica

Diagnostic Tests

Diagnostic tests

512

516

Diagnostic Tests

Diagnostic tests in education are measurement methods that aim at identifying specific aspects in the field of teaching and learning. In psychology and medicine, the term *diagnostic* refers to making a medical diagnosis of a disease. The term also sometimes refers to any test that produces a profile of scores with strengths and weaknesses. This entry does not discuss the term in those contexts. Testing in this context means assessing one or more clearly defined criteria such as school achievement, student motivation, or learning disabilities. In a scientific understanding, tests need to be constructed using specific methods, need to meet specific quality criteria, and need to be conducted under specific rules. Teachers and educators need to evaluate and analyze diagnostic test results in order to use them for their teaching and instructional program. Results of diagnostic tests illustrate the extent to which the measured aspect exists and can be used to optimize teaching and learning in educational contexts.

Diagnostic tests are a major part of and great possibility for assessing relevant aspects of teaching and learning in educational systems. In order to optimize teaching, the assessment of learning-relevant aspects is indispensable. In this regard, and with reference to high-stakes testing, educational measurement, and evaluation, diagnostic tests have great influence on and strongly result in the changes of educational systems and instructional programs.

This entry initially describes functions and forms of diagnostic tests in educational purposes. More precisely, diagnostic subjects as well as different possibilities to assess the relevant data are presented. Subsequently, special characteristics of diagnostic tests are described that enable a classification of different test formats. This includes the description of the classical quality different test formats. This includes the description of the classical quality criteria for tests, the comparison of formal/standardized and informal tests, tests at an individual or group level, norm-referenced and criterion-referenced tests, as well as summative and formative tests. These classification aspects lead to describing possible uses of diagnostic tests in the school context as well as competence that are required by teachers and educators who conduct these diagnostic tests—the so-called diagnostic competence. An overview of three diagnostic tests exemplifies the described functions and classifies the relevant test characteristics.

Functions and Forms of Diagnostic Tests

Diagnostic tests in educational contexts are used in different situations and with different purposes. As a result of the test, each examinee is generally assigned a measurement value that indicates the extent to which the criterion is given. Therefore, a test is usually a relation between empirically tested persons and measurement scores. A diagnostic test in education aims to assess criteria in the field of teaching and learning. This can be learning conditions such as knowledge, competences, motivation, self-regulation, learning disabilities, and learning achievements, or more general, such as intelligence or school qualification.

At an individual level, they may be used to assess students’ learning characteristics such as knowledge, skills and competences, motivation, learning difficulties, or learning and work behavior. Teachers can use the results to enrich their judgments about individual students as well as adapt and optimize their teaching with regard to their students’ weaknesses and strengths. The purpose of these individual-level assessments is to reach results of each individual student, such as learning outcomes or special learning difficulties. Examples for these individual student assessments are class tests that assess the students’ learning achievement at the end of a teaching unit or tests that assess specific learning disabilities such as dyscalculia or dyslexia.

Other tests provide results at a group level. These tests use the individual students’ results not only to describe specific learning outcomes on the individual student’s level but also to summarize statements that can serve monitoring purposes at a program, school, or system level. Here, students’ outcomes are aggregated to compare results on each of the respective levels. Diagnostic tests that are usually used in educational contexts are achievement tests, learning disability tests, intelligence tests, school qualification tests,

tests, learning disability tests, intelligence tests, school qualification tests, cognitive attention tests, or social tests.

Test formats can differ. However, each test aims to assess criterion-relevant data from the test person(s) or object in focus. This data collection can happen in different forms, such as through paper-and-pencil tests, observations, interviews, oral questioning, writing samples, and others.

Quality Criteria of Diagnostic Tests

The accurateness of a test’s result can be judged by specific quality criteria. These criteria—validity, reliability, and objectivity—come from the classical test theory, describe the quality of a measurement instrument (or diagnostic test), and enable a flawless assessment of a construct.

Results of a diagnostic test are *valid* if the test measures exactly what it is supposed to measure. Thus, a diagnostic test that ought to assess reading abilities is valid if it measures the student’s reading ability and not the student’s ability to read graphs or any other skills. The quality criterion *reliability* judges whether a test produces dependable and stable results. The focus exclusively lies on the exactness of measurement. Finally, a test is *objective* if its results are independent of the test administration and different persons reach equal results when administering the test. These numeric results allow for comparing the tested persons.

In this regard, diagnostic tests are distinguished in terms of their consideration of the quality criteria.

Characteristics of Diagnostic Tests and Distinctive Features

Formal diagnostic tests are usually created by expert test developers and closely account for all quality criteria listed in the previous section. As an equivalent term, formal tests are often described as *standardized* tests. Standardized tests are applied to a representative sample under exactly the same terms and conditions and provide a norm as reference for the individual test results. *Informal* tests, in contrast, are usually developed by teachers and often do not consider the quality criteria as intensely. In addition, formal and informal tests can be distinguished by their reference to a standard sample.

These tests are called norm-referenced tests and aim at comparing test results. In this regard, a comparison of results between an individual and other individuals is called a social norm-referenced test, and a comparison between individuals test result and their result at a previous testing time refers to an individual norm. Many diagnostic norm-referenced tests enable a ranking of the individual test result in relation to the standardized value of a representative sample. This comparison can describe how many students of the representative sample scored above or below the assessed student.

These norm-referenced tests can be applied in different educational contexts. Teachers could use the results of their students to verify their grades or to adjust their teaching with regard to the students' heterogeneous learning requirements. In addition, the teacher can check the students' learning results in comparison to the results of a representative student sample. This may result in an adaption of teaching methods to optimize terms and conditions for teaching and learning in class.

However, these tests enable statements about students' rank in comparison to their peers but do not give information about the extent to which the tested student actually reached the content-related learning goal. If none of the students in the sample that represents the norm reached the learning goal, a student may achieve good test results compared to this social norm despite not achieving sufficient results with regard to content. In this regard, criterion-referenced tests give information about the extent to which the assessed criterion was reached. Test items assess whether a teaching aim was reached or not, and the resulting test score of a student then quantifies the learning result referring to the specific criterion. Examples of criterion-referenced tests are final exams or tests to access specific subjects in university settings. These tests often define a minimum score that students have to achieve in order to pass the content requirements. Although norm-referenced tests are usually distinguished from criterion-referenced tests, criterion-referenced tests can also refer to a norm, in this case a factual norm. Concluding, the difference between norm-referenced tests and criterion-referenced tests can be visualized by a student who is taking a numeracy test and correctly solves 65% of the given tasks. As a norm-referenced test, this result provides information about the position of the students' results with regard to their classmates. The result may be that the students test score lies in the best quarter of the class. However, with regard to the assessed criterion, the students' may only pass the test if they correctly solve 80% of the test questions.

may only pass the test if they correctly solve 80% of the test questions. Therefore, the result of the criterion-referenced test is that the student failed the test.

Informal tests are usually not distinguished in norm-or criterion-referenced tests but can refer to norms or criteria as well.

Another aspect for distinguishing diagnostic tests is their use of the findings. In this regard, summative tests are distinguished from formative tests. In summative tests, results are used to describe the learning outcome of the tested person(s) at a specific moment in time. For example, the master’s degree describes the result of a student’s university studies. This result presents the learning outcome and can be used as a qualifying report in the application process. With regard to tests at a group level, results from summative tests are used to describe outcomes of an educational system. International standardized tests that assess a representative number of students in several countries to compare and describe the outcomes of different educational systems are an example of summative tests at a group level.

Results of formative tests, in contrast, are used to adapt the following teaching and learning actions. Thus, formative tests assess certain aspects in the process of teaching and learning such as learning difficulties, students’ misconceptions, or the effectiveness of specific teaching methods to optimize teaching and subsequent learning. Therefore, one of the main characteristics of formative tests is to provide feedback that improves educational practices. An example for a formative test is a class test that assesses the students’ preliminary learning results to a specific learning content. The teachers use the results of the test to gain knowledge about the current learning stage of their pupils and to develop teaching strategies based on the deficits that become obvious through the test results.

However, formative as well as summative tests primarily and equally assess data regarding specific aspects of teaching and learning such as learning outcomes. They only differ in the way these data are used. In summative tests, results are used to show outcomes, whereas results of formative tests are used to adapt and optimize teaching and learning. In this regard, results from international standardized tests that were presented as an example for summative tests can serve formative test purposes. Their results provide feedback for the respective educational system and can be used to adapt and optimize structures. This differentiation, therefore, only refers to the use of the results, not to the data assessment or testing situation itself.

Various Uses of Diagnostic Tests in Schools and the Diagnostic Competence of Teachers

The different diagnostic tests that serve different diagnostic purposes are developed and used by different parties in an educational context. For example, large standardized international studies that compare learning achievements across countries are usually developed by external organizations or research institutions. Diagnostic tests that serve monitoring purposes at a national level are often developed by national departments or ministries, and finally, teachers develop diagnostic tests to assess individual and learning-relevant aspects of their students in class. This test development requires diagnostic competence of the teachers. In addition, teachers need to be able to evaluate, interpret, and implement the different diagnostic information that they obtain through the various diagnostic tests. With regard to standardized tests at an individual level, teachers need the ability to choose diagnostic tests adequately, they must use them appropriately and correctly, and they need to evaluate and use the data in the most efficient way. However, these diagnostic tests at the individual level are usually not applicable during class, and teachers do not have enough time during class to assess specific learning criteria of only one student.

Other diagnostic tests that do not focus on students, their knowledge, abilities, and motivation as a research target might assess the teacher or the teaching quality. More precisely, teachers can develop and use observation instruments as diagnostic tests and assess their colleagues, or students may judge the instruction as well. However, this form of teacher observation and assessment is a common method in teacher education. Here, teacher educators assess their students’ teachers using specific diagnostic observation tests. Again, teachers can use this information to adapt their teaching.

Examples of Diagnostic Tests That Are Used in Educational Contexts

The Test of English as a Foreign Language (TOEFL)

The TOEFL is a standardized diagnostic test that assesses the English language abilities of nonnative speakers of English. The test has either the format of a paper-and-pencil test or can be accessed electronically. With regard to content, the test assesses the examinee’s skills in the areas of reading, listening, speaking, and writing. In this regard, the test result is given as a score, with one score for each of the four content areas as well as a total score that summarizes the results across all areas. However, a minimum test score that indicates whether the tested person’s English language abilities are sufficient or not is not defined. Therefore, the score itself does not allow for interpreting whether the test was passed or failed. But the total score provides information about the extent of the English language abilities. Some institutions such as universities or companies declare certain score limits that enable or deny educational access.

With regard to the classification criteria, the TOEFL is a formal diagnostic test that is criterion referenced and provides results at an individual level. The results from the TOEFL provide summative findings and are not used in school but rather are provided by a nonprofit organization. However, educational institutions such as universities use these test scores as entrance requirements.

The Programme for International Student Assessment (PISA)

The PISA study is a standardized diagnostic test that assesses 15-year-old students in several participating countries. It is developed and conducted by the Organisation for Economic Co-operation and Development and aims to assess the students’ knowledge and skills in mathematics, reading, and science. More precisely, the test measures to what extent students at the end of their compulsory education are able to solve realistic and problem-solving tasks and can, thus, show that they are able to be a reflective and independent part of society. A score is generated for each of the assessed areas and each of the participating countries, and results are presented as rankings.

In this regard, the PISA study is a formal diagnostic test that provides results at a group level—more precisely, at a system level. It is a norm-referenced test and provides summative results. However, the findings of the PISA study caused extensive change and development in some educational systems, and therefore, the summative results initiated formative processes.

The Kaufman Assessment Battery for Children—2

The Kaufman Assessment Battery for Children – 2

The Kaufman Assessment Battery for Children, second edition, is an intelligence test that assesses children and young adults' cognitive and information-processing abilities. The test was developed to test persons aged 3–18 years and enables the test administration to choose between different theoretical models. Several tasks from different subtests require the children who are being assessed to receive and process information. Subscales refer to children's sequential (short-term memory), simultaneous (visual processing), learning (long-term storage and retrieval), and planning (fluid reasoning) skills, whereas verbal ability and specific relevant knowledge are only part of the test when choosing one specific theoretical model. The test's results are reported as score summaries and scale profiles. This intelligence test is a norm-referenced diagnostic test that provides the norm of a large, representative sample. This test can be classified as a formal and summative test at an individual level.

Intelligence tests are used in educational contexts if students attract attention because of their highly above average or below average performances.

Jessica Hoth

See also Achievement Tests; Classical Test Theory; Classroom Assessment; Criterion-Referenced Interpretation; Norm-Referenced Interpretation; Tests

Further Readings

Black, P. J., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5–31.

Cunningham, G. K. (1998). Assessment in the classroom: Constructing and interpreting tests. London/Washington: Falmer Press.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2015). Teacher assessment literacy: A review of international standards and measures. Educational Assessment, Evaluation and Accountability, 28(3), 251–272.

Joint Advisory Committee, Centre for Research in Applied Measurement and

Evaluation, … University of Alberta. (1993). Principles for Fair Student Assessment Practices for Education in Canada. Edmonton, Canada: Joint Advisory Committee. Retrieved from www.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf.

Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman assessment battery for children, second edition: Examiners manual.

Nusche, D. (2008). Assessment of learning outcomes in higher education: A comparative review of selected practices (OECD Education Working Papers, No. 15). OECD Publishing. Retrieved from http://dx.doi.org/10.1787/244257272573

OECD Programme for International Student Assessment. (2009). PISA 2009 assessment framework: Key competencies in reading, mathematics and science. Paris, France: OECD.

Phye, G. D. (Ed.). (1997). Handbook of classroom assessment: Learning, achievement, and adjustment. San Diego, CA: Academic Press.

Anne Corinne Huggins-Manley Anne Corinne Huggins-Manley Huggins-Manley, Anne Corinne

Differential Item Functioning Differential item functioning

516

521

Differential Item Functioning

Differential item functioning (DIF) is formally defined as a lack of equality between two group’s conditional probability functions that relate a trait of measurement to an item’s response data. DIF indicates that examinees who are equal on the trait of measurement, but differ according to some external variable, show differential performance on a test item. For example, DIF is said to be present if groups of examinees who are matched in quantitative aptitude (trait of measurement), but vary in country of origin (external variable), show differential performance on a test item that is intended to measure quantitative aptitude. The presence of DIF is viewed as problematic because it implies that some factor external to the trait being measured is influencing responses to test items in different ways for different groups of examinees, providing an advantage to one or more groups. DIF is a commonly explored phenomenon in educational measurement data because it assists in the statistical process of gauging fairness at the item level, which contributes to the overarching evaluation of measurement validity. Educational research on DIF includes developing methods to estimate DIF; comparing and contrasting the wide variety of DIF methods available in the literature; connecting DIF to other psychometric phenomena; conducting DIF analysis on particular test items and/or across particular groups of examinees; and overcoming challenges to estimating, interpreting, and addressing DIF in practice. The remainder of this entry introduces the concept of DIF in relation to other educational measurement concepts, defines various manifestations of DIF in data, describes some select methods for evaluating DIF, and reviews some of the challenges to evaluating DIF in practice.

The Concept of DIF and Relationships to Other

Educational Measurement Concepts

DIF indicates that two (or more) groups display conditional differences in item responses. Ultimately, these functions can only differ if some secondary factor is playing a role in item responses, and the groups have different distributions on that secondary factor. For example, an item that is intended to measure science ability may also measure language proficiency as a nuisance trait (i.e., an unintended trait of measurement). If two (or more) groups of examinees have different distributions of language proficiency (e.g., one group has a lower mean language proficiency), then DIF is expected to be present across those groups in the item response data. This conceptual understanding of DIF exemplifies the connections between DIF, dimensionality, fairness, and validity. DIF is one of many ways that failure to measure a single trait in the same manner across different groups of examinees manifests itself in test outputs. Hence, the *Standards for Educational and Psychological Testing* refers to DIF analysis as a necessary part of evaluating test fairness.

DIF is encompassed in the broader statistical phenomenon of measurement invariance. Measurement invariance refers to the equality of parameters in a statistical model across various groups, and the parameters of concern could be item-level parameters, structural model parameters, model fit parameters, parameters related to external relationships, and more. DIF is only concerned with item-level parameters and a lack of invariance in such parameters, making it a specific type of violation of measurement invariance.

It is important to separate the definition of DIF from the definition of group mean differences on a trait of measurement. DIF refers to *conditional* group variance in item performance, whereas group mean differences refer to *unconditional* group variance in item performance. These two phenomena can occur together or separately in any combination because they come about for very different reasons. Group mean differences can occur because of true differences in the population of examinees (e.g., females outperforming males on a reading comprehension test item because that group has higher reading comprehension in the population) or nonrepresentative sampling (e.g., even though females and males are equal in mean reading comprehension in the population, the sampling procedure selected more able females than males). DIF, however, refers to item performance differences across groups for examinees who are equal on the trait being measured, and it arises as a result of deviation from unidimensional measurement across or within groups. Regardless of population mean differences between groups or sampling techniques, group comparisons for examinees of equal (or approximately equal) levels on the trait being measured may or may not show DIF in an item.

The concept of DIF is also distinct from the related term of item bias. Item bias refers to a test item that has some underlying mechanism providing an unfair advantage or disadvantage to a group(s) of examinees. Notice that this is a more general term than DIF because it can encompass fairness related to item content, to observed item performance, to external criteria, and more. DIF is a more specific term, as it refers only to a statistical phenomenon manifested in the observed item data. Also, item bias necessarily indicates a lack of fairness, while the presence of DIF is not necessarily considered unfair. If the secondary trait that is the source of DIF can be located in an item and subsequently shown to be a meaningful part of the intended trait of measurement and/or intended test use, then the presence of DIF may not signify a lack of fairness. For example, if an item on a quantitative aptitude test for college admissions displays DIF, it is determined that the cause of DIF is language proficiency differences between groups, *and* it is decided that language proficiency is an important aspect of quantitative aptitude in future college performance, then the presence of DIF may be considered a valid presence of conditional group differences in item performance.

Manifestations of DIF in Data

The definition and concept of DIF are consistent across different types of items, different unidimensional traits of measurement, different groups of examinees, and more. However, DIF can be manifested in multiple ways in a particular item’s data. For this section, assume a binary, ordered item (i.e., 0 is a lower score than 1, and those are the only possible scores) and the case of two groups of examinees. The unequal conditional probability functions between the groups are necessarily associated with the groups having different conditional probabilities of scoring 1 on that item. The group with the higher conditional probability of scoring 1 is considered advantaged by the item, while the other group is considered disadvantaged. But that advantage is not necessarily consistent across different levels of the trait of measurement. Rather, DIF has a particular direction (i.e., which group is advantaged) and magnitude (i.e., how much is a group advantaged), both of which can vary across levels of the trait of measurement. For example, for an item on a test that is intended to measure measurement. For example, for an item on a test that is intended to measure reading comprehension, it may be that the presence of DIF is associated with an advantage for one group at lower levels of reading comprehension but no advantage for any group at moderate and high levels of reading comprehension.

Uniform DIF refers to the case in which the direction and magnitude of advantage is consistent across levels of the trait of measurement. For example, uniform DIF would be present if students from private schools were advantaged on a physics item across all levels of physics ability in the trait distribution, and the magnitude of that advantage was consistent across levels of physics ability. Nonuniform DIF refers to the case in which the direction of advantage is consistent across levels of the trait of measurement, but the magnitude is not. For example, nonuniform DIF would be present if students from private schools were benefited on a physics item across all levels of physics ability, but the advantage was much larger for examinees with higher levels of physics ability. Crossing DIF is a special type of nonuniform DIF in which not only is the magnitude of the advantage inconsistent across levels of the trait but also the direction of the advantage varies across levels of the trait. For example, crossing DIF would be present if students from private schools were advantaged on a physics item at a lower level of the physics ability distribution, but disadvantaged at moderate and high levels of the physics distribution.

There are additional classifications of DIF manifestation, particularly with respect to different testing and examinee situations. For example, when dealing with polytomous items, the manifestation of advantage can take on more complex forms. Also, when evaluating DIF across more than two groups, there are more intricacies related to the nature of the advantage. But even in those cases, there are persistent questions about whether or not the direction and magnitude of the DIF effect are consistent across trait levels because most methods for estimating DIF vary in their performance depending on the type of DIF manifestation in the data.

Methods for Evaluating DIF

There are a plethora of methods that have been proposed for evaluating DIF in item response data. In the remainder of this section, a few commonly used DIF methods are briefly introduced. Readers should bear in mind that this is only a small sample of the many methods in the literature and that details about the utility of these methods are not fully described. For example, some methods are only appropriate for particular types of items, grouping variables, sample sizes, and manifestations of DIF. Some methods provide a direct hypothesis test for DIF, some methods result mainly in effect sizes of DIF, and some provide both. Readers should access additional readings to fully explore the various DIF methods.

Nonparametric Methods

Nonparametric methods are available for evaluating DIF, and they benefit from smaller sample size requirements and fewer assumptions as compared to some parametric DIF methods. Contingency table approaches to nonparametric methods are commonly used for estimating DIF effects. First, examinees are divided into strata based on their trait level so that item performance is only compared within strata (i.e., item performance comparisons are roughly conditioned on trait level). Then, within the strata, the proportions of 0 and 1 item responses are disaggregated to each of the levels of the grouping variable (e.g., females and males for a gender grouping variable). For an item to be free of DIF, one would expect no relationship between the grouping variable and the proportions of 0 and 1 item responses within strata. Restated, one would expect that for examinees of roughly the same trait level, the grouping variable does not relate to item responses.

A specific DIF statistic that can be calculated from the conditional contingency tables is the standardized *p* difference, which involves calculating differences in conditional proportions of examinees scoring 1 within strata and obtaining a weighted average of those differences across strata. There are signed and unsigned versions of this index, which vary in their ability to detect different manifestations of DIF. Other methods utilize odds ratios (i.e., the odds of scoring a 1 as opposed to a 0 within group) rather than proportions within strata and then combine those odds ratios across strata. The Mantel–Haenszel common log-odds ratio is a popular DIF method for binary items based on this type of approach. These types of contingency table procedures have been adapted to the case of polytomous item DIF analysis with, for example, the Liu–Agresti estimator of the cumulative common odds ratio.

Parametric Methods

Some DIF methods take advantage of parametric statistical models, namely, item response theory (IRT) models. While they often require larger sample sizes (including the concern of within-group sample sizes), these methods can benefit from the estimation of the underlying latent trait that can be used for conditioning response probabilities. The general approach is to fit an IRT model to examinee groups separately and then compare the item parameters obtained across groups. For example, area measures for binary items estimate the signed or unsigned area between group-level functions based on item parameters. The differential functioning of items and tests approach extends these area methods to incorporate, among other things, a weighted difference in the group-level functions across the trait levels. Alternatively, differences in group-level item parameters can be evaluated with a likelihood ratio test. This method compares a model in which an item is free to have different parameters for groups to a model in which the groups are restricted to the same set of item parameters. These IRT-based methods have been extended to polytomous items, albeit with some complications in estimation and application.

Logistic regression methods can be used for DIF analysis of both binary and polytomous items, with the latter case introducing some complexity. These are considered parametric methods in that the DIF effect is estimated through slope parameters. However, regression models generally do not require as large of a sample size as the IRT-based methods previously discussed. The general approach is to treat the item data as an outcome variable in a logistic regression, with predictors that include the trait of measurement (often with total test scores as proxies), the grouping variable, and an interaction between the two. The slope estimates for the grouping and interaction variables indicate the presence or absence of DIF as well as provide information on the particular manifestation of DIF in the item data (e.g., nonuniform vs. uniform DIF).

DIF methods are continuously being developed as psychometric theories and models are advanced in the literature. For example, DIF methods are available for multidimensional IRT models, diagnostic classification models, testlet and bifactor models, and more. While these DIF methods have been adapted to the particular measurement models of interest, their core focus is consistent with the definition and concept of DIF discussed earlier.

Challenges to Evaluating DIF in Practice

Conducting a DIF analysis is rarely a straightforward endeavor of applying a DIF method and interpreting the statistical results. Rather, the process of DIF evaluation is complex and iterative, and it requires not only statistical expertise but also critical thinking skills and an ability to synthesize the results into an interpretable outcome and solution that is appropriate for a particular measurement context and test use. This complexity is, in part, due to the fact that DIF results need to be connected to issues of test fairness, bias, and validity, each of which are multifaceted concepts surrounded by much uncertainty in practice. The complexity is also due to data analytic challenges, some of which are discussed in the remainder of this section.

An initial challenge in any DIF analysis is to select the groups of examinees to evaluate. The general recommendation is to evaluate groups for which there is some substantive interest related to test fairness and validity. Demographic groups (e.g., gender, ethnicity) are often examined because historical contexts related to fairness and equity beyond just the scope of testing often revolve around such groups. A more specific recommendation is to focus on groups for whom the test use is particularly important. For example, if a test is to be used to identify students in need of reading remediation, it may be particularly important that the test item data display similar conditional performance across students with and without reading disabilities. A vastly different approach to selecting groups for DIF analysis is to estimate latent groups from the data that display DIF (i.e., locate the groups that show DIF rather than specify them *a priori*), but challenges are associated with interpreting the nature of those groups.

Another ubiquitous challenge in all DIF evaluations is matching the examinees on the underlying trait of measurement. Most often, a test is administered because examinees have unknown locations on the trait of measurement (e.g., one administers a math test because one does not yet know the math ability of the examinees). DIF evaluation requires the data analyst to condition examinees on these unknown locations. For nonparametric methods and logistic regression methods, total test scores are often used for this matching. However, there is a core concern here; if items have DIF, then total test scores are systematically biased estimates of true scores. Purification processes are meant to alleviate these concerns in that the DIF analysis becomes an iterative process of testing for DIF, removing items that have DIF from the trait proxy (e.g., the total test score), retesting for DIF, and so on, until a DIF conclusion is obtained. The simultaneous item bias test method for DIF detection incorporates procedures for purification in the method itself. Many other DIF methods simply estimate DIF effects while leaving the purification process to the analyst. IRT-based DIF methods use the latent trait as a proxy rather than the total test score. This can have some benefits over using total test scores, but the solution is by no means a have some benefits over using total test scores, but the solution is by no means a panacea to the problem of testing for DIF with a proxy trait estimate that may have come from item response data that contains DIF.

Once a DIF effect has been estimated for particular groups of matched examinees, interpreting DIF results is a matter of good judgment. Many DIF methods are associated with hypothesis tests that may suffer from larger than desired Type I or Type II error rates in particular conditions, and/or they result in effect size estimates that are difficult to interpret. Also, it is commonplace for multiple (if not many) items on a single test to display DIF, for the advantage of the DIF to vary in direction and magnitude across items, and for the DIF effect to have no obvious source in item content. Roughly speaking, the first challenge is to connect the statistical DIF results to the item content for the purposes of identifying the source of DIF. The next challenge is to explore the extent to which such DIF impacts test score interpretations and/or uses to determine whether it has some meaningful, negative impact on issues related to fairness and validity. If a DIF situation is considered unacceptable for a particular item and test, a following challenge is to make decisions for how to remedy it. For example, practitioners may consider item removal or revision, lowering the stakes of the test use, removing some groups of examinees from the population for which the test should be used, or altering the nature of test score interpretation.

Anne Corinne Huggins-Manley

See also Ethical Issues in Testing; Item Analysis; Mantel–Haenszel Test; Measurement Invariance; Psychometrics; *Standards for Educational and Psychological Testing*; Test Bias; Validity

Further Readings

Camilli, G. (2006). *Test fairness*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 221–256). Westport, CT: American Council on Education and Praeger.

Camilli, G., & Shepard, L. A. (1993). Methods for identifying biased test items (Vol. 4). Thousand Oaks, CA: Sage.

Holland, P. W., & Wainer, H. (Eds.). (1993). Differential item functioning. New York, NY: Routledge.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing bias. Applied Psychological Measurement, 17, 297–334.

Osterlind, S. J., & Everson, H. T. (2009). Differential item functioning (2nd ed.). Thousand Oaks, CA: Sage.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics: Psychometrics (Vol. 26, pp. 125–167). Amsterdam, The Netherlands: Elsevier.

Howard T. Everson Howard T. Everson Everson, Howard T.

Difficulty Index

Difficulty index

521

523

Difficulty Index

When designing and developing educational and psychological tests, questionnaires, and assessments, measurement specialists attend directly to the qualities of the items making up the test or questionnaire. For example, it would be important to know whether an item is too easy or too difficult for the intended audience and uses of the test. Thus, estimating accurately an item’s *difficulty index* is important for good measurement and high-quality test design.

Estimating Item Difficulty

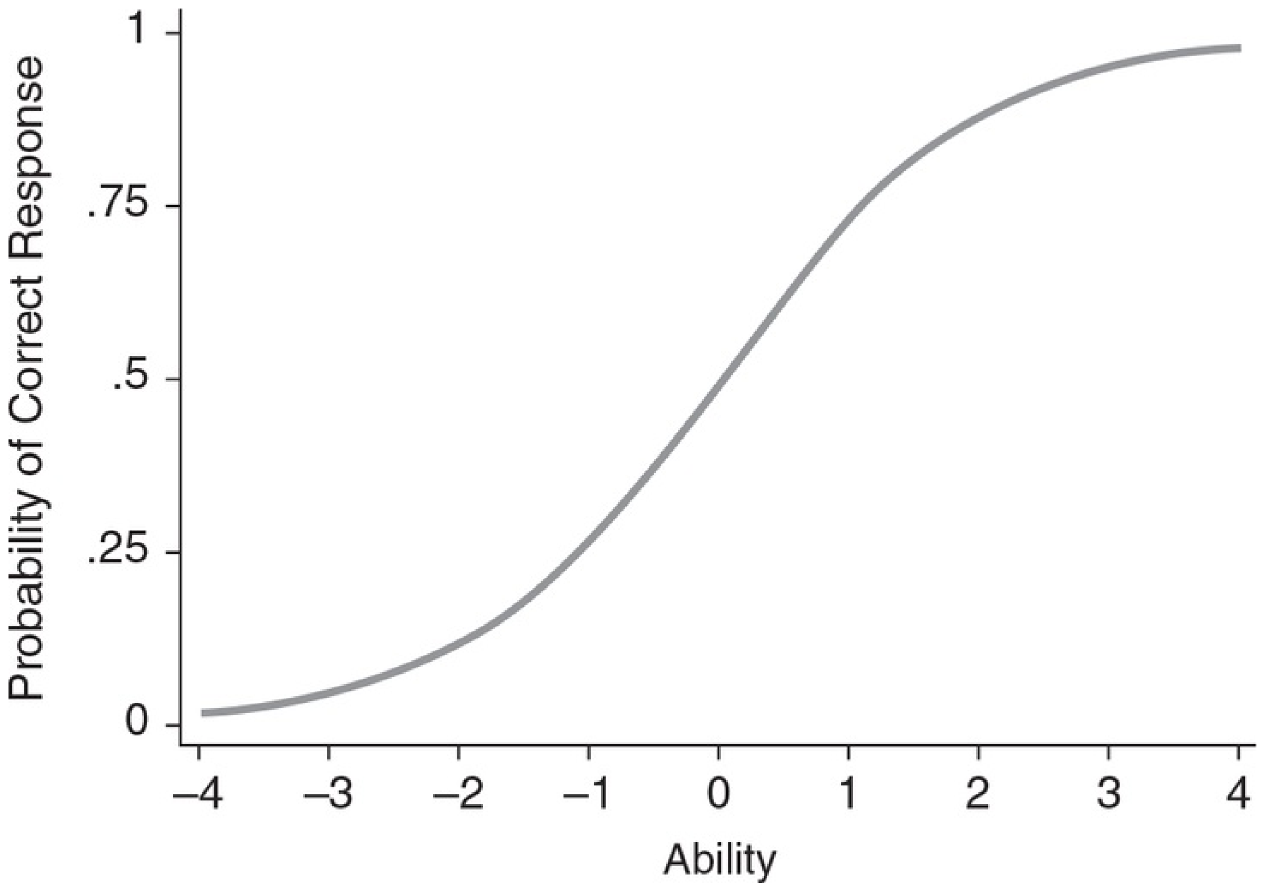
Contemporary test development practices rely on two broad statistical approaches for estimating item difficulty—classical test theory (CTT) and item response theory (IRT). The CTT approach draws on traditional statistical methods for estimating item difficulty. In the CTT framework, the proportion of examinees answering an item correctly or endorsing a particular response option on a questionnaire serves as the difficulty index. This is referred to as an item’s *p* value, and it ranges between 0.0 and 1.0, with higher values indicating a greater proportion of examinees responding correctly to (or endorsing) the item. Depending on the design, say a criterion-referenced test versus a norm-referenced test, measurement specialists may create a test using items having a range of *p* values—seeking an appropriate mix of easy, moderately difficult, or very difficult test items. Thus, an item’s *p* value is one of the most useful, and most frequently reported, item statistics.

Item *p* values, however, are highly sample dependent. The underlying or latent ability levels of the sample examinees interact with estimates of the difficulty of the test items. An item may appear to be much harder (or easier) in one sample of examinees than in another. As a consequence, CTT methods often lead to perplexing and unintended shifts in item difficulty estimates from sample to sample. Estimates of item difficulty that are independent of the ability levels of the sample examinees would be more helpful.

To address this problem, psychometric specialists working largely during the latter half of the 20th century developed a series of statistical methods referred to collectively as IRT. The IRT framework rests on the idea that measurement specialists are interested largely in measuring cognitive abilities, personality traits, and other psychological characteristics that are not directly observable or latent. From this perspective, a test is simply a collection of items designed to measure a person’s level or standing on the latent trait. Thus, when designing the test, the developer is interested in how each individual item relates to the latent trait and how the group of items relates to that trait or ability. IRT models make the study of these relationships more tractable.

The IRT framework assumes the relationship between item performance and the latent ability can be modeled by a one-, two-, or three-parameter logistic function. For simplicity, the focus here is on the one-parameter (the difficulty index) model. Typically, two assumptions underpin an IRT model—the first assumes a unidimensional structure of the test data (measuring one primary construct or latent ability) and the other relates to the mathematical (logistic) form of the item characteristic function or curve (denoted as the ICC). Figure 1 shows the general form of item characteristic functions for the one-parameter logistic model.

Figure 1 One-parameter item characteristic curve

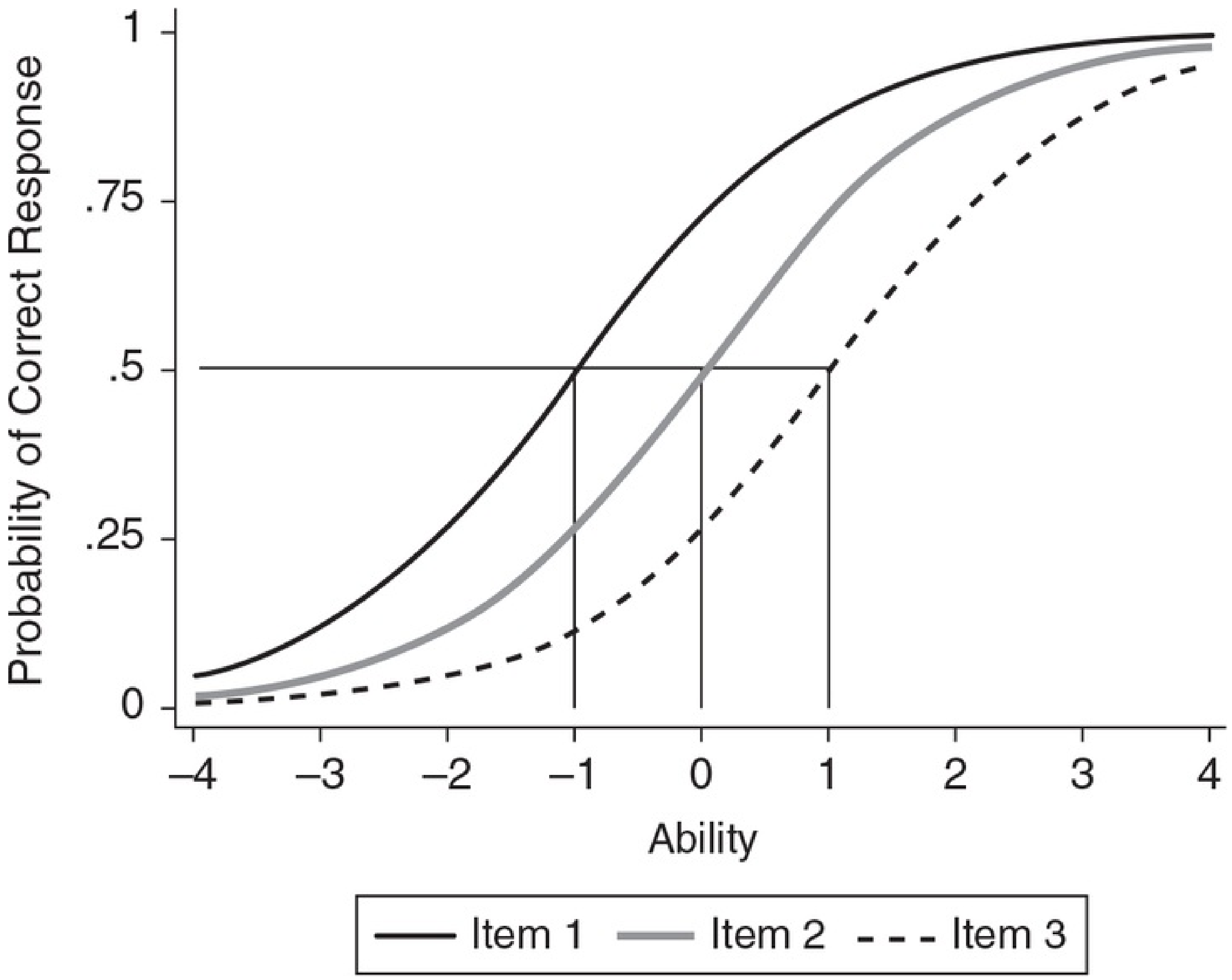


The item characteristic function (the difficulty index) is generated from the expression in Equation 1.

$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}.$$

In this model P_{ij} (θ_j, b_i) gives the probability of a correct response to item *i* as a function of ability (denoted by θ). The *b* parameter, the difficulty index, is the point on the ability scale (θ) where an examinee has a .5 probability of a correct answer. By varying the items' *b* parameters, many *S*-shaped curves or ICCs can be generated to fit actual test data. A typical set of ICCs for 3 items with varying difficulty indices is shown in Figure 2.

Figure 2 Item characteristic curves for three items of varying difficulty



For example, Figure 2 plots the ICCs for Items q1, q2, and q3, with difficulty parameters −1, 0, and 1, respectively. Item q1 is the least difficult, and Item q3 is the most difficult. Notice the change in item difficulty (the *b* parameter) shifts the ICC along the ability (θ) scale. The probability of success on Item q1 is higher than the probability of success for the other two items at any ability level. We can say Item q1 is less difficult than the others because a person would need only an ability level greater than or equal to −1 on this ability scale to be expected to succeed on Item q1. On the other hand, a person would need an ability level above 0 to be expected to succeed on item q2 and an ability level above 1 to be expected to succeed on item q3. In designing an instrument intended to differentiate between all levels of a latent trait, a researcher should try to have items with difficulty indices (*b* parameters) spread across the full range of the latent trait.

Summary

The CTT framework expresses the difficulty index as a *p* value—the proportion of examinees answering the item correctly. However, the *p* value is sample dependent. IRT was developed to address the sample dependency problem, and it offers a statistical method for modeling the relationship between an item characteristic and the examinee's ability by using a one-, two-, or three-parameter logistic function. In a one-parameter IRT model, the *b* parameter represents the item difficulty index.

Howard T. Everson

See also Criterion-Referenced Interpretation; High-Stakes Tests; Item Analysis; Norm-Referenced Interpretation

Further Readings

Haladyna, T. M., & Rodriquez, M. C. (2013). Developing and validating test items. New York, NY: Routledge.

Mellenbergh, G. J. (2011). A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests. The Hague, the Netherlands: Eleven International.

Osterlind, S. J. (2010). Modern measurement: Theory, principles, and applications of mental appraisal (2nd ed.). Boston, MA: Pearson.

Satoko Siegel Satoko Siegel Siegel, Satoko

Discourse Analysis

Discourse analysis

523

525

Discourse Analysis

Discourse analysis is a broad term for the study of language usage. Discourse analysis has been utilized in the humanities and social sciences, including education, linguistics, sociology, anthropology, cognitive psychology, social psychology, communication, and artificial intelligence. As such, discourse analysis includes a range of topics, such as linguistics styles and rhetoric, speakers' and hearers' cognition, and language in social contexts. Given that discourse analysis involves both theoretical and methodological elements, each part of the discipline has its own definition of discourse and discourse analysis as well as its own assumptions and methodologies. Data for discourse analysis are also widely ranged; they can be informal or formal conversations, in private or institutional contexts, and in spoken or written versions.

Through discourse analysis, education researchers try to make sense of the ways in which people make meaning in educational contexts. This entry focuses on describing four approaches of discourse analysis that have been largely utilized in educational research: (1) ethnography of communication, (2) interactional sociolinguistic, (3) conversation analysis, and (4) critical discourse analysis. Each approach is explained with its theory, characteristics, key concepts, and methodology.

Ethnography of Communication

Ethnography of communication, formerly called the ethnography of speaking, was developed by Dell Hymes. This approach was derived from the disciplines of anthropology and linguistics. Ethnography of communication perceives language usage as more than grammatical knowledge; it pays attention to “way of speaking,” which is culturally specific interaction. Such “way of speaking” is shared in the speech community, which is a group of people who share norms and expectations in their language usage. Ethnography of communication researchers conduct ethnographic fieldwork. Specifically, by talking to community members, observing events, and being involved in activities, researchers aim to investigate communicative patterns of the speech community to gain an understanding of how a group of people make sense of their interactions and experiences.

In the theory of the ethnography of communication, the notion of the communicative competence is a key. Communicative competence refers to grammatical knowledge as well as cultural and social knowledge on how and when to use language with grammar in an appropriate way. In relation to communicative competence, Hymes developed a framework, called SPEAKING, for an analysis of a speech event. SPEAKING is an acronym; each letter is an abbreviation for a different element of communication: S (setting and scene, including the time and place), P (participant, including identity and personal characteristics), E (ends, including the purpose of the event), A (act and sequence), K (key or the tone of the language), I (instrumentalities or the linguistic code, such as dialect), N (norms of intention and interpretation), and G (genre of the event). Utilizing the SPEAKING model, ethnography of communication seeks to discover a holistic explanation of a cultural group’s communicative competence in speech situations, events, and acts.

Interactional Sociolinguistics

Having its roots in the ethnography of communication, interactional sociolinguistics comes from a variety of academic disciplines, including anthropology, linguistics, and sociology. Linguistic anthropologist John Gumperz and sociologist Erving Goffman are the founders of this approach. Focusing on culture, society, and language, interactional sociolinguistics seeks a better understanding of how people signal and interpret meanings in face-to-face social interactions.

Contextualization cues, developed by Gumperz, are an important concept to interactional sociolinguistics. Contextualization cues are signaling mechanisms in which people imply how they mean on what they say in their speech. There are four levels of cues: (1) code, dialect, or style; (2) prosodic features; (3)

lexical and syntactic options, formulaic expressions; and (4) conversational openings, closing, and sequencing strategies. Contextualization cues are crucial for conversational inference, which is the situated process of interpretation. Speakers’ cultural backgrounds heavily influence the usage and understandings of contextualization cues, so that cues are most likely to be culturally and socially specific. In such cases, these cues are subtle and used subconsciously.

Interactional sociolinguistics scholars are interested in interaction sequences in naturally occurring conversations; they take a microanalytic approach to analyze discourse, while also paying attention to sociocultural contexts that influence the interactions. The analysis of interactional sociolinguistics provides an account of speakers’ intentions, interpretations, and assumptions that speakers bring into interactions, which contain a variety of contextualization cues to convey their intended meanings. Code-switching, intercultural communication, cross-cultural miscommunication, identity construction, gender differences, and politeness are a few research areas that utilize the interactional sociolinguistics approach.

Conversation Analysis

Influenced by Harold Garfinkel’s ethnomethodology, conversation analysis was developed by three people: Emanuel Schegloff, Harvey Sacks, and Gail Jefferson. Conversation analysis is the approach to the study of human social interaction across disciplines of sociology, linguistics, and communication. Based on the belief that a conversation is managed through interactions between speakers, conversation analysis aims at uncovering how speakers make systematic solutions for structural problems in conversation. In other words, this approach is interested in revealing the sequential feature of talk regarding how a speaker and hearer construct a conversation. The core research concerns of conversation analysis include turn-taking, topic management, information receipt, and opening and closing talk.

Unlike ethnography of communication and interactional sociolinguistics, conversation analysis does not require researchers to be in the field to collect data; neither participant observation nor ethnographic interviews are conducted. For this approach, naturally occurring conversations, which are video or audio recorded, are utilized for data with or without researchers’ involvement in the conversation. Concomitantly, the recordings are transcribed in detail to analyze the ways in which speakers construct the interaction and constitute the context. Conversation analysis utilizes an inductive data-driven analysis: researchers identify recurring patterns in the conversation to analyze to develop a rule or model to explain the occurrence of the pattern. The transcriptions are also utilized as a part of the conversation analysis report to allow people to check the analysis presented.

Critical Discourse Analysis

Critical discourse analysis is distinguished from other discourse analysis approaches by its critical focus and its approach to discourses in social and political contexts. Linguistics scholars Norman Fairclough, Ruth Wodak, and Teun van Dijk are the contributors to the establishment of this approach. By bringing social theory and discourse analysis together, critical discourse analysis defines language usage as a social practice. Influenced by the work of Karl Marx, Stuart Hall, Jürgen Habermas, and Michel Foucault, critical discourse analysis aims to reveal how social power relations are constructed, negotiated, maintained, and reinforced through language usage. In other words, scholars who conduct critical discourse analysis are interested in describing and interpreting discourse structures in the social and political contexts. As such, notions of power, ideology, reproduction, social orders, and institutions are focused in the critical discourse studies through the analysis of language in use.

In the critical discourse analysis, there is no single form of approach; critical discourse analysts utilize a variety of approaches and techniques, such as discourse-historical methods, systemic functional linguistics, and a sociocognitive approach. The technique that a particular study employs is up to researchers’ theoretical backgrounds and the focus of the study. There is no accepted standard of data collection and analysis in this approach.

One of the common features of studies that utilize critical discourse analysis is that they are problem oriented. Unlike other sociolinguistic approaches, critical discourse analysis takes a deductive approach—the researcher has interests and questions in advance. Another distinguished feature of critical discourse analysis from other discourse analysis methods is that the researcher becomes an advocate for the people who experience social and political inequality. Critical discourse analysis is committed to political intervention and social change by raising people’s awareness about specific issues. In other words, critical discourse analysis aims to investigate how discourse works in the social world as well as interject in institutional, social, or political controversies.

In educational settings, critical discourse analysis focuses on the ways in which social relations, identity, knowledge, and power are negotiated through written and spoken texts in communities, schools, and classrooms. Critical discourse analysis has been utilized in many different education areas, such as education policy studies, community education, and art education. Schools and classrooms are critical sites to investigate from the microlevel of classroom talk to the macrolevels of reproduced social structures.

Satoko Siegel

See also Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Cazden, C. B., John, V. P., & Hymes, D. H. (Eds.). (1972). Functions of language in the classroom. New York, NY: Teachers College Press.

Fairclough, N. (1992). Discourse and social change. Cambridge, UK: Polity.

Gumperz, J. J. (1982). Discourse strategies, studies in interactional sociolinguistics 1. Cambridge, UK: Cambridge University Press.

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), Directions in sociolinguistics: The ethnography of communication (pp. 35–71). New York, NY: Holt, Rinehart and Winston.

Sciffrin, D., Tannen, D., & Hamilton, H. E. (Eds.). (2001). The handbook of discourse analysis. Malden, MA: Blackwell.

David W. Stockburger David W. Stockburger Stockburger, David W.

Discriminant Function Analysis

Discriminant function analysis

525

532

Discriminant Function Analysis

Discriminant function analysis is used to predict group membership based on a linear combination of interval predictor variables. The procedure begins with a set of observations, whereby both group membership and the values of the predictor variables are known, with the end result being a linear combination of the interval variables that allows prediction of group membership. The way in which the interval variables combine allows a greater understanding and simplification of a multivariate data set. Discriminant analysis, based on matrix theory, is an established technology that has the advantage of a clearly defined decision-making process. Machine learning techniques such as neural networks may be used alternatively for predicting group membership from similar data, often with more accurate predictions, as long as the statistician is willing to accept decision-making without much insight into the process.

For example, a researcher might have a large data set of information from a high school about its former students. Each student belongs to a single group: (a) did not graduate from high school, (b) graduated from high school or obtained a General Educational Development, and (c) attended. The researcher wishes to predict student outcome group using interval predictor variables such as grade point average, attendance, degree of participation in various extracurricular activities (e.g., band, athletics), weekly amount of screen time, and parental educational level. Given this complex multivariate data set and the discriminant function analysis procedure, the researcher can find a subset of variables that in a linear combination allows prediction of group membership. As a bonus, the relative importance of each variable in this subset is part of the output. Often researchers are satisfied with this understanding of the data set and stop at this point.

Discriminant function analysis is a sibling to multivariate analysis of variance as both share the same canonical analysis parent. Where multivariate analysis of variance received the classical hypothesis testing gene, discriminant function analysis often contains the Bayesian probability gene, but in many other respects, they are almost identical.

This entry explains the procedure by breaking it down into its component parts and then assembling them into a whole. The two main component parts in discriminant function analysis are implicit in the title: discriminating between groups and functional analysis. Because knowledge of how to discriminate between groups is necessary for an understanding of the later functional analysis, it is presented first.

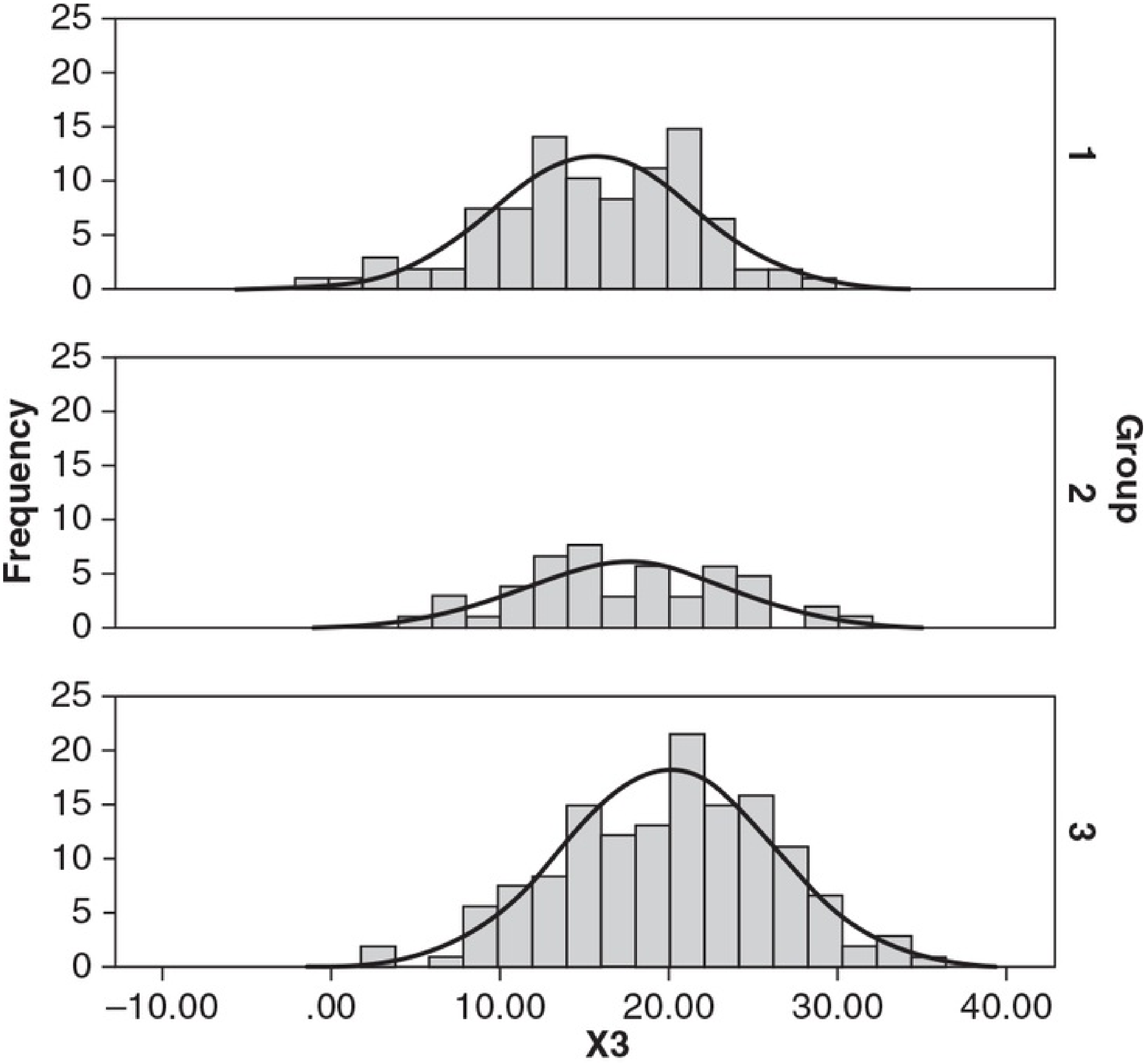
Discriminating Between Groups

Discriminating Between Groups With a Single Variable

The simplest case of discriminant function analysis is the prediction of group membership based on a single variable. An example might be the prediction of successful completion of high school based on the attendance record alone. For the rest of this section, the example uses three simulated groups with *N*s equal to 100, 50, and 150, respectively.

In the example (Figure 1), histograms are drawn separately for each of the three groups. Second, overlapping normal curve models are shown where the normal curve parameters μ and σ are estimated by the mean and standard deviation of the three groups. An analysis of variance shows that the three means are statistically different from each other, but only limited discrimination between groups is possible.

Figure 1 Modeling group membership



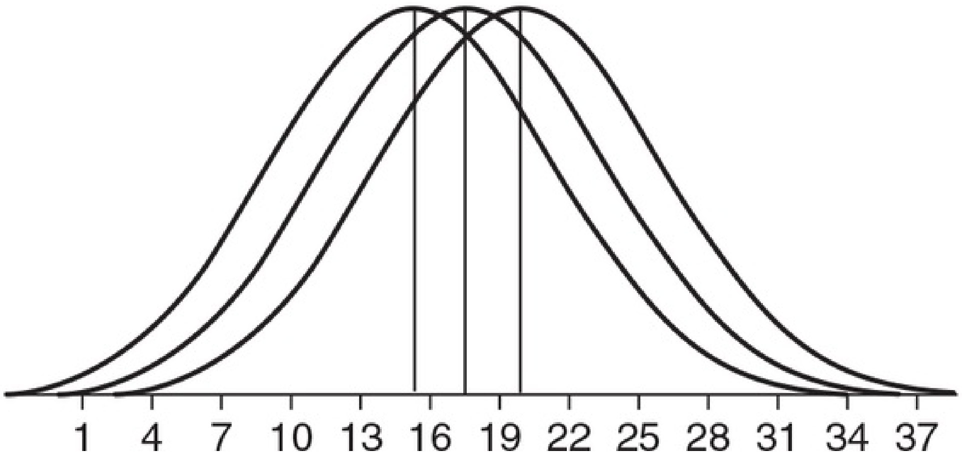
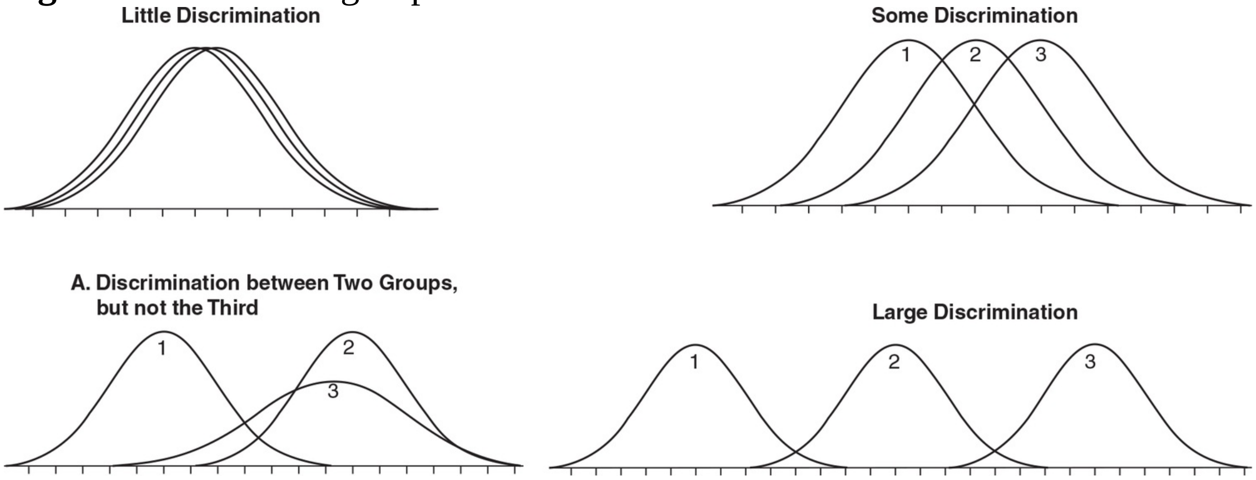


Figure 2 shows various possibilities for overlapping group probability models, from little or no discrimination to almost perfect discrimination between groups. Note that the greater the difference between means relative to the within-group variability, the better the discrimination between groups.

Figure 2 Varieties of group discrimination

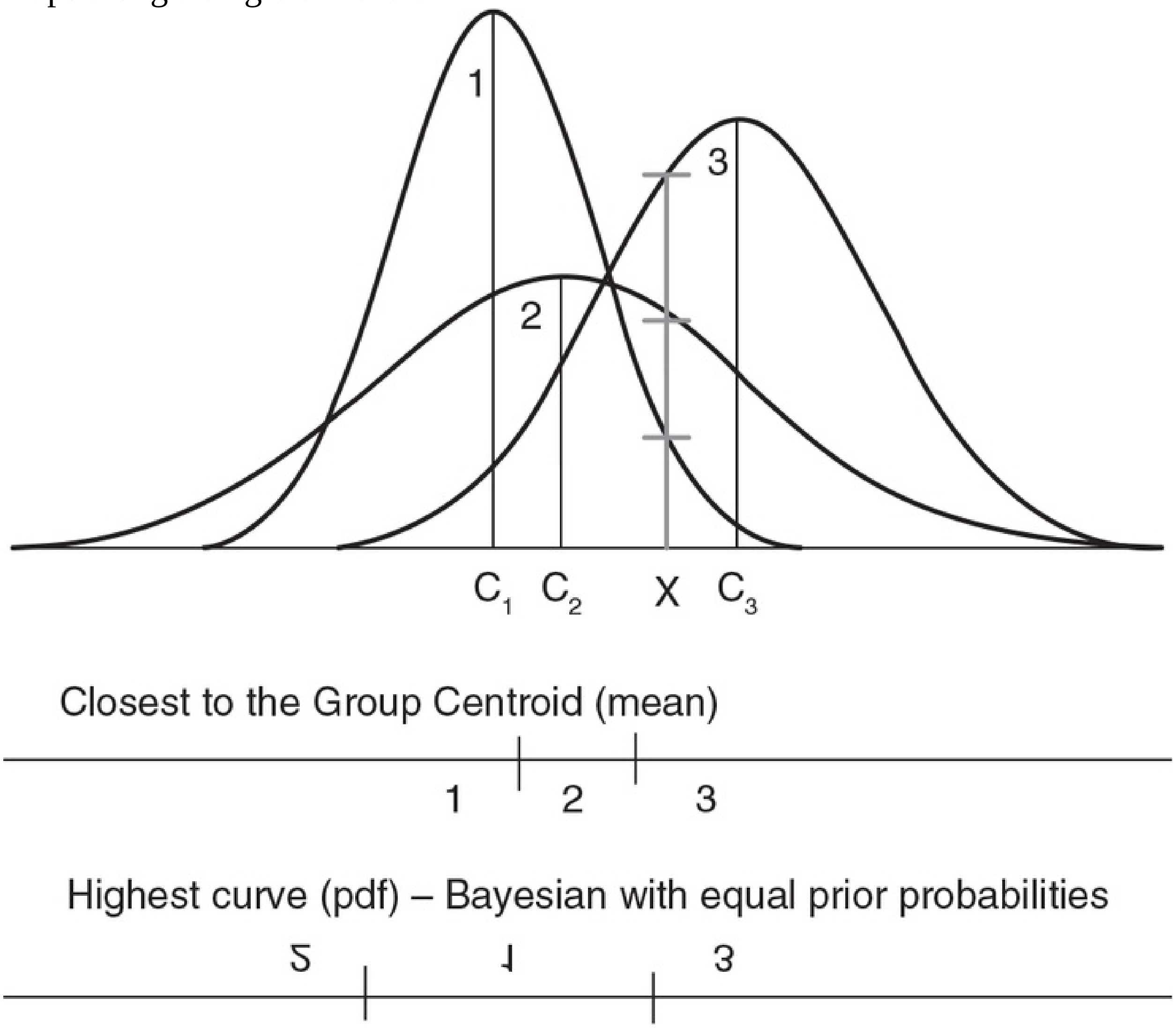


Given that means and standard deviations can be calculated for each group, different classification schemes can be devised to classify scores based on a single variable. One possibility is to simply measure the distance of a particular score from each of the group means and select the group that has the smallest distance. (In discriminant function analysis, group means are called *centroids*.) The advantage of this system is that no distributional assumptions are necessary.

Although not absolutely necessary to perform a discriminant function analysis, Bayes' theorem offers a distinct improvement over distance measures. Bayes' theorem modifies existing probabilities, called *prior probabilities*, into posterior probabilities using evidence based on the collected data. In the case of discriminant function analysis, prior probabilities are the likelihood of belonging to a particular group before the interval variables are known and are generally considered to be subjective probability estimates. Prior probabilities are symbolized as $P(G)$. For example, $P(G_1)$ is the Dyslexia prior probability of belonging to Group 1. In discriminant function analysis software programs (e.g., SPSS), the default option is to set all prior probabilities as equally likely. For example, if there were three groups, each of the three prior probabilities would be set to .33333.… Optionally, the prior probabilities can be set to the relative frequency of each group. In the example data with *N*s of 100, 50, and 150, the prior probabilities would be set to .333…, .16666…, and .75, respectively. Since prior probabilities are subjective, it would also be possible to set them based on cost of misclassification. For example, if misclassification as Group 1 membership is costly, the prior probability might be set to .10 rather than .333.

The probability models of the predictor variables for each group can be used to provide the conditional probability estimates of a score (D) given membership in a particular group, $P(D|G)$. Using the PDF of the probability model, the height of the curve at the data point can be used as an estimate of this probability. Figure 3 illustrates this concept at the data point x, where $P(D = \text{x}|G_1) < P(D = \text{x}|G_2) < P(D = \text{x}|G_3)$.

Figure 3 Classification based on probability models with different territorial maps along a single dimension



Bayes’ theorem provides a means to transform prior probabilities into posterior probabilities given the conditional probabilities $P(D|G)$. Posterior probabilities are the probability of belonging to a group given the prior and conditional probabilities. In the case of discriminant function analysis, prior probabilities $P(G)$ are transformed into posterior probabilities of group membership given a particular score $P(G|D)$. The formula for computing $P(G|D)$ using Bayes' theorem is as follows:

$$P\left(G_j \mid D\right)=\frac{P\left(D \mid G_j\right)P\left(G_j\right)}{\sum_i^{\text{Groups}} P\left(D \mid G_i\right)P\left(G_i\right)}.$$

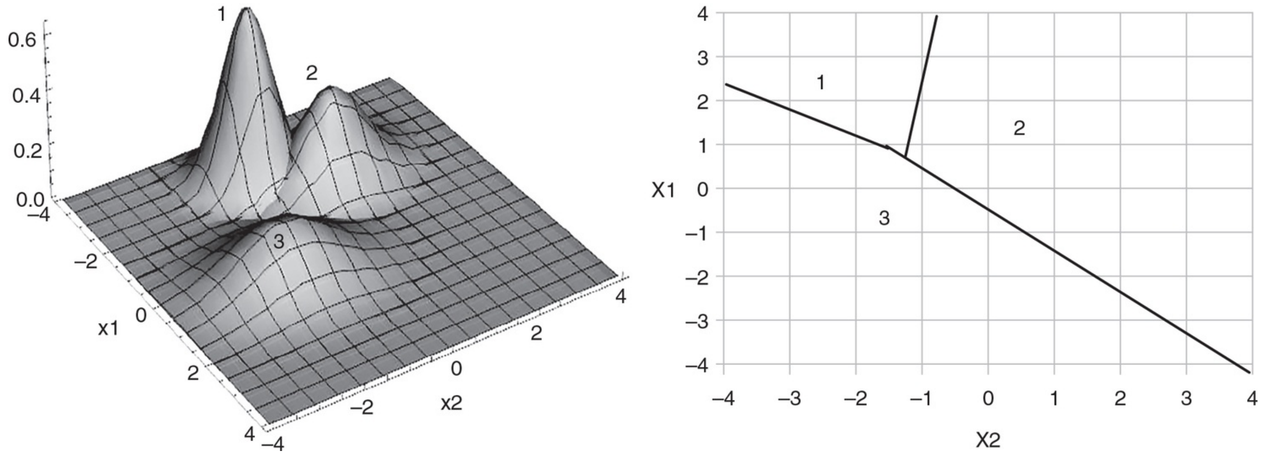
The Bayesian classification system works by computing the posterior probability at a given data point for each group and then selecting the group with the largest posterior probability.

If equal prior probabilities are used, then $P(G_{\mathrm{i}})$ is constant for all groups and can be canceled from the formula. Since the denominator is the same for all groups, the classification system will select the group with the largest $P(D|G)$. In the case of the normal curve examples of conditional distributions presented in Figure 3, at any given point on the *x* axis, the selected group would correspond to the group with the highest curve. This is reflected on the last territorial map on the figure. Note how different it is from the classification system based on distances from each mean. If unequal prior probabilities are used, then the posterior probabilities are weighted by the prior probabilities and the territorial maps will necessarily change.

Discriminating Between Groups With Multiple Variables

In some cases, especially with multiple groups and complex multivariate data, discrimination between groups along a single dimension is not feasible, and multiple dimensions must be used to ensure reasonably correct classification results. A visual representation of a fairly simple situation with two dimensions and three groups is presented in Figure 4. Note that better classification results can be obtained using two dimensions than any single dimension.

Figure 4 Bivariate normal distribution with territorial map



Conceptually, the classification methods are fairly straightforward extensions of the classification systems along a single dimension, although visual representations become much more problematic, especially in three or more dimensions.

Various methods of computing distances from the group centroids can be used, and the minimum distance can be used as a classification system. The advantage of using distance measures is that no distributional assumptions are necessary.

When using a Bayesian classification system, distributional assumptions are necessary. One common distributional assumption is a multivariate normal distribution. The requirements for a multivariate normal distribution are much more stringent and complex than for a univariate normal distribution and therefore harder to meet. For example, both X1 and X2 could be normally distributed, but the combination might not be a bivariate normal distribution. The multivariate normal assumption becomes even more problematic with many more variables. If the distributional assumptions are acceptable, then the Bayesian classification system proceeds in a manner like discriminating between groups with a single variable. The advantage to using a Bayesian classification system is that posterior probabilities of belonging to each group are available.

Linear Functions

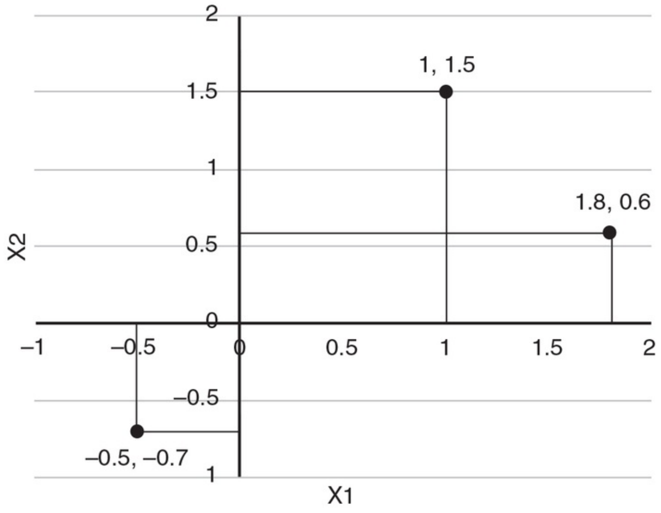
It is only when there are two or more predictor variables that the power of discriminant function analysis becomes apparent. Basically, the procedure discovers linear combinations of the predictor variables that best discriminate between the groups by using matrix operations that are available in canonical analysis. The matrix procedure discovers the linear combination of variables that minimizes the within-group variability and, in the process, maximizes the between-group variability. While a matrix presentation can be beautiful in its apparent simplicity, as some of the additional resources show, what is really occurring beneath the surface can be difficult to fathom if one is not familiar with matrix operations. Thus, this presentation visually focuses on the underlying concepts rather than a mathematically precise formulation.

Changing Structure Using Linear Functions

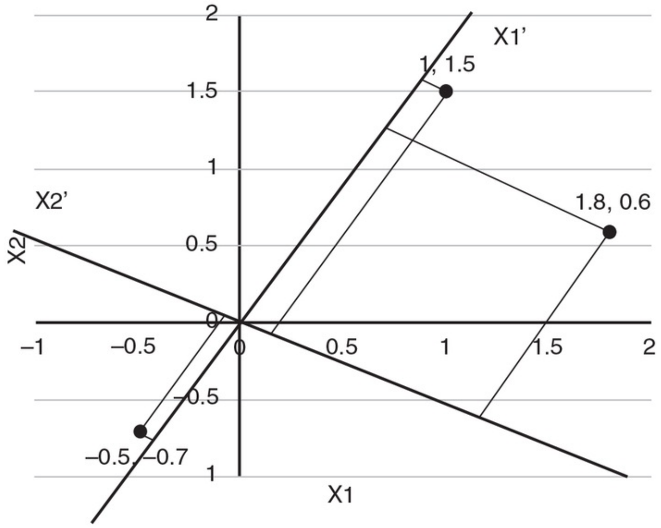
The effect of linear transformations can be observed in Figure 5. Three points, (X1, X2) = (1, 1.5), (1.8, 0.6), and (−0.5, −0.7), are first displayed on their original axis. Note that the population variability of X1 (1.36) and X2 (1.22) as projected onto their respective axes is approximately equal. The sum of the two variances is 2.58.

Figure 5 Dimension reduction—two dimensions to one

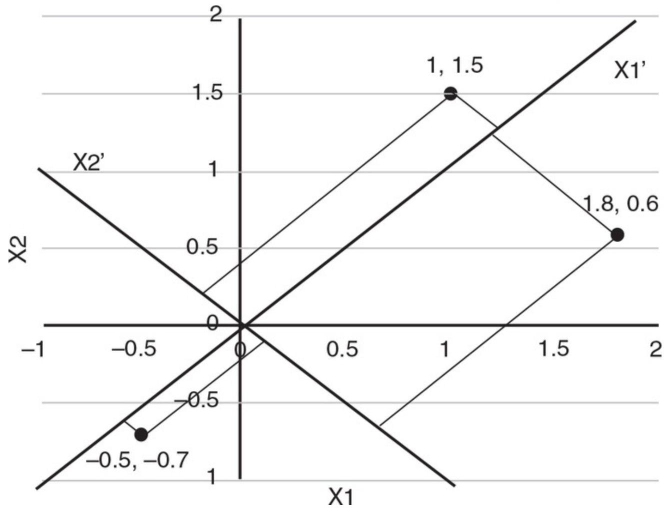
Original Data on X_1 and X_2



Rotated Data $a_1 = 0.5$, $a_2 = .886$, $b_1 = –0.5$. $b_2 = .886$



Rotated Data $a_1 = 0.707$, $a_2 = 0.707$, $b_1 = –0.707$ $b_2 = 0.707$



These points can be transformed by the following formulas:

$$a_1 \mathrm{X1} + a_2 \mathrm{X2}$$

$$b_1 \mathrm{X1} + b_2 \mathrm{X2}$$

$$a_1^2 + a_2^2 = 1, \;\; b_1^2 + b_2^2 = 1,$$

where

$$a_1 = 0.5, a_2 = .866, b_1 = 0.5, b_2 = -.866$$

and projected on the new axes X1′ and X2′ in Figure 5. The three points now become (1.80, 0.12), (−0.86, −0.08), and (1.42, 1.26) using the two transformations. The population variance for X1′ and X2′ is 2.06 and .52, respectively, along the new axes. Note that the sum of the two variances is 2.58, the same as for the original axes. Thus, by using linear transformations with constraints, the variance can be partitioned differently along different axes.

A second linear transformation using:

$$a_1 = 0.707, a_2 = .707, b_1 = 0.707, b_2 = -.707$$

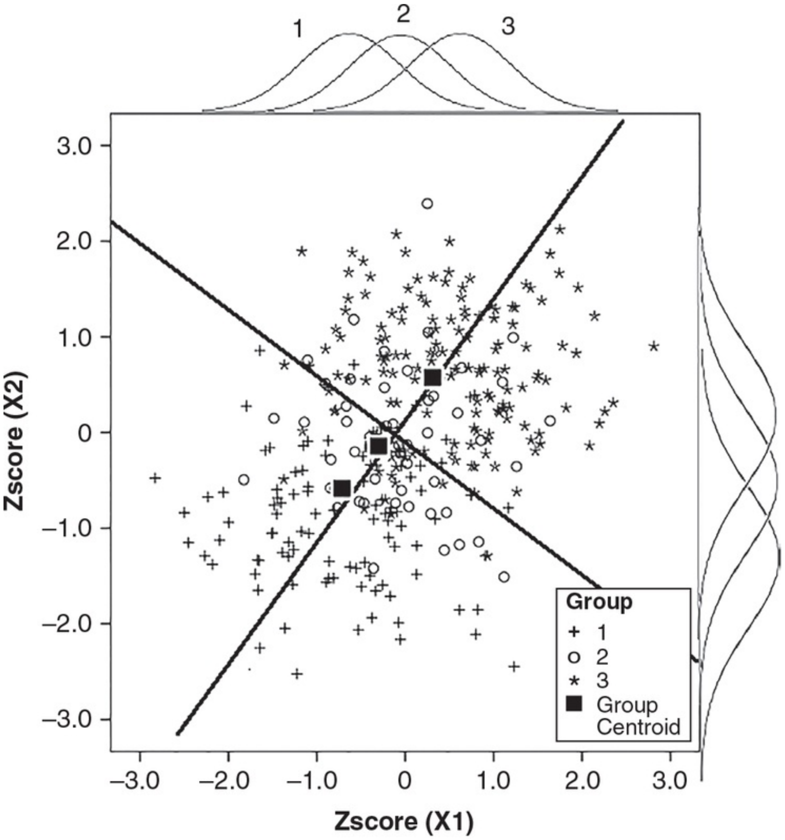
can also be observed in Figure 5. The three points now become (1.77, −0.36), (−0.85, 0.14), and (1.70, 0.85). The population variance for X1′ is 2.22 and for X2′ is .36 for a total of 2.58, the same as the previous axes.

There are values for *a* and *b* such that the variability for X1′ is a maximum and X2′ is a minimum. That is what the matrix operations of discriminant function analysis provide. Basically, it finds an axis (a single dimension) in multidimensional space that maximizes the discrimination between groups. Given the first axis is set, it then finds a second axis (dimension) that maximizes the remaining discrimination between groups. The second axis is orthogonal to the first. The procedure continues until it runs out of groups or variables. At some point, the inclusion of dimensions provides very little additional discriminatory ability and allows the researcher to interpret a much smaller set of variables than the original multivariate data.

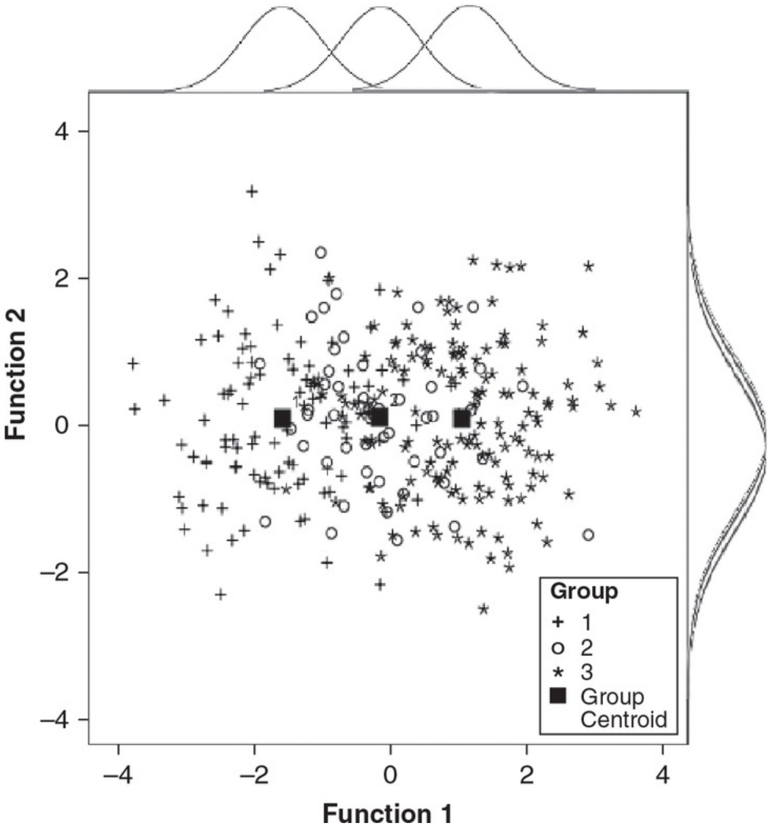
Rotating the Axes to Maximize Discrimination—An Example

An example of the application of discriminant function analysis may be the best manner to illustrate how the procedure works. In this example, there are three groups (1, 2, and 3) and two variables (X1 and X2). Because differential variability of the interval variables can affect the results greatly, the first step in the analysis is to standardize the variables. The scatterplot of the standardized variables for this example appears in Figure 6. The means for the three groups are plotted on the graph and are called group centroids.

Figure 6 Example data scatterplot before and after rotation



Canonical Discriminant Functions



From the marginal distributions in Figure 6, it can be seen that individually, both X1 and X2 somewhat discriminate between the three groups, but the distributions have considerable overlap. Although it would be possible to sequentially apply Bayes' theorem using the two variables, discriminant function analysis first finds a linear combination of the two variables that best discriminates between all groups and then generates a second function that contains whatever is left over.

Applying discriminant function analysis to these data, the first decision is how many factors or dimensions are to be included in the analysis. Inferential and model building techniques are typically used to make the decision, but they are beyond the scope of this entry. For this example, a significant Wilks's λ test and squared canonical correlation greater than .10 (Pedhazur, 1973) suggest using a discriminant function analysis that would result in a single factor. The squared canonical correlation for each discriminant function can be interpreted as the proportion of variability that the discriminant function describes, similar to R^2 in multiple regression. For the discriminant function analysis on the example data, this would result in a single factor.

Even though the analysis would suggest that only a single factor be analyzed, both factors are presented below for completeness sake. The two discriminant functions from the "Standardized Canonical Discriminant Function Coefficients" table are

$$\text{Factor 1} = 0.572 \times z\text{X1} + 0.836 \times z\text{X3}$$

$$\text{Factor 2} = 0.821 \times z\text{X1} - 0.549 \times z\text{X3}.$$

The bottom line is that Factor 1 would be computed for all records, and then a classification system would be employed to classify into appropriate groups. Most statistical packages optionally allow these additional discriminant variables to be created. The results of the applied classification system (equal prior and Bayesian decision process) can be seen in the contingency table. Note that the application of discriminant function analysis in the example resulted in a 71% correct classification.

A scatterplot of both discriminant functions is presented in Figure 6. Note the position of group centroids along the Factor 1 axis and the marginal discrimination of the two functions. The discriminant function coefficients are essentially the β weights of each variable for the discriminant function. They describe the relative importance of that variable in constructing the function, although they must be interpreted with caution, as they have similar issues as the interpretation of β weights in multiple regression. It can be seen in the example that X2 (0.836) contributes to the function to a greater extent than X1 (0.572).

To make predictions using the results of discriminant function analysis, the raw scores need to be standardized using the means and standard deviations of the original data set. Following that, the discriminant functions are computed for each record, and then the classification system is applied relative to the conditional distributions.

If there are more than two variables and two groups, the procedure results in additional discriminant functions equal to the lesser of the number of groups minus one or the number of interval variables. For example, when there are three groups and two interval variables, the procedure will produce two discriminant functions. In almost all cases, however, the procedure will reduce the dimensionality of the original data.

As with any multivariate system of analysis, the more the groups and variables, the greater the complexity of analysis. With three groups and three variables, the first discriminant function would be the line through the multidimensional space that minimized the within-group variance. The second line would be perpendicular to the first and would minimize the within-group residual variability from the first discriminant function. The third discriminant function would be a line perpendicular to the first two and again minimize the within-group residual variability from the first two discriminant functions.

Limitations

Discriminant function analysis has been around since its origin in 1936 with two defined groups by R. A. Fisher. It was later extended by others to include more than two groups. Because of the computational difficulty of the analysis, it was not extensively used until computers became widely available. It has the advantage of describing a complex decision process with a few parameters and producing results that can be interpreted.

The linear models of discriminant function analysis are also its main disadvantage, as many relationships in the real world are not linear. The use of programs that can be trained to use multiple “if-then” statements or neural networks that learn complex relationships with large data sets and estimation of thousands of parameters have eclipsed the use of linear models. The accuracy of these types of programs is generally greater than linear models but comes at a cost to the researcher of not understanding the “why” of the decisions.

A second major disadvantage of discriminant function analysis is the reliance on the assumption of multivariate normal distributions for classification. Although classification decisions can be made without reference to this assumption, when it is made, it is almost certain to be incorrect. How robust the system is with respect to this assumption can be checked with use of two data sets, one for training and one for testing.

Discriminant function analysis offers a powerful tool to discriminate between groups based on creating new variables, called discriminant functions, using linear models of existing interval variables. Measures of accuracy of prediction along with the manner in which the variables combine provide the statistician with a means of understanding multivariate data.

David W. Stockburger

See also Bayesian Statistics; Canonical Correlation; Logistic Regression; Multivariate Analysis of Variance

Further Readings

Johnson, R. A., & Wichern, D. W. (1982). Applied multivariate statistical analysis (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Lantz, B. (2013). Machine learning with R. Birmingham, UK: Packt.

Pedhazur, E. J. (1973). Multiple regression in behavioral research explanation and prediction (3rd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Van de Geer, J. P. (1971). Introduction to multivariate analysis for the social

sciences. W. H.

Howard T. Everson Howard T. Everson Everson, Howard T.

Discrimination Index Discrimination index

532

534

Discrimination Index

When developing educational and psychological tests, particularly for high-stakes uses, psychometricians look carefully at the statistical properties of the items making up the test. In general, the goal is to calibrate statistically two key item indices—the items' difficulty and their ability to discriminate among and between examinees. With respect to the latter characteristic, measurement specialists look at how well the test items discriminate among and between examinees with varying levels of the abilities measured by the test. Consider, for example, a test designed to select students for admission to college. This test ought to include items that discriminate between high-and low-achieving examinees—offering sound evidence in support of the colleges' selection decisions.

Estimating the Discrimination Index

When creating tests, psychometric specialists often employ two psychometric methods for estimating a test item's discrimination index—one based on classical test theory (CTT) and the other on item response theory (IRT). The CTT approach draws on traditional statistical methods such as the correlational analyses for estimating item discrimination indices. The CTT framework, for example, offers two related methods for computing this index—the *biserial* correlation coefficient and the *point biserial* correlation coefficient. These methods quantify the relationship between an examinee's performance on a given item (correct or incorrect) and the examinee's score on the overall test. For purposes of discussion, only the *point biserial* correlation is described here.

The *point biserial* correlation coefficient, referred to as r_{pb}, is used to estimate the correlation of quantitatively and continuously measured variables (e.g., a test score) and the dichotomous variable (e.g., the binary item score for a correct or incorrect response). This correlation is expressed as:

$$r_{pb} = (Y_1 - Y_0) \bullet sqrt(pq) / \sigma_Y,$$

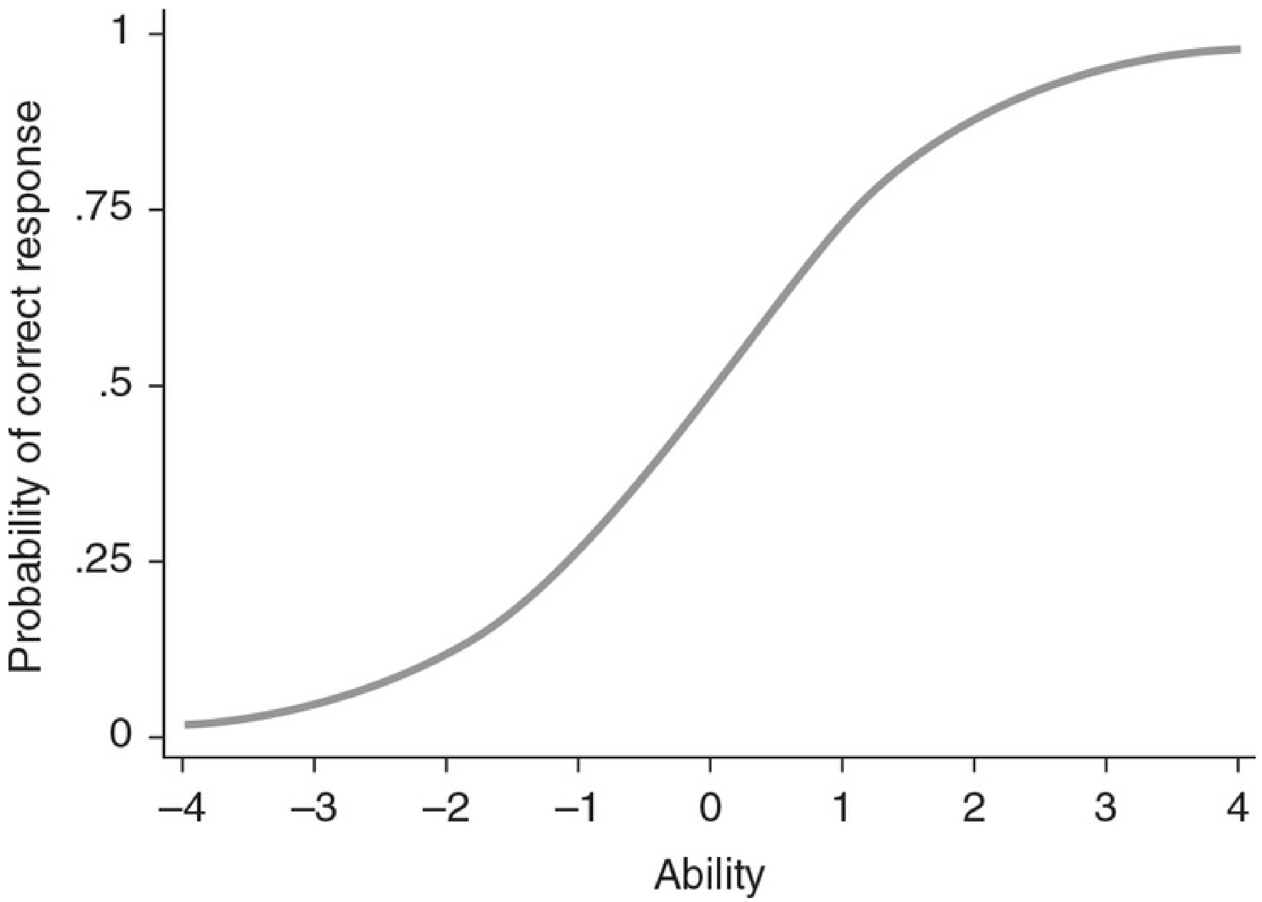
where Y_0 and Y_1 are the Y score means for data pairs with an x score of 0 and 1, respectively, $q = 1 - p$ and p are the proportions of data pairs with x scores of 0 and 1, respectively, and σ_Y is the population standard deviation for the y data. The possible range of the discrimination index is −1.0 to 1.0; however, if an item has a discrimination index below 0.0, it suggests the higher ability examinees are getting the item wrong, whereas paradoxically the lower ability examinees are getting the item right. Similarly, a negative discrimination value suggests the test item is likely measuring something other than the targeted test construct.

Like other CTT indices, an item's discrimination index is sample dependent. The underlying or latent ability levels of the sample examinees interact with the estimates of the difficulty of the test items and, in turn, with the calibration of item discrimination. To address this and other sample dependency problems, psychometric specialists have developed a series of statistical models referred to collectively as IRT.

The IRT framework rests on the idea that the constructs of interest (e.g., cognitive ability, personality characteristics, and attitudes) are latent and not directly observable. Test developers are interested in how each item relates to the latent trait, and how the entire group of items relates to that trait or ability.

The IRT approach assumes the relationship between item characteristics and the latent ability can be modeled, for example, by a two-parameter logistic function (the parameters provide estimates of the difficulty and the discrimination indices). Two assumptions undergird most IRT models—the first is that a unidimensional structure of the test data (measuring one primary construct) and the other relates to the mathematical (logistic) form of the item characteristic function or curve. Figure 1 shows the general form of item characteristic function or curve for a one-parameter model.

Figure 1 One parameter item characteristic curve



In this model, the item characteristic function (the difficulty index) is generated from the expression in Equation 1.

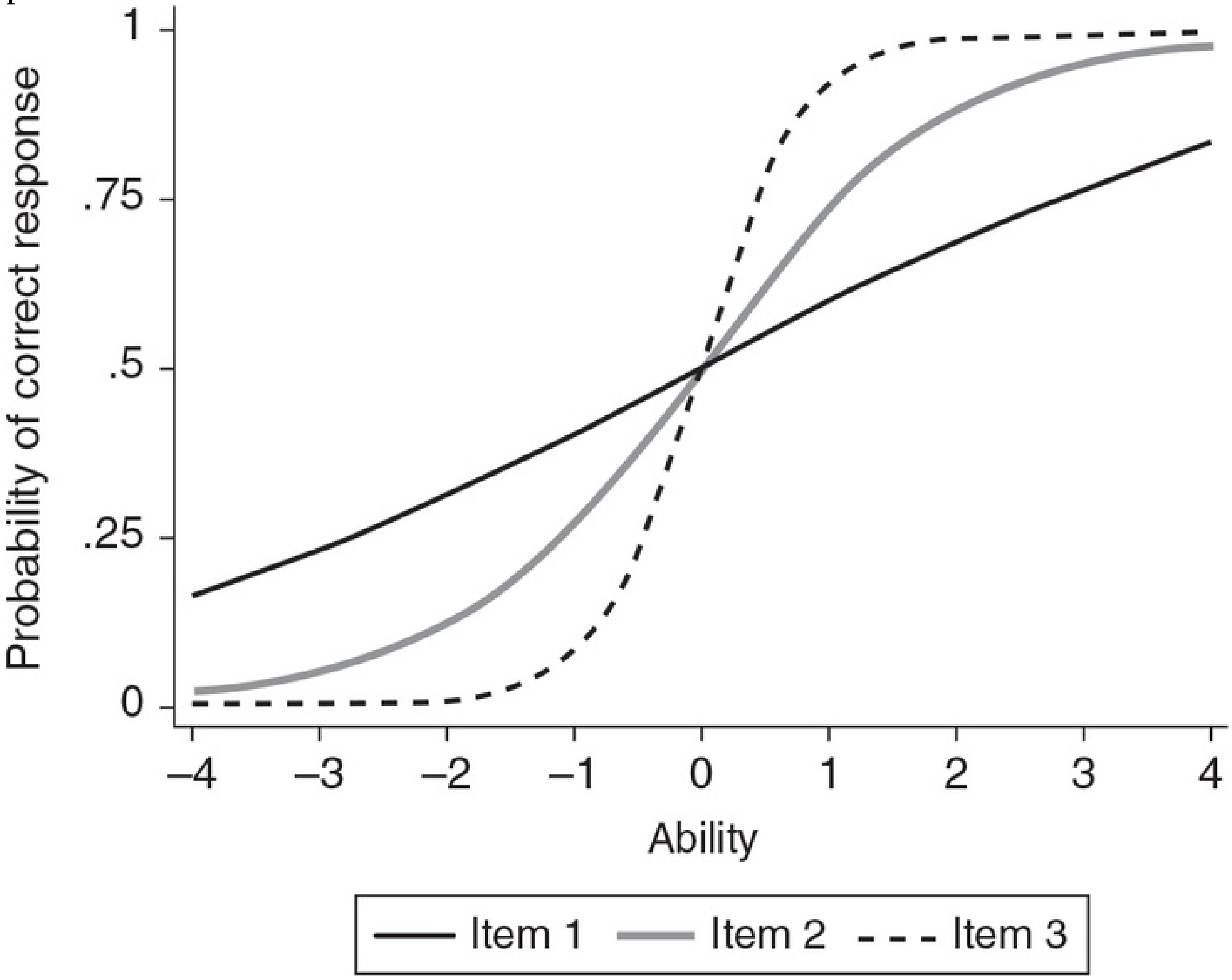
$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}.$$

Here P_{ij} (θ_j, b_i) gives the probability of a correct response to item *i* as a function of ability (denoted θ). The *b* parameter, the difficulty index, is the point on the ability scale (θ) where an examinee has a .5 probability of a correct answer. This one-parameter model is expanded to include a second item parameter, the discrimination index. In the two-parameter model, the probability of a correct response is estimated in Equation 2:

$$P(x_j = 1 | \theta_k, a_j, b_j) = \frac{1}{1 + \exp[-1.7 a_j(\theta_k - b_j)]} \equiv P_{jl}(\theta_k),$$

where x_j is the score on item j (1 = *correct* and 0 = *incorrect*); $a_j > 0$, is the slope parameter of Item j, estimating its discrimination index; and b_j is the parameter of Item j estimates the item's difficulty. A high a_j parameter value suggests an item differentiates well between examinees. Item characteristic curves for 3 items—*Item 1, Item 2, and Item 3*—with different discrimination parameter values are shown in Figure 2.

Figure 2 Item characteristic curves for three items with varying discrimination parameters



The difficulty parameter values for these three items are all zero. The discrimination parameter values are 0.3, 1, and 2, respectively. Notice that as the discrimination parameter increases, the item characteristic function or curve becomes steeper around 0. *Item 3* appears to differentiate examinees with ability values around zero more efficiently than *Item 1*. Using an IRT model, we can estimate the discrimination index, for each item on a test trait.

Summary

The CTT framework expresses the discrimination index as a correlation coefficient and, to a large degree, is sample dependent. IRT addresses the sample dependency problem by offering a statistical approach that models the relationship between an item's discrimination index and the examinees' ability using a two-parameter logistic function. In the IRT framework, the *a* parameter serves as the item discrimination index.

Howard T. Everson

See also Criterion-Referenced Interpretation; Difficulty Index; Item Analysis; Item Information Function; Norm-Referenced Interpretation

Further Readings

Haladyna, T. M., & Rodriquez, M. C. (2013). Developing and validating test items. New York, NY: Routledge.

Mellenbergh, G. J. (2011). A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests. The Hague, the Netherlands: Eleven International.

Osterlind, S. J. (2010). Modern measurement: Theory, principles, and applications of mental appraisal. (2nd ed.). Boston, MA: Pearson.

Carl B. Hancock Carl B. Hancock Hancock, Carl B.

Dissertations

Dissertations

534

537

Dissertations

A dissertation is an extended scholarly document written to describe an original and independent research study conducted by a doctoral student after completing all coursework and in partial fulfillment of the requirements for the doctor of education or doctor of philosophy degree. Written for an expert audience, a typical dissertation begins by introducing an argument that leads readers toward a set of questions the student intends to answer and investigate. A thorough review of prior scholarship situates the investigation, and the dissertation continues with a description of the study's procedures, analyses, and results. The document generally concludes with a discussion of the findings while considering prior research, theory, and practice. A dissertation is generally expected to make an important and original contribution to a discipline and therefore serves as a significant source of novel information for researchers and practitioners. It also provides evidence of the student's ability to conduct research, which is confirmed by the student's doctoral committee. This entry explores the history of the dissertation, role of the doctoral committee, general contents of the document, an alternative dissertation format, expected outcomes, and publication.

Origins and History

Beginning as early as the middle ages, contemporary dissertations evolved from the oral public debates, also known as academic disputations, held at colleges. The debates served as both a research method and a test of students' knowledge. Participants included faculty presiders, who introduced the topic of the debate, and students of various ranks, who debated the merits of various positions with one another. A written summary of the debate, called the *question disputata*, explained the topic, lines of reasoning, counterarguments, and final conclusion, which was rendered by the presider. Authorship of the document was often indeterminable and could range from the faculty presider to one or more of the student participants. These writings were frequently exchanged for books and manuscripts from other institutions, thereby disseminating innovative and interdisciplinary ideas among scholars. During the age of enlightenment, the creation of written disputations, penned by a single author, subsumed oral debates but retained the general style and purpose of the earlier form. The 19th and 20th centuries saw the emergence of the modern research university and its focus on scientific thinking, which advanced the rigor, scope, and sophistication of the methods and analyses documented in dissertations. During the 21st century, the integrated dissertation emerged as an alternative to the lengthy monograph. The former was a set of peer-reviewed manuscripts investigating a theme and bookended by an introduction and conclusion, while the latter concentrated on a single topic. The second decade of the 21st century gave rise to the acceptance of large-scale projects substituting the written dissertation. Examples include original music compositions, curated archives of artifacts, hand-drawn illustrations, sophisticated websites, and innovative computer code. In general, during the first 15 years of the 21st century, 819,450 dissertations from around the world, including 144,375 (17.6%) education dissertations, were added to the 1.76 million dissertations from the 20th century in the ProQuest Dissertations … Theses database.

Doctoral Committees

The purpose of the doctoral or dissertation committee is to supervise the work of doctoral student, direct the selection of a topic, evaluate the scientific merit of the study, and assess the quality of the final manuscript—on behalf of the community of scholars. Three to five faculty, who are experts in their respective areas, comprise the typical doctoral committee. Sometimes, one member serves as an outside committee member who specializes in a different academic field than the student and committee members. Another member of the committee serves as the student’s dissertation advisor. The advisor is the de facto chair of the dissertation committee, directs the dissertation, provides support and resources, and works closely with the doctoral student on the dissertation. Occasionally, more than one committee member will share the advisor role and serve as coadvisors.

Dissertation Process

A typical dissertation requires 1–3 years to complete. Effort on the dissertation begins after a student completes the necessary coursework and required exams to become a doctoral candidate. The dissertation advisor helps the student select a project that makes an important and original contribution to knowledge and can be completed within a reasonable amount of time. A written proposal for the project, referred to as a prospectus, is then presented to the doctoral committee for approval. A typical prospectus includes:

- a proposed title,
- introduction,
- selective literature review,
- study purpose,
- research questions,
- research plan,
- proposed methods,
- instruments and measures, and
- bibliography.

If the study involves collecting data from or working with children, students, or adults, then approval is secured from a university committee, the institutional review board, for human subjects review before the research begins. Throughout the process of writing and conducting research, the doctoral student works under the supervision of the advisor and with consultation with other doctoral committee members. When the dissertation is complete, the committee attends a formal defense of the dissertation, queries about the contents of document, and provides additional feedback. It is typical for the dissertation defense to include an oral presentation for the committee and public before the committee privately weighs the merits of the dissertation and renders a decision to pass, pass with revisions, reschedule, or fail. Some dissertation defenses have included live streaming of the oral presentation and live tweeting or posting of major points of the defense on the Internet.

Topic Selection

Dissertation topics are expected to be specific, novel, germane, and manageable. In general, there are two approaches to topic selection. The student can pursue a study that fits in with the student’s major professor’s research agenda, as is common in science departments, or strike out independently to pursue a unique research topic. Although working within an existing research program offers security for the student, an independent project promotes researcher independence and encourages a strong sense of ownership over the dissertation by the doctoral candidate.

Parts and Contents

The main body of a dissertation typically includes five chapters:

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: Method/Methodology/Procedures
- Chapter 4: Results
- Chapter 5: Discussion/Conclusion

The introduction includes an argument for investigating a research problem and presents an original thesis in response. Additional contents include background research, research questions, guiding questions, hypothesis, definitions, the purpose/objective of the study, and a description of the scope and significance of the dissertation.

The literature review contains a comprehensive critique of prior scholarship surrounding the research problem and thesis. The chapter provides a deep understanding of the topic, frames the topic within a broader field, and explores important methods, debates, and theories. Typically, the contents of the chapter position the dissertation in the field, thereby ensuring the study makes a unique contribution to the literature.

The methods or methodology chapter outlines the design of the study and describes the selection of participants, methods used to collect data, procedures for data analysis, and plan for addressing assumptions, limitations, validity, and rigor.

The results chapter presents the collected evidence and details the findings of the analysis by showing what was discovered in support or in refutation of the thesis and in answer to the research questions. Most quantitative dissertations rely on and in answer to the research questions. Most quantitative dissertations rely on the use of data tables, graphs, and statistics to summarize their findings, whereas qualitative dissertations tend to provide descriptive evidence such as excerpts from transcripts and other artifacts.

The discussion chapter summarizes the main results of the study, relates them to the research questions, and places the findings within the body of prior research and theories in the topic area. The discussion typically closes with conclusions and implications that may be applied to the field of study.

Additional materials in a dissertation include front matter, such as the title page, copyright page, abstract, table of contents, list of tables, list of figures, glossary of terms, acknowledgments of assistance, and a dedication, whereas the back matter includes the bibliography, appendices, and an abbreviated curriculum vitae about the dissertation’s author.

Integrated Dissertations

Although the vast majority of dissertations involve the process and format described in the previous section, there has been growth in the use of alternatives. There is increased pressure for doctoral students entering academia to prepare for the “publish or perish” culture of higher education by publishing journal articles prior to graduation. The tradition of producing a “book-like” manuscript does not map well onto that preparation, nor does it for graduates who will pursue positions in nonacademic settings such as industry or government, where there is an expectation for experience with products and portfolios.

The integrated dissertation emerged as a more appropriate type of preparatory experience. It takes on the format of a cumulative or portfolio dissertation and is a collection of research manuscripts that form a cohesive body of research rather than the traditional single study. Typically, the first and last chapters tie the collection together with an overarching context, rationale, and conclusion. Manuscripts used in the portfolio typically have been published or accepted for publication before the student defends the dissertation.

Outcomes

Successful completion of a dissertation results in a culminating product and an experience that serves as the conclusion of a doctoral program. As such, the process helps candidates hone writing skills, develop strategies for investigating research problems, interact in the community of researchers, and identify needed research within a discipline. It often results in the candidate creating a unique niche in the candidate's selected field and becoming an "expert" in an area, sometimes surpassing the knowledge of the subject held by the dissertation advisor and committee. In some cases, the dissertation has a significant impact that leads to changes such as the modification of a concept, development of a new method, or solution to a persistent challenge plaguing a discipline.

Publication and Distribution

Because dissertations are expected to extend the boundaries of knowledge and understanding, it is expected that the document is of publishable quality. In some fields, the dissertation itself can be edited and published as a book with an academic or professional press. In other fields, the student is expected to rewrite the dissertation into a journal length article and submit it for publication.

Prior to the 21st century, bound physical copies of dissertations were submitted to college and university libraries, and older dissertations were archived using microfilm technologies. Today, most dissertations in the United States are electronically uploaded to ProQuest's Dissertation … Thesis database. Once submitted, the dissertation's abstract, title, and metadata are immediately archived and the document becomes available in print, digital, and microfilm forms.

On occasion, a delayed release of a dissertation, referred to as an embargo, can be used to delay release and publication of the work to limit access to the complete text, although the general information and the abstract will be accessible. An embargo is requested by an author to provide time to publish the dissertation as a book or chapters as articles before the dissertation is publicly available.

Other publication options for the dissertation include submitting to open-access publishers and acquiring a Creative Commons copyright, so electronic versions of the dissertation can be shared and distributed online or posted to other online repositories.

Carl B. Hancock

See also Abstracts; Human Subjects Protections; Journal Articles; Methods Section; Pilot Studies; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Breimer, D. D., Damen, J., Freedman, J. S., Hofstede, M., Katgert, J., Noordermeer, T., & Weijers, O. (2005). Hora Est! On dissertations. Leiden, the Netherlands: Leiden University. Retrieved from https://www.lorentz.leidenuniv.nl/history/proefschriften/hora_est.pdf

Buckler, S., & Walliman, N. (2016). Your dissertation in education (2nd ed.). Thousand Oaks, CA: Sage.

Clark, W. (2006). Academic charisma and the origins of the research university. Chicago, IL: The University of Chicago Press.

Cone, J. D., & Foster, S. L. (2006). Dissertations and theses from start to finish: Psychology and related fields (2nd ed.). Washington, DC: American Psychological Association.

Roberts, C. M. (2010). The dissertation journey: A practical and comprehensive guide to planning, writing, and defending your dissertation (2nd ed.). Thousand Oaks, CA: Sage.

Websites

ProQuest. http://www.proquest.com/products-services/dissertations/

Swapna Kumar Swapna Kumar Kumar, Swapna

Michael Kung Michael Kung Kung, Michael

Distance Learning

Distance learning

537

539

Distance Learning

Classroom learning presumes the presence of a teacher and learner in a physical space where teaching and learning takes place. When the learner and the teacher are not located in the same physical space, and use a communications medium to interact with each other over a distance, it is termed *distance education* or *distance learning*. Distance learning was practiced in various forms (e.g., correspondence courses or programs) in the 19th century but has experienced exponential growth since the early 1990s with the advent of the Internet, leading to increased research about theories, design, development, infrastructure, implementation, evaluation, quality, use, and support structures of distance learning. This entry begins with a brief overview of the development of distance learning, describes key features in its implementation, and ends with a short summary of research in distance learning.

Development of Distance Learning

Many forms of communication media have been used for teacher–learner communication in distance learning through the ages, from print materials, audio tapes, videotapes, CD ROMs, to lately, the Internet. Distance learning courses involve the creation of materials for distance learners in a specified format or medium. In the case of print materials or audio tapes/videotapes in the past, these were mailed to learners, who reviewed the materials, completed assignments, and mailed them back to the teachers. The teachers then graded the materials and provided feedback that was mailed back to the learners. Internet and communication technologies have speeded up this process tremendously. The materials created for learners are now available in an online space, learners can access the materials at any time, interact with instructors and peers using e-mail or other forms of communication, submit assignments within that space, and receive feedback electronically. This form of distance learning, also termed *e-learning* or *online learning,* enables learners to interact with and learn from each other, where before they only interacted with the instructor(s). Advances in technologies also enable the use of multiple media in online learning (e.g., videos, audios, and online texts) that can help diverse learners understand content in different ways. Online courses are hosted within a learning management system that represents a closed and protected online classroom space available to learners enrolled in the course. Learning management systems include several areas for teachers to make content (e.g., documents, videos, and links to websites) available to students and create quizzes that students can take, where students can interact with each other and with the teachers (discussion forums, group rooms), and for assignment submission and feedback.

In the early 2010s, massive open online courses (MOOCs) hosted on platforms such as Coursera, Udacity, and EdX gained popularity because learners from around the globe could access them or enroll in them and learn free of cost. Despite having high interest and high enrollment rates, MOOCs report high dropout rates and low completion rates. MOOCs provide an opportunity for universal education through distance learning and for increased informal learning but are designed such that the learner is responsible for course completion. Critics of MOOCs also claim that one of the reasons why MOOCs have low completion rates is because they are unable to replicate the intellectual community found in a physical classroom.

Distance Learning Implementation

Communications media that are used for distance learning are characterized as synchronous and asynchronous media. E-mail, a commonly used medium of communication today, is asynchronous because the communication does not occur at the same time. A telephone conversation or a video-based chat (e.g., Skype) is a form of synchronous communication because the message is sent and received simultaneously. Although asynchronous media such as e-mail, messaging, or discussion forums are used more widely, synchronous media in the form of virtual classrooms (where the instructor and students meet at the same time or students meet in real time to complete class projects) and video same time or students meet in real time to complete class projects) and video classrooms (where the instructor is projected on a large screen into a classroom of students) are also prevalent in distance learning. Asynchronous distance learning provides flexibility and allows students time and opportunity to review materials, reflect, and then participate. Synchronous distance learning enables students and instructors to interact in real time and feel more connected to each other.

Distance learning can be isolating and difficult for learners not only because they are at a geographical distance from the instructor and their peers but also because they experience transactional distance, described by Michael Moore as a psychological distance that distance learners experience and that can be reduced through increased interactions in the communications medium. Learners in an online course need more support than learners in a classroom because it is often difficult to ask a question of the teacher or ask a peer to clarify their doubts during a class meeting. Distance learning materials and online courses have to be designed for learners to be able to learn the content effectively and understand what is expected of them. A combination of pedagogical approaches and the process of instructional design are used for this purpose. Institutions of higher education have teams of instructional designers who assist faculty members in designing materials for distance learning based on learning theories and research in educational psychology, multimedia design principles, emerging technologies, and instructional strategies for distance learning. Distance learning courses can be paced courses, in which learners go through the course together and interact with each other while completing activities at a specific pace, or self-paced courses, in which they complete all learning materials at their own pace and at their convenience. Paced courses offer more opportunities for learners to feel connected to their peers and their instructor and receive feedback on their learning, while self-paced courses are beneficial for learners who have other commitments and would like to study on their own time.

Distance learning makes it possible for learners to study from anywhere, thus increasing equity and access to education for all student populations, especially working professionals, those living at a distance from quality educational institutions and those caring for others. It is increasingly a part of teaching and learning at all levels, in all disciplines, and multiple contexts, such as K–12 schools, virtual schools, corporate education, vocational education, post-secondary institutions, the military, and nonprofit education. Institutions embarking on or implementing distance learning have to consider accreditation and compliance procedures; strategic partnerships and collaborations;

institutional resources and infrastructure; usability, intellectual property, and accessibility policies; and support for students, faculty, and course development, among other factors. Organizations in various countries provide benchmarks and quality indicators to guide institutions in distance learning initiatives (e.g., Australian Council on Open, Distance, and e-Learning; European Association of Distance Teaching Universities). In the United States, the Institute for Higher Education Policy, Quality Matters, and the Online Learning Consortium provide guidelines and resources for assuring quality and creating successful online learning programs.

Research on Distance Learning

The American Journal of Distance Education, published in 1987, was the first scholarly journal focused on distance learning and was followed by several journals devoted to various aspects of distance learning. Early research on distance learning focused on comparisons between classroom instruction and distance learning, attempting to establish whether learners in distance learning environments demonstrated the same learning outcomes as those in classroom environments. However, research since the 2000s has centered on various aspects of distance learning as an independent educational phenomenon. Information technology, education (e.g., educational technology, educational leadership), psychology, learning sciences, neuroscience, communication sciences, and media sciences are some of the leading disciplines that contribute to scholarship about distance learning. Research on distance learning encompasses a wide range of topics (e.g., the design, prototyping, and usability of technologies used for distance learning); theoretical frameworks, cultural approaches, and models of distance learning; learner diversity, preferences, interactions, self-regulation, and support in distance learning; human–computer interaction, cognition, and community-building in distance learning environments; and management, strategic planning, scalability, and equity in distance education. The research designs were qualitative, quantitative, and mixed methods; however, after the early 2000s research in distance learning saw an increased use of design-based research, learning analytics, social network analysis, and neuroscience methods. The Babson Research Group, the Online Learning Consortium, the Association of Educational and Communications Technology, the International Society of the Learning Sciences, and the International Association for K–12 Online Learning are excellent sources of reports on developments and research in distance learning in the United States.

Swapna Kumar and Michael Kung

See also Learning Theories; Professional Learning Communities; Self-Directed Learning; Social Network Analysis; Universal Design in Education

Further Readings

Anderson, T. (Ed.). (2008). Theory and practice of online learning. Edmonton, Canada: Athabasca University Press.

Moore, M. G. (1993). Theory of transactional distance. In D. Keegan (Ed.), Theoretical principles of distance education (pp. 22–38). London, UK: Routledge.

Moore, M. G. (Ed.). (2013). Handbook of distance education. New York, NY: Routledge.

Moore, M. G., & Kearsley, G. (2011). Distance education: A systems view of online learning. Belmont, CA: Wadsworth.

Zawacki-Richter, O., & Anderson, T. (Eds.). (2014). Online distance education. Towards a research agenda. Edmonton, Canada: Athabasca University Press.

Stella Bollmann Stella Bollmann Bollmann, Stella

Distributions

Distributions

539

544

Distributions

A probability distribution provides probabilities for all possible values of a (random) variable. For example, if the variable *X* is gender, the probabilities might be 0.5 for *X* = male and 0.5 for *X* = female. This assignment can be represented in a graphical illustration or in a mathematical formula. There are two types of distributions: discrete and continuous. Whether the distribution is discrete or continuous depends on the random variable.

Random Variables

Technically speaking, a probability distribution is a representation of the probabilities of all possible outcomes of a random phenomenon. A random phenomenon can be an experiment or a measurement, for example. In this entry, the example of a random phenomenon will be to flip a coin twice and record each flip. The set of all possible outcomes of the random phenomenon is called the *sample space*. Flipping a coin twice yields the following sample space: S = {HH, HT, TH, TT}, where H stands for head and T for tail. Dependent on the information of the data one is interested in, a *random variable* assigns a number to each outcome of the sample space. For example, if we are interested in the number of heads in this random phenomenon, we get a random variable that has a 2 for the outcome in which we observe HH, a 1 for both outcomes HT and TH, and a 0 for TT. Another example for a random variable could be to record if the first flip is a head (1) or not (0). Then we would assign a 1 to the outcome HH and HT and a 0 to TH and TT.

If a random variable is discrete, any outcome can have a natural number assigned to it. No further number can be added between these two. For example, a random variable that indicates how many items of an exam were answered correctly is discrete.

A random variable is called continuous whenever there are theoretically an infinite number of values between any two values. Therefore, it is not possible to assign a natural number to any possible outcome. Here, decimal numbers are used. One example of a continuous random variable would be if the time a person takes to complete an exam was measured.

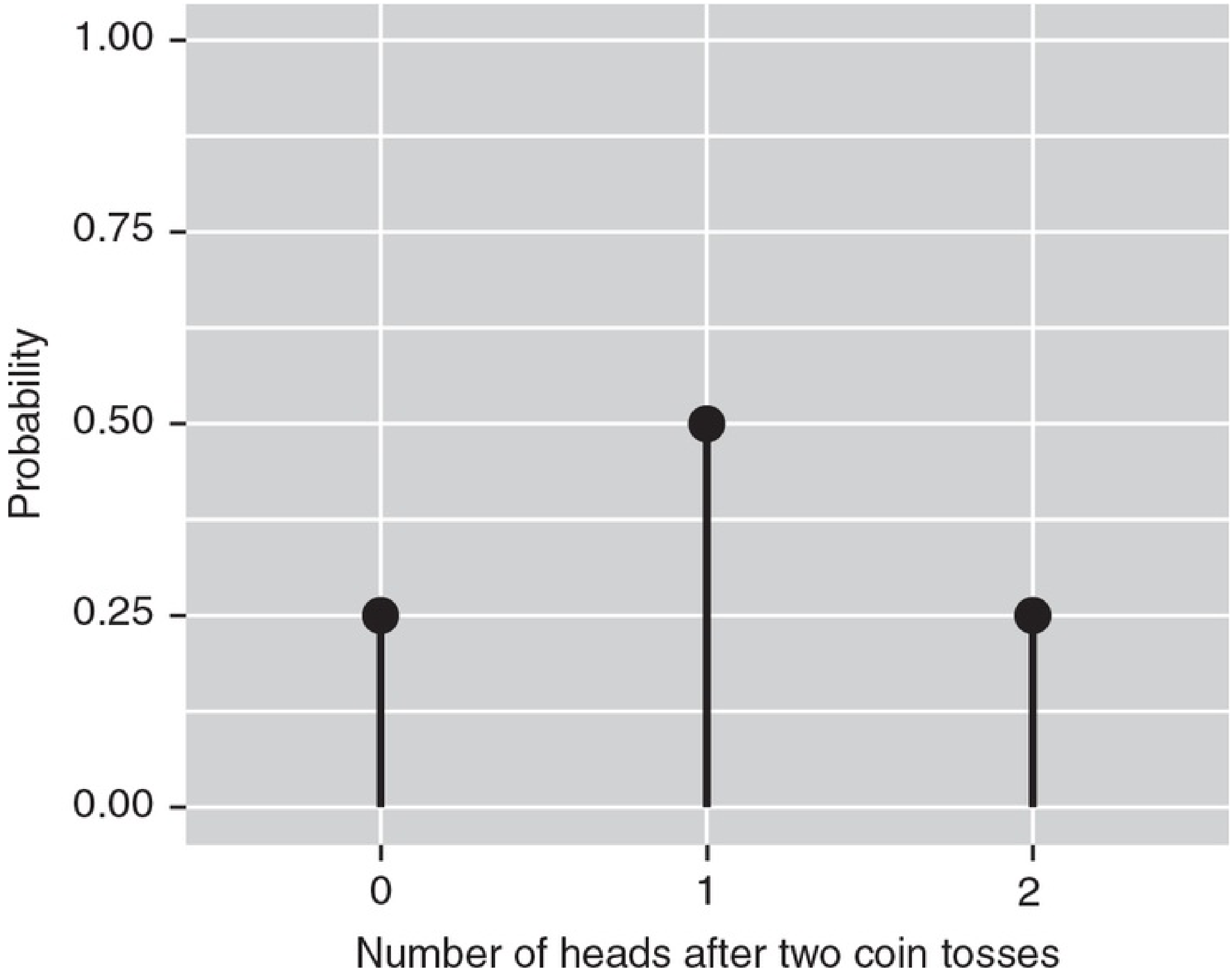
In the coin flipping experiment, the random variables are discrete. One example for a continuous random variable for this example would be the amount of time it takes to flip a coin until head appears once.

X can be used to represent the random variable, while x represents a particular value of the variable.

Discrete Probability Distributions

If X is a discrete random variable, each value x has a specific probability $P(X = x) = p(x)$. Probabilities are always indicated in numbers between 0 and 1 with 1 being the certain outcome. Formally, discrete probability distributions are described by probability functions. These functions are formulas that assign a probability to each single outcome of a discrete random variable. The sum of all probabilities of all single events of one random variable is 1. It describes the probability that any of the possible outcomes of the random variable is observed and this probability is 1. For the random variable number of heads flipping a coin twice, we get the following probabilities: $P(X = 0) = 1/4$, $P(X = 1) = 1/2$, and $P(X = 2) = 1/4$, and the probability distribution takes the form shown in Figure 1.

Figure 1 Probability distribution of the discrete random variable number of heads after two coin tosses.

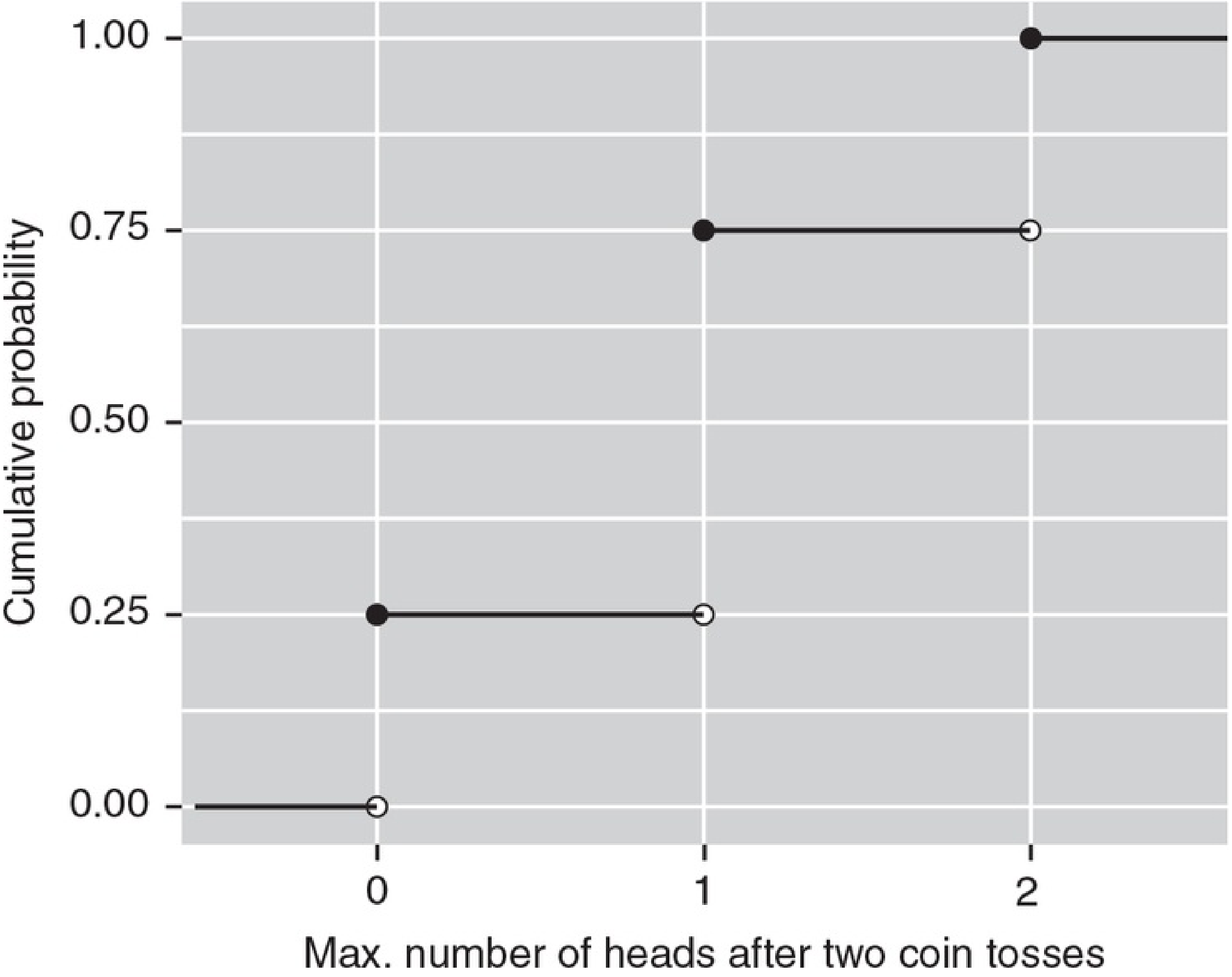


Continuous Probability Distributions

In the case of a continuous random variable, there are theoretically an infinite number of possible outcomes. Therefore, the probability of any individual outcome is 0. For example, the probability that the first head is observed exactly after 2.54344… minutes is 0. The continuous equivalent to the probability function is the probability density function. It describes the continuous probability distribution. The function values of the probability density function are not to be interpreted as probabilities, they are called *densities*. Probabilities can only be determined for intervals of outcomes, and areas under the probability density function are interpreted as probabilities. Probabilities of intervals of outcomes are calculated by integrating over the probability density function, the continuous equivalent to the probability function. It describes the continuous probability distribution. The entire area under the distribution is equal to 1. Equivalent to the discrete case, it represents the probability that any of the outcomes is observed. One example for a continuous probability distribution is the Gaussian or normal distribution. In research, probability distributions are needed for most hypothesis testing. The idea is to find the right distribution of the test statistic under the null hypothesis in order to be able to determine the critical value for the test.

The cumulative distribution function (CDF) of a random variable X, written $P(x)$, gives the probability of observing a value of the variable that is less than or equal to a particular value. For example, the probability for the example that the number of heads is less than or equal to 1 is $P(X \leq 1) = P(X=0) + \mathrm{P}(X=1) = 0.25 + 0.5 = 0.75$. This CDF can also be represented graphically (see Figure 2).

Figure 2 Cumulative distribution function of the random variable number of heads after two coin tosses.

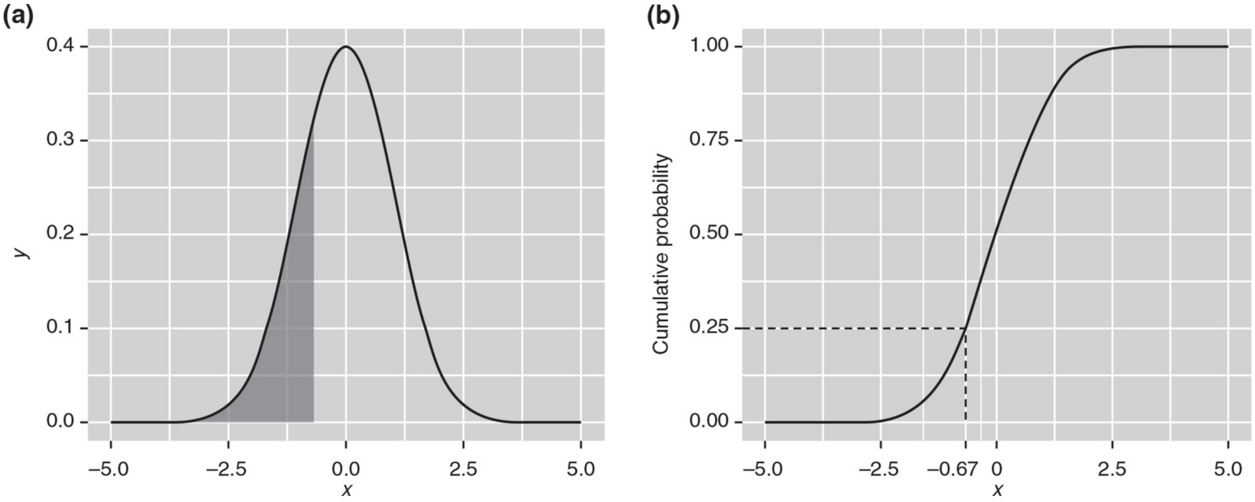


For continuous random variables, the CDF is represented by an integral over the probability distribution:

$$F(b) = P(X \geq b) = \int_{-\infty}^{b} f(x) dx.$$

For example, the probability $P(X \leq -0.67)$ can be represented graphically as shown in Figure 3.

Figure 3 (a) Graphical representation of the probability $P(X = -0.67)$ as the area under the probability density function of the normal distribution. (b) Graphical representation of the probability $P(X = -0.67)$ as the function value of a cumulative distribution function.



Expected Value and Variance

Typically, probability distributions are characterized by their expected value and variance. The expected value determines the center of the probability distribution, while the variance specifies how spread out the distribution is around its expected value. The expected value is equivalent to the arithmetic mean for empirical data. For the arithmetic mean, each value x_j is multiplied by its relative frequency, and they are all summed. This principle can be applied to any discrete random variable X and yields the equation for the expected value of the distribution of the random variable X:

$$E(X) = \sum_{j=1}^{m} x_i \cdot p(x_i).$$

The expected value characterizes the sum over all possible realizations *m*, each multiplied by the probability of its occurrence. For example, in a lottery, we have 10 lots out of which 1 is a winning of US$5 and 2 are winnings of US$2. One lot costs US$1. The possible values of the random variable that describes the wins and losses are {−1, 1, 4}, while the value −1 has a probability of 0.7, the value 1 has a probability of 0.2, and 4 has a probability of 0.1. Then the expected value of the distribution of the random variable is

The formula of the expected value for the continuous case is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

The variance of empirical data can be described by taking the difference between each possible value x_j and the arithmetic mean , multiplied by the relative frequency $f(x_j)$ of this value, and they are all summed. If we apply this principle to the discrete random variable *X*, we get:

$$\text{Var}(X) = \sum_{j=1}^{m} \left(x_j - E(X)\right)^2 \cdot p(x_j)$$

and for the continuous case:

$$\text{Var}(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot f(x) dx.$$

The Joint Probability Distribution

The *joint probability distribution* of the two random variables X_1 and X_2 gives the probabilities of each possible combination of values for the two variables. It is denoted by $p(x_i, x_2)$, which stands for the probability that X_1 takes the value x_1 and X_2 takes the value x_2 at the same time. The same terminology applies for both discrete and continuous random variables, and it can be extended to the case of numerous random variables writing $p(x_1, x_2, \ldots, x_n)$. One example of a joint distribution can be constructed for the coin-flipping experiment where we flip a coin twice: Let X_1 indicate the number of heads and X_2 indicates whether both coins are the same or not (with 1 if they are the same and 0 if they are different).

Examples of Distributions

This section examines one example of a discrete distribution and one of a continuous distribution.

Binomial Distribution

The binomial distribution is a discrete distribution that is based on the Bernoulli distribution. Any random variable with two possible realizations 0 and 1 is called Bernoulli distributed if the probability for 1 is *p* and for 0 is $1 - p$. As *p* is a probability, it can take any value between 0 and 1. One example of a Bernoulli experiment could be to draw a random person from the population and record the person’s gender. The probability that this person is male is *p* (not necessarily 0.5) and consequently the probability for this person to be female is $1 - p$. For the binomial distribution, a number of independent Bernoulli experiments are considered. One example could be to draw more than one person from the population and record the person’s gender. The random variable *X* that counts the number of males in that sample is called binomial distributed. Another example would be to count the number of heads in not only two but *n* independent tosses. The probability of observing exactly *x* heads and $n - x$ tails is given by the probability function of the binomial distribution:

$$P(X = x) = p(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}.$$

The factor p^x determines the probability of exactly *x* heads and $(1 - p)^{n - x}$ is the probability of observing exactly $n - x$ tails. The coefficient is the number of possible arrangements of *x* heads and $n - x$ tails. It is called the binomial coefficient.

p and *n* are called the parameters of the distribution. By fixing *p* and *n* each to a particular value, the distribution is defined. For a random variable *X* that is binomial distributed with parameters *p* and *n*, we write:

$$X \sim B(n;p).$$

The expected value of the binomial distribution is *np* and its variance is *np*(1 − *p*).

Let us consider the example of flipping a coin 10 times with the probability of heads being *p* = .5. We say the random variable *X*, number of heads, is binomially distributed with *n* = 10 and *p* = .5.

$$X \sim B(n=10;p=.5).$$

The probability of observing, for example, 10 heads can be calculated by substituting *n* and *p* in the formula for the binomial distribution:

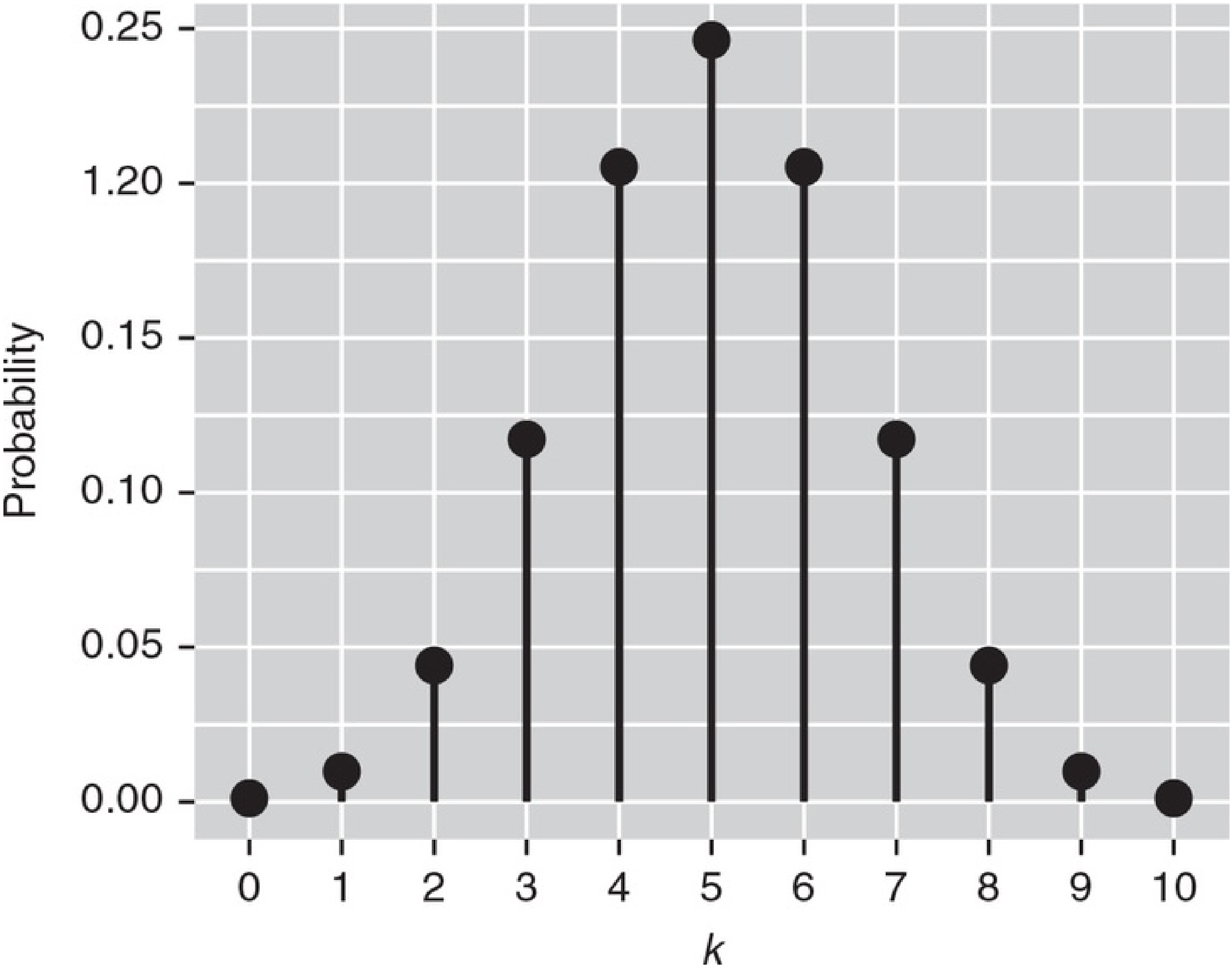
$$P(X=10)=\begin{pmatrix}10\\10\end{pmatrix}0.5^{10}\bullet 0.5^{0}$$

$$=\frac{10!}{10!(10-10)!}\bullet 0.5^{10}\bullet 1$$

$$=\frac{10!}{10!}\bullet 0.5^{10}=0.000977.$$

A graphical representation of this example is shown in Figure 4.

Figure 4 Binomial distribution with *n* = 10 and *p* = .5.



The expected value of our example is $E(X) = n \cdot p = 10 \cdot 0.5 = 5$ and the variance of our example is $\mathrm{Var}(X) = n \cdot p \cdot (1 - p) = 10 \cdot 0.5 \cdot 0.5 = 2.5$.

In the social sciences, the binomial distribution is mostly used for hypotheses relating to variables that only have two values. One common example is the distribution of males and females in a certain population, for example, among psychology students.

Normal Distribution

In social sciences, the normal distribution is the most commonly used continuous distribution.

It has the following probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

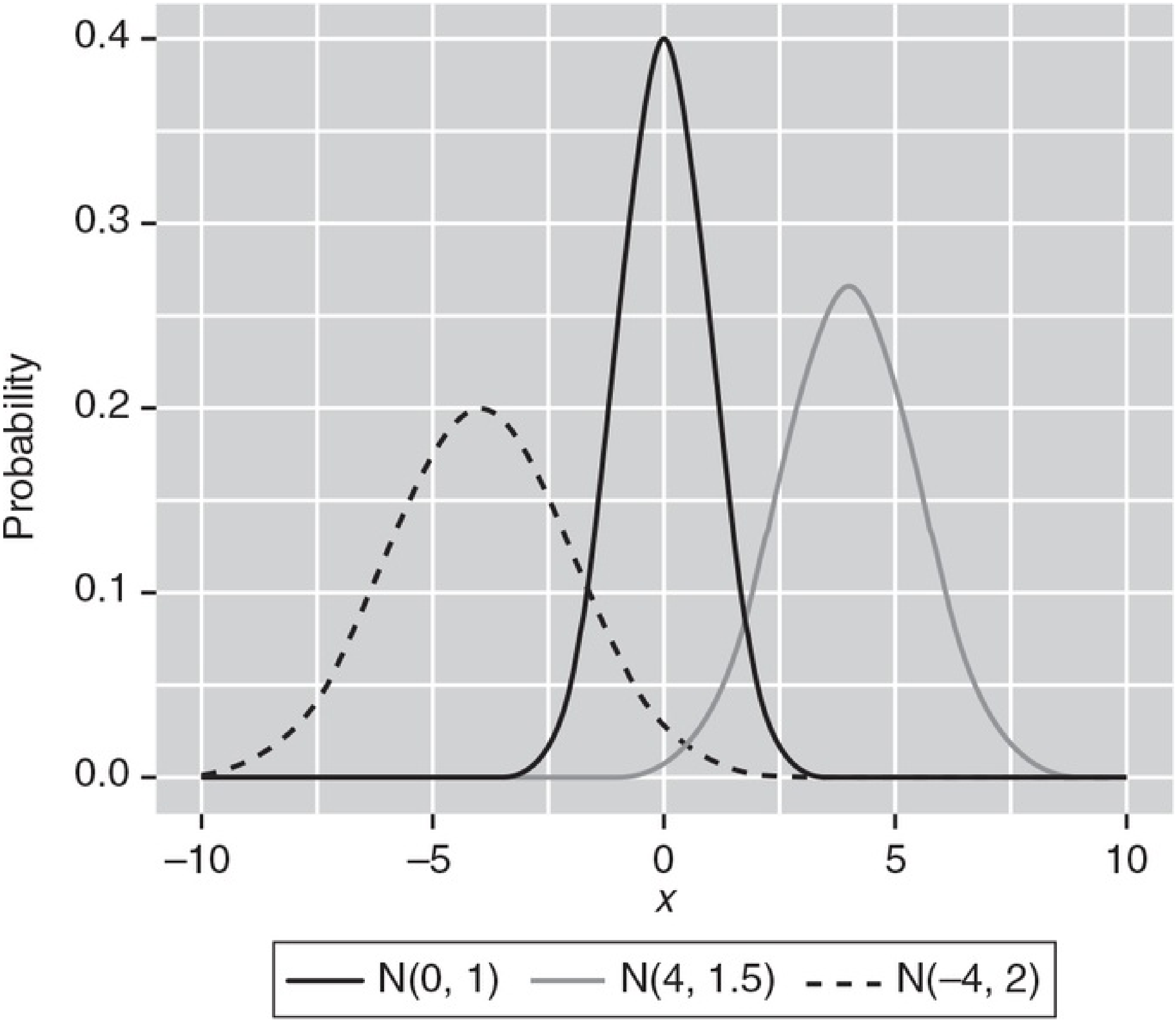
For the random variable that is normally distributed, we can write:

$$X \sim N\left(\mu;\sigma^2\right).$$

Examples of graphical representations of three different shapes of normal distributions can be obtained from Figure 5. The specific shape of the normal distribution is determined by the two parameters μ and σ^2; μ is the expected value (mean) and determines the location of the peak of the distribution and σ^2 is the variance and determines the width of the distribution. Important characteristics of the normal distribution are that it has one maximum at $x = \mu$ and that it is symmetric around μ. A particularly important distribution in the social sciences is the distribution of a standard normally distributed random variable $Z \sim N(0;1)$. It is, for example, important for the calculation of confidence intervals of normally distributed random variables. Any normally distributed random variable can be transformed into a standard normally distributed random variable by using the following formula:

$$Z = \frac{X-\mu}{\sigma}.$$

Figure 5 Probability distributions of three different random variables.



The CDF of the standard normal distribution is tabulated in most statistics books and can be used to determine the probabilities of a random variable being realized in a specific interval around its mean. For example, the probability that a normally distributed random variable is realized in an interval from μ − σ to μ + σ is 0.683. Since the normal distribution is symmetric, the probability that the observed value is below μ − σ is exactly equal to the probability that the observation is above μ + σ; they are both .

In the social sciences, the normal distribution is very commonly used for a couple of reasons. As mentioned earlier, one important application of distributions in research is for hypothesis testing. Different distributions can be assumed for that purpose but for most commonly used statistical methods (e.g., *t* test, analysis of variance, and most regression analyses), the normal distribution is the standard. One reason for that is that the distribution of many phenomena can be approximated by the normal distribution (e.g., height, weight, and intelligence). Another reason is that the normal distribution is the distribution of the arithmetic means of all possible realizations when repeating a random experiment very often ($n - \infty$). Furthermore, many statistical methods (as, e.g., the *t* test and the analysis of variance) assume normally distributed random variables and cannot be properly conducted when this assumption is violated.

Another application of distributions is for modeling data. For example, random measurement errors can be approximated by the normal distribution with mean 0. This principle was derived from the properties of random errors in physics where it is known that the sum over all steps in a process leads to a normal distribution no matter how the individual steps were distributed and therefore the random error over all these steps is normally distributed.

Stella Bollmann

See also *F* Distribution; Normal Distribution

Further Readings

Agresti, A., & Finlay, B. (2009). Probability distributions. In Statistical methods for the social sciences (pp. 73–106). San Francisco, CA: Dellen Publishing Company.

Field, A. (2016). Probability and frequency distributions. In An adventure in statistics: The reality enigma (pp. 228–247). Thousand Oaks, CA: Sage.

Fox, J. (2015). Appendix D: Probability and estimation. In Applied regression analysis and generalized linear models. Thousand Oaks, CA: Sage.

Judith M. S. Gross Judith M. S. Gross Gross, Judith M. S.

Document Analysis Document analysis

544

548

Document Analysis

Document analysis is a form of qualitative research that uses a systematic procedure to analyze documentary evidence and answer specific research questions. Similar to other methods of analysis in qualitative research, document analysis requires repeated review, examination, and interpretation of the data in order to gain meaning and empirical knowledge of the construct being studied. Document analysis can be conducted as a stand-alone study or as a component of a larger qualitative or mixed methods study, where it is often used to triangulate findings gathered from another data source (e.g., interview or focus group transcripts, observation, surveys). When used in triangulation, documents can corroborate or refute, elucidate, or expand on findings across other data sources, which helps to guard against bias. When used as a stand-alone study, it can answer questions about policy, past events, cultural context, organizations, activities, groups, and more. Document analysis is a viable independent research method and should not be considered merely as a supplement to other methods. This entry explores the method of using documents as data and describes the two primary types of data. It also explains how to identify document sources and select an appropriate sample of documents. Creating a system of document management is then discussed, followed by the process of data analysis and some limitations to document analysis.

Using Documents as Data

Documents of all types can be useful for the researcher in uncovering insights, developing theory, and gaining a greater understanding of the topic of study. However, it is important to remember that all documents exist within the context of their creation, meaning that the social, economic, political, and cultural influences of the time and place of their creation contribute to their representation of the construct being studied. Documents provide a record of our existence and our activities from the author’s point of view, and they may be published or unpublished, public or private, hard copy or electronic, textual or visual. There are many types of documents that can be used, which include both primary and secondary sources of data.

Types of Documents

There are numerous types of documents that researchers may use. Documents consist of words and images that have been created or recorded without the influence of the researcher and for a purpose other than the research study. Some examples of types of documents that can be used include policies and regulations, papers about the operation and history of an organization, reports, budgets, newsletters, meeting minutes and agendas, organizational charts, presentations, manuals/handbooks, book chapters, journal articles, white papers, brochures and pamphlets, advertisements, photos, letters/emails, diaries, newspaper articles, posters, event programs, webpages, and maps and charts. These documents may be sorted into two types of data: primary and secondary.

Primary Types of Data

Primary types of data are documents that provide a first-hand account of an event or occurrence, without interpretation or analysis. For example, the minutes from the board meetings of a nonprofit organization would be considered a primary source of data because a meeting member generated the document as a record of the meeting. Other examples of primary sources of data include, but are not limited to, personal letters or e-mails, photos, policies, newspaper articles, and advertisements. Primary sources may not be published and may require more extensive research and greater time to identify and retrieve them from personal, public, or historical archives.

Secondary Types of Data

Secondary types of data include documents that were developed as a result of an analysis of primary sources and interpretation of the construct of interest. These documents were created for the purpose of sharing the interpretation with a wider audience and are often published in the public domain. Therefore wider audience and are often published in the public domain. Therefore, common secondary sources of documentary data are book chapters, research articles, dissertations, literature reviews, and webpages, among others.

Document Sources and Sampling

Document Sources

Source for documents will depend on the research question. There are many large document repositories in the United States such as the National Archives and the Library of Congress. However, for most education research studies, the sources will be more targeted, with the data sources tied to the context of the topic of study. Examples of sources of primary types of data include organization files, personal records and communications, school records, city or county records, and newspaper archives. Examples of sources for secondary types of data include public and university library collections and periodical databases, websites, and clearinghouses.

One way to narrow down the list of potential document sources is through conducting a preliminary search. This can be especially helpful if the potential sources include library databases or clearinghouses with immense amounts of potential documents to peruse. Conducting a preliminary search of the potential sources facilitates identifying additional keywords to add to the search list, clarifying and defining inclusionary and exclusionary criteria, and determining which document sources yield the greatest quantity and quality of documents related to the research question. When conducting the preliminary search, the researcher conducts a simple cursory review of the results returned, investigating closely enough to see patterns and determine quality and quantity of relevant results. This process helps the researcher to identify the most appropriate and fruitful sources to use for sampling.

Document Sampling

It is important to approach document sampling in a systematic way. Keeping a record of decisions made regarding sources searched, methods used, keywords used, and results returned can help to ensure a systematic approach and provide an audit trail for use by other researchers. Additionally, considering the potentially enormous selection of available documents related to the research question and creating a set of clearly defined inclusionary and exclusionary criteria focus the selection of documents and ensure authenticity and representativeness of those identified for the sample.

Inclusionary Criteria

Inclusionary criteria are essential for ensuring systematic document selection and reducing irrelevant data collection. One parameter to consider when developing the inclusionary criteria is determining the age of the documents. For example, the researcher may be seeking the most current representation of the construct of study and as such may choose to limit the time period for document creation to the last 5 years. Or, conversely, the researcher may be seeking to understand an event from the past and will narrow the period to a select time frame in which that event occurred. Another parameter to consider is geographic representation. With globalization and digital libraries, researchers have access to information worldwide. Depending on the question and context in which the answer will be presented, the researcher may choose to limit the data collection nationally, regionally, or even locally. A third criterion could include narrowing the types of documents included in the sample. For example, if analyzing a new policy and its regulations to better understand its future implementation, the researcher would choose to include only official government documents, as opposed to opinion-based editorials about the new policy.

Exclusionary Criteria

Exclusionary criteria are important to narrowing the list of potential documents down to the final sample. Exclusionary criteria help to ensure a systematic approach to final selection while ensuring representativeness, topic and content relevance, and appropriateness. At this stage in the data collection, a more in-depth review of the documents gathered is required. The researcher ensures that the documents gathered in the initial sample using the inclusionary criteria were appropriate to the study and reasonable for addressing the research question. Exclusionary criteria can be used to narrow search results and limit redundant representations in the sample. For example, the researcher may choose to allow only one document per author to ensure diversity of perspective represented in the sample.

Once both the inclusionary and exclusionary criteria have been applied, the resulting documents will comprise the final sample.

Data Collection and Management

The researcher must devise and develop a system for managing and organizing the document sample. It is important to track essential “demographics” on each document to ensure that the collection is representative and on point for the research question. The development of a table identifying essential demographics on each document will help identify them, the context in which they were developed, and place them in relationship to each other. Some of the demographics recommended for collection on each document in the sample include title, author, audience, purpose, date produced, context, and source of the document. In addition to helping the researcher to understand the relationship of the documents in the sample to each other, it also helps ensure representativeness by providing a means of sorting the document records by different demographics to reveal any overrepresentation of author, document type, or source in the sample.

Data Analysis

Analysis of documents may use any number of epistemological or ontological approaches, including but not limited to content analysis, semiotics, discourse analysis, interpretative analysis, conversation analysis, or grounded theory. The approach used and the stance taken by the researcher should be reflective of the purpose of the study and the research question. Regardless of which approach is taken with the analysis, there are some common basic steps in the qualitative document analysis process: coding and categorizing, interpretation, and thematic analysis.

Coding and Categorizing

Identifying the most appropriate codes to use to analyze the documents can be done inductively, deductively, or some combination of the two. Inductively, the researcher selects a small, diverse subsample of documents (6–12) from the whole and conducts open coding, line-by-line, to determine the most appropriate codes with which to code the whole sample of documents. Deductively, the researcher may begin with a collection of initial codes derived from the keywords used in the literature search and the knowledge gained in conducting the literature review for the study. This collection of predetermined categories then form the basis for the codes used to analyze the small document subsample. Either method will result in the development of an initial codebook to apply to the remaining documents.

Once the initial codebook is fully established, with codes organized by categories and subcategories, it should be tested against the subsample again to assure its appropriateness and completeness. Next, the codebook is used to analyze the whole document sample during the focused coding stage. In document analysis, the focused coding stage may involve the use of a worksheet or protocol to facilitate data extraction, ensure systematic analysis, and reduce time burden. With documents, often it is unnecessary to code line-by-line once a codebook is established. This is because documents are developed for innumerable purposes, and many will ultimately have large sections of unrelated information, unlike when analyzing interview or focus group data. Therefore, the use of a worksheet facilitates a more expedient and focused coding of the full sample, while still allowing for the inductive emergence of new codes and revision of existing codes based on comparison with coded data and definitions.

Interpretation

Because the documents gathered for the sample represent a diverse collection of different types of documents, developed for different purposes and for different audiences, the researcher will need to engage in interpretation of the data in order to concisely code and categorize. Sometimes authors do not use the same vocabulary to describe the same concept, requiring the researcher to interpret the meaning based on the other content within the document. In other documents, the author may merely describe the construct under study without ever naming it, so the researcher must then interpret the author’s description and determine how to categorize it.

Thematic Analysis

Thematic analysis is in essence pattern recognition. Throughout coding and categorizing, the researcher reviews the coded data and thinks about how it is connected, looking for big ideas that permeate the data and links within and across categories. These links, or big ideas, recurring throughout the data are themes that describe an aspect of the construct or experience under investigation and answer the question “How?” Researchers may choose to use a conceptual framework to display the findings in a cohesive, visual representation of both the content and the themes identified. A conceptual framework is how the researcher shares how the data are organized and connected, and what it says about the construct under study.

Limitations

Although there are numerous advantages to using documents as a data source (e.g., cost-effective and efficient, readily available, nonreactive, stable, broad coverage), there are also limitations to be considered when using this research method. First, the documents are produced for some other purpose entirely and may lack sufficient detail to adequately answer the question. Sometimes, this can be overcome through a broad representation in the sample of documents by different authors and written for different audiences (or diversified by other means appropriate to the research question) on the topic of interest. Second, low retrievability of the documents may present a barrier, particularly for some primary types of data that may not be publically available or are deliberately blocked. Fortunately, in the digital age, with many documents electronically stored, organizations and authors who may be archiving such personal documentation may be easily located and contacted, increasing retrievability rates. Finally, documents may be biased, as they are representative of the author's perspective, and the sample as a whole may suffer from selectivity if it represents an incomplete collection or limited selection of the available documents on the topic. Defining clear inclusionary and exclusionary criteria and using a table or database to manage the document sample can help to reduce an overly biased or selective document sample.

Judith M. S. Gross

See also Case Study Method; Conceptual Framework; Content Analysis; Narrative Research; Nvivo; Qualitative Research Methods; Systematic Sampling; Triangulation

Further Readings

Altheide, D. L. (2000). Tracking discourse and qualitative document analysis. Poetics, 27, 287–299.

Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27–40. doi:10.3316/QRJ0902027

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. Thousand Oaks, CA: Sage.

Fitzgerald, T. (2007). Documents and documentary analysis: Reading between the lines. In A. R. Briggs & M. Coleman (Eds.), Research methods in educational leadership and management (2nd ed., pp. 278–294). Thousand Oaks, CA: Sage.

Krippendorf, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.

Merriam, S. (1998). Qualitative research and case study applications in education. San Francisco, CA: Jossey-Bass Publishers.

Kyrsten M. Costlow Kyrsten M. Costlow Costlow, Kyrsten M.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Double-Blind Design Double-blind design

548

549

Double-Blind Design

Double-blind design refers to an experimental methodology with treatment and control groups where neither participants nor researchers, including investigators and outcome assessors alike, know who belongs to the treatment group and who belongs to the control group. This entry describes inconsistencies in blinding terminology, the use of double-blind design in randomized controlled trials, and the importance of the double-blind design in minimizing biases.

Blinding is used in various study designs but is most often associated with randomized controlled trials. Double-blind designs are used to minimize participant and researcher biases, which threaten the validity of a research study. Due to ambiguity in blinding terminology, researchers are encouraged to specifically report which individuals remain blind in a given study.

Blinding terminology varies across researchers, journals, and textbooks, leading to inconsistent definitions. Blinding, also known as masking, describes the process of withholding knowledge of intervention assignments from participants, investigators, or outcome assessors. Definitions of *double blind* and *single blind* disagree on which of these groups of individuals remain blind in each design. Participants, investigators, and assessors typically all remain blind in double-blind designs, but not all studies accord with this definition. To avoid confusion across definitions, researchers can replace basic terminology with specific descriptions of blinding procedures.

Double-blind designs are often associated with randomized controlled trials or studies where participants are randomly assigned to a treatment (intervention) or control (placebo). The placebo effect occurs when an individual’s behavior changes in response to a fake treatment (placebo) simply because that person expects a change. Double-blind designs control for the placebo effect because participants do not know whether they are receiving the treatment or placebo and therefore have equal expectations. Without blinding, participant biases may occur, where participants alter their behavior according to the results expected of their group. In educational research, for example, students who know they are placed in an academic intervention group may work harder to confirm expected academic improvements. Keeping students blind to group membership thus reduces potential participant biases.

Double-blind designs also minimize researcher biases, which occur when researchers (even unconsciously) influence results to confirm their expectations. When researchers allow their expectations to influence participant behavior, this creates self-fulfilling prophecies in participants who may confirm the study's expected results. For example, a teacher in an intervention group who expects his or her students to improve academically may unconsciously provide his or her students with enhanced attention and enthusiasm.

Researcher biases also occur when researchers gather and interpret data in ways that might confirm their expectations. For example, an outcome assessor may look for and exaggerate academic improvements in an intervention classroom compared to a control classroom. Participant and researcher biases threaten internal validity, and double-blind designs can reduce this threat. Blinding procedures can therefore be effective in reducing bias but must be reported clearly in light of inconsistent definitions.

Kyrsten M. Costlow and Marc H. Bornstein

See also Internal Validity; Interviewer Bias; Placebo Effect; Random Assignment; Scientific Method; Selection Bias; Triple-Blind Studies; Validity

Further Readings

Devereaux, P. J., Manns, B. J., Ghali, W. A., Quan, H., Lacchetti, C., Montori, V. M., … … Guyatt, G. H. (2001). Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA, 285(15), 2000–2003.

Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomised trials: Hiding who got what. The Lancet, 359(9307), 696–700.

Alex J. Bowers Alex J. Bowers Bowers, Alex J.

Dropouts

Dropouts

549

550

Dropouts

The term *dropout* refers to a student in school who fails to complete the full course of curriculum and instruction for a degree or diploma; it covers the full spectrum of students who stop attending classes at some point between enrolling in school and their planned graduation. The reasons and process of students dropping out and how schools can help students persist to completion have become a focus of education institutions globally, especially for K–12 schooling and the completion of a high school diploma. Reducing dropout rates is important because studies have shown that students who drop out of school are more likely to experience negative overall life outcomes, such as lower rates of employment and pay, less job security, poorer health, and higher rates of incarceration and unemployment. This entry describes the major issues in students dropping out of high schools in the United States, focusing first on the history and predictors of dropping out, then turning to the question of why students drop out. The entry concludes with recent research that shows that there are at least three different subgroups of students who drop out, each with different reasons and possible interventions.

The History and Predictors of Dropping Out

A student dropping out of high school has long been seen as an issue across K–12 schooling. Students who drop out not only fail to receive instruction and curriculum that the taxpayer has paid for, but students without a high school diploma face tough challenges in the marketplace and in finding high-paying and long-lasting careers. In the early 20th century, the majority of students dropped out of secondary school, or never attended, as only about 20% of all students in the United States graduated with a high school diploma in the early 1920s. During the remainder of that century and into the 21st century, great strides have been made in improving graduation rates, with dropout rates now nationally less than 10% on average. However, there are disparities in dropout rates across the United States, as the national dropout rate hides localized issues of high dropout rates in many urban communities as well as historically underserved communities, such as Hispanic and African American students, as well as students living in urban or rural poverty.

Predictors of which students are most likely to drop out, known as *at-risk predictors*, have traditionally focused on two main areas: demographics and student performance. For demographics, beyond student context factors—such as ethnicity and poverty—research has focused on student family history, showing that children of parents who did not complete high school or who have siblings who also dropped out are more likely to drop out themselves. However, in searching for a means to help students persist in schools, student context factors are mostly beyond the influence of schools. Thus, much of the at-risk prediction literature has focused on student performance in the schooling system to attempt to identify factors that are associated early with student likelihood of dropping out at a later date. These performance factors include low or failing grades in middle school or ninth grade, high absences, and multiple discipline reports, suspensions, or expulsions. Additionally, the practice of retaining students after course failure in grade for a second year is well-known to predict student dropout, as the majority of students retained in high school do not earn their high school diploma.

Why Do Students Drop Out?

The research on why students drop out focuses on two main areas: voluntary dropout and discharge. In the voluntary dropout research, researchers focus on what is termed the *life course perspective*. A student’s decision to drop out of school is typically not based on a single incident, issue, or period of time, but rather is based on a student’s cumulative experience in school. The life course perspective posits that while students are overall resilient and will continue with their schooling over many years, by the time they reach high school, they may have encountered multiple course failures, disproportionate disciplinary policies, family strife, or a multitude of these factors. Cumulative effects build over time, and thus by high school, a student may drop out. Alternatively, from the discharge perspective, some schools may disproportionally push students out who may have low test scores or discipline issues. Researchers have termed the schools in which a large number or majority of the students drop out as “dropout factories,” many of which are in high-need neighborhoods, identifying the schools nationally by name with the intent to focus resources and professional development on the needs of the schools, teachers, and administrators so they can help find ways for students to persist to high school graduation.

Interventions

To date, while there is some preliminary positive evidence on interventions to help students persist in school who are likely to drop out, dropout intervention studies are known to have high variability. Recent research has found that there are at least three very different types of students who drop out and that the differences between these subgroups lead to differential effects in the intervention literature. The three types of dropout groups are the jaded, the quiet, and the involved.

The jaded group includes only one third of all dropouts but represents the popular stereotype of a student who does not like school and performs poorly. These students are typified by decreasing grades and require interventions that provide mentorship to them and link them back to the purpose of schooling. Quiet dropouts are the majority, and while they like school, their teachers, and their friends, their grades are low yet increasing over time, but not increasing fast enough for these students to pass their courses and graduate. These students require academic tutoring to help reduce their dropout rate. The final group, the involved, makes up only about 9% of all dropouts. Members of this group tend to drop out late in the high school process due to either a life event, such as family divorce or pregnancy, or a mistake in their transcript. These students need administrative help with making sure they have the appropriate courses and credits to receive a diploma.

Alex J. Bowers

See also Adolescence; Motivation; State Standards

Further Readings

Bowers, A. J., & Sprott, R. (2012). Why tenth graders fail to finish high school:

A dropout typology latent class analysis. Journal of Education for Students Placed at Risk, 17(3), 129–148.

Dupéré, V., Leventhal, T., Dion, E., Crosnoe, R., Archambault, I., & Janosz, M. (2015). Stressors and turning points in high school and dropout: A stress process, life course framework. Review of Educational Research, 85(4), 591–629.

Rumberger, R. W. (2011). Dropping out: Why students drop out of high school and what can be done about it. Cambridge, MA: Harvard University Press.

Brandon LeBeau Brandon LeBeau LeBeau, Brandon

Dummy Variables

Dummy variables

551

552

Dummy Variables

Dummy variables, sometimes referred to as indicator variables, are a common data preparation step to represent categorical (or qualitative) variables as a series of dichotomous (i.e., 0/1) variables. This technique is useful to recreate an analysis of variance model in a regression framework, which is achieved by creating $c - 1$ new dichotomous variables from a categorical variable, where c represents the number of groups, categories, or levels of the original categorical variable. For example, a variable representing a high school graduated student (i.e., graduated *vs*. did not graduate) was created by assigning a value of 1 if the student graduated from high school or 0 otherwise. This entry explores in more detail the creation, interpretation, and reasons for using dummy variables.

Creating Dummy Variables

Creating dummy variables is an important data preparation step that is mostly used for fitting linear regression models; however, it is also useful for graphical or tabular displays. When creating dummy variables for tables or figures, it is helpful to create dummy variables for all the categories of the original variable.

For example, suppose the grade level of eight students were collected. This variable could be represented as the grade level each student is currently in (as shown in the left most column of the matrix shown in Figure 1). These eight students were in Grades 7, 8, or 9. The grade level of a student is represented by an integer; however, you could argue that the variable is only ordinal in nature. This suggests that the differences between the values on the scale are not consistent; for example, the difference (growth) between seventh and eighth grade is not the same as between eighth and ninth grade. In these situations, dummy variables offer an alternate representation of the data.

Figure 1 Matrix of dummy variables, showing categorical variables

$$\begin{bmatrix} Grade \\ 7 \\ 7 \\ 8 \\ 9 \\ 7 \\ 9 \\ 8 \\ 7 \end{bmatrix} = \begin{bmatrix} Grade_7 & Grade_8 & Grade_9 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

The dummy variables created from the grade level of the eight students are shown in the right matrix of Figure 1, labeled as Grade$_7$, Grade$_8$, and Grade$_9$, representing three variables for Grades 7, 8, and 9, respectively. As can be seen in Figure 1, to create the Grade$_7$ variable, any students who were recorded to be in Grade 7 in the left side of the equation are now represented by a value of 1 in the right side of the equation, whereas any other grade is represented with a 0. Similar logic was used to create the Grade$_8$ and Grade$_9$ variables. Dummy variables are also referred to as indicator variables, as these new variables in the right matrix indicate the group (i.e., grade) the student belongs to.

When creating dummy variables for regression models, only $c - 1$ new dichotomous variables are needed when an intercept is included in the linear model. (In fact, because of the perfect relationship created using all categories as predictors, software will have difficulty calculating the statistics). This is shown in Figure 2, where only the variables Grade$_7$ and Grade$_8$ were created. The dummy variable not coded (i.e., Grade$_9$ from the previous matrix) is referred to as the reference group. From a mathematical perspective, it does not matter which level of the categorical variable is used as the reference group; instead, the decision regarding the level to be used as the reference group is driven by the research question of interest. More information on this will be given in the following section on interpreting dummy variables.

Figure 2 Matrix of dummy variables, showing dichotomous variables

$$\begin{bmatrix} \mathit{Grade} \\ 7 \\ 7 \\ 8 \\ 9 \\ 7 \\ 9 \\ 8 \\ 7 \end{bmatrix} = \begin{bmatrix} \mathit{Grade}_7 & \mathit{Grade}_8 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Interpreting Dummy Variables

Dummy variables entered into regression models are interpreted as mean adjustments from the reference group. A researcher exploring whether the average time spent studying differs for the three grades can lead to the following linear model:

$$Y_j = \beta_0 + \beta_1 \mathrm{Grade}_{7j} + \beta_2 \mathrm{Grade}_{8j} + \in_j .$$

In Equation 1, Y_j is the response variable (e.g., hours studying); β_0, β_1, and β_2 are regression coefficients for the intercept, Grade_{7j} and Grade_{8j}, respectively; and $\in_j$ is error. The regression coefficient for the intercept, β_0, represents the average value of Y_j, given that the covariates equal 0. In this example, this occurs when the variables Grade_{7j} and Grade_{8j} both equal 0. Therefore, β_0 in Equation 1 would represent the average number of hours studying for ninth graders. Regression coefficients for continuous predictors are interpreted as the average change in Y_j for a one-unit change in the covariate. As the variable Grade_{7j} only takes values of 0 or 1, β_1 would represent the mean adjustment needed to move from Grade 9 (the reference group) to Grade 7. A similar interpretation for β_2 would be made. The average time each grade spent studying could then be found from the following equations:

$$\mu_{\text{grade}7} = \beta_0 + \beta_1 . \mu_{\text{grade}8} = \beta_0 + \beta_2 . \mu_{\text{grade}9} = \beta_0 .$$

Negative estimates for β_1 or β_2 would represent means for Grade 7 or 8 being less than Grade 9, whereas positive estimates for β_1 or β_2 would represent means for Grade 7 or 8 being larger than Grade 9.

There are alternative ways to code dummy variables in addition to that shown in the previous examples. The most common alternatives are referred to as *effects coding* and *contrast coding*. Melissa Hardy provides a discussion about these two alternative dummy coding schemes in her book *Regression With Dummy Variables*.

Limitations

There are two concerns to keep in mind when using dummy variables in regression models. First, always use $c - 1$ dummy variables to represent a categorical variable, where c is the number of categories or levels. As John Fox discusses in his text *Applied Regression Analysis and Generalized Linear Models*, if all c dummy variables are included in the regression model along with an intercept, the model will be overparameterized and unique estimates cannot be found. This problem can be corrected using $c - 1$ dummy variables or not including an intercept in the model. For most conditions, using $c - 1$ dummy variables is recommended. Second, standardized regression coefficients for dummy variables are not directly interpretable due to their dichotomous nature.

Instead if a standardized metric is desired, it is better to standardize the response variable and quantitative covariates and leave the dummy variables in their dichotomous form.

Brandon LeBeau

See also Analysis of Variance; Data; Levels of Measurement; Multiple Linear Regression; Simple Linear Regression

Further Readings

Hardy, M. A. (1993). Regression with dummy variables. Sage Publications.

Fox, J. (2015). Applied regression analysis and generalized linear models. Sage Publications.

Suits, D. B. (1957). Use of dummy variables in regression equations. Journal of the American Statistical Association, 52(280), 548–551.

Thomas Ledermann Thomas Ledermann Ledermann, Thomas

Robert A. Ackerman Robert A. Ackerman Ackerman, Robert A.

Dyadic Data Analysis Dyadic data analysis

553

555

Dyadic Data Analysis

Dyadic data analysis refers to the analysis of data from pairs of people, called dyads, using statistical methods. Typical examples of dyads include romantic couples and twins. The link between two dyad members can be interactive (such as between a tutor and a student), genetic (such as between two siblings), experimental (such as when two persons are paired in terms of certain characteristics), or yoked (such as when two persons are exposed to the same external influences).

Dyadic data analysis is often very different from the analysis of individual data. Various models have been developed that allow researchers to address a wide range of questions, including the assessment of associations between variables, the analysis of similarity effects, and the study of change at the individual and dyadic levels. This entry looks at the different models and how they are used and also discusses concepts important to dyadic data analysis.

An important distinction with implications for analyses is whether dyad members are distinguishable or indistinguishable. Members are distinguishable when there is a variable that enables a meaningful classification of the dyad members into two different groups (categories), such as gender in heterosexual couples or family role in mother–daughter dyads. Dyad members are indistinguishable, sometimes called exchangeable, if there is no such distinguishing variable. Same-sex twins or lesbian couples are typical examples of indistinguishable members.

Nonindependence

The concept of nonindependence is the most fundamental to dyadic analysis. Nonindependence occurs when the scores of two dyad members are statistically related. In heterosexual couples, for example, husbands and wives are typically similar in many respects, including education, attitudes, and personality characteristics. Nonindependence can be positive, reflecting similarity between dyad members, or negative, reflecting dissimilarity. Negative nonindependence occurs less frequently but can be expected if there is compensation within dyads (e.g., the more I do, the less my partner needs to do) or competition between members (e.g., the happier I am with how I performed, the less happy my partner is with his or her performance).

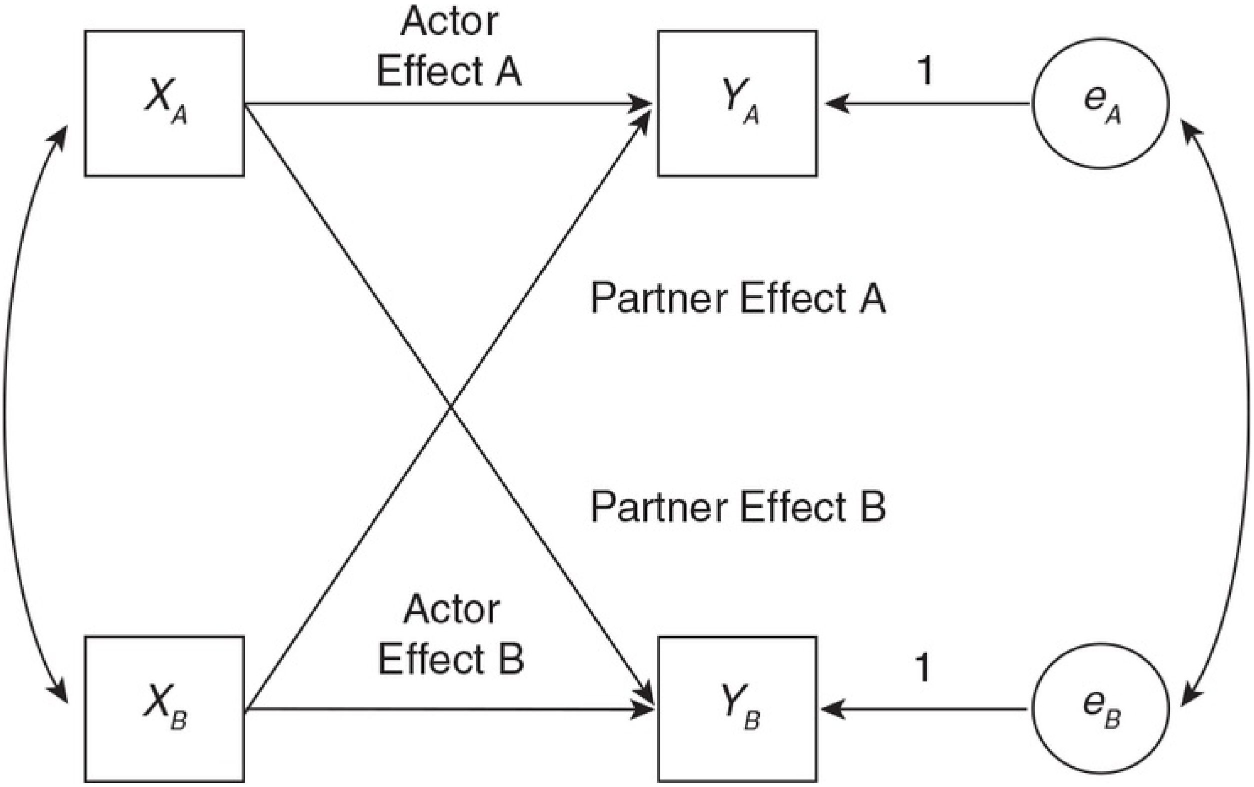
Nonindependence can be assessed when members are distinguishable or indistinguishable. When members are distinguishable and the variable measured in both members is continuous, the Pearson correlation between the members' scores is often used to measure nonindependence. When members are indistinguishable, nonindependence can be assessed by the intraclass correlation coefficient, which can be calculated using one-way analysis of variance or multilevel modeling (MLM), or the pairwise correlation, which requires a data structure known as pairwise or double-entry structure.

Dyadic Models

The dyadic data are analyzed using specific techniques because the group size is only two. In the last 3 decades, a wide range of models have been developed to study dyads, with two models dominating the field: the actor–partner interdependence model (APIM) and the dyadic growth curve model. Importantly, most dyadic models require that the same set of variables is measured in both members.

Figure 1 displays the APIM for two variables, *X* and *Y*, both measured in member *A* and member *B*, which might be caregiver and patient. This model allows researchers to predict a person's outcome by the person's own predictor and the partner's predictor. The path from the person's predictor *X* to that same person's outcome *Y* is called the actor effect and the path from the partner's predictor *X* to the person's outcome *Y* is called the partner effect. With distinguishable members, there are two actor effects and two partner effects, one for each type of dyad member. With indistinguishable members, there is only one actor effect and one partner effect.

Figure 1 The Actor–Partner Interdependence Model for member A and member B

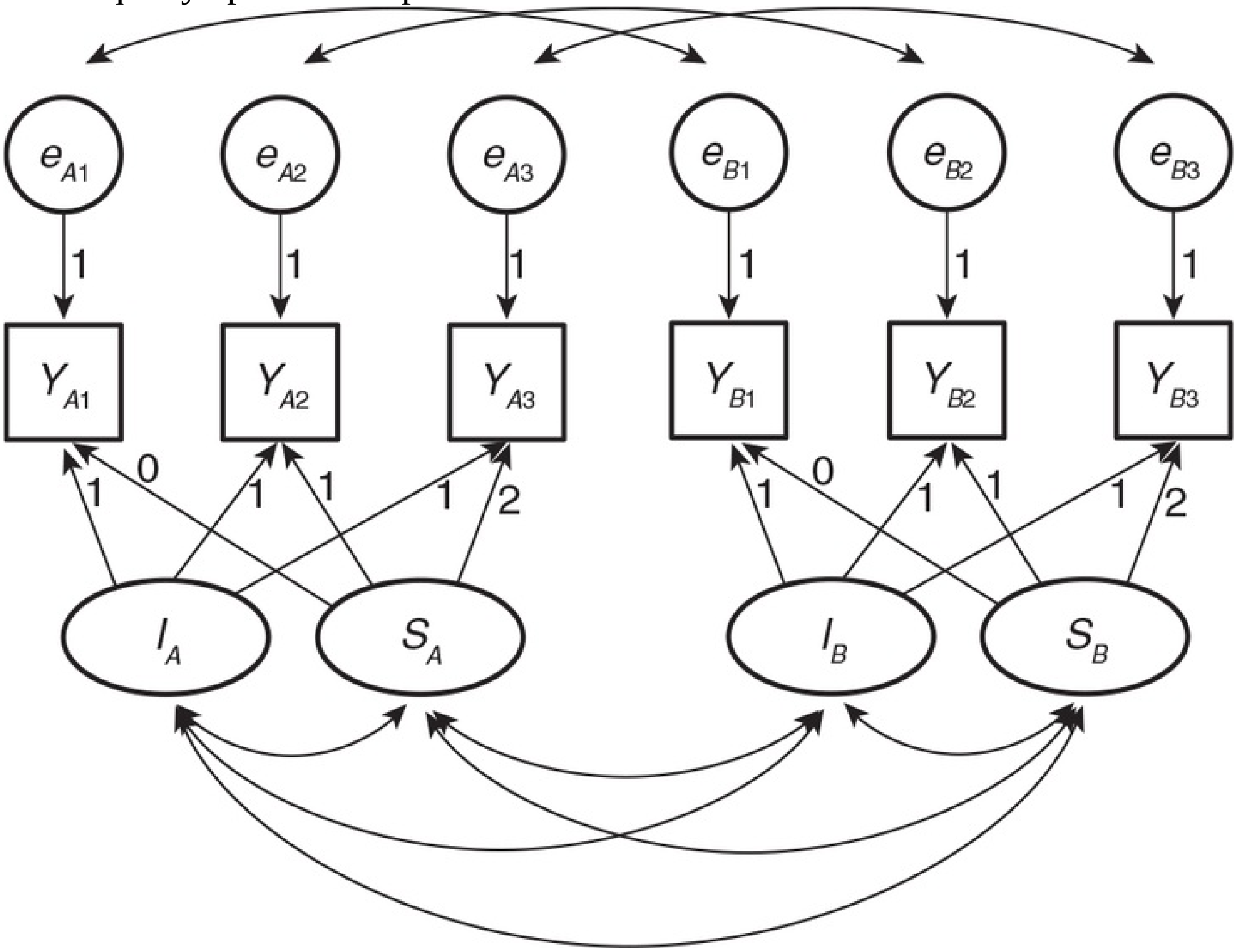


As indicated by the double-headed arrows, there is a correlation between the members' predictors and the members' residuals e_A and e_B. The APIM can also be used when *X* and *Y* are the same variable measured at two time points. In this case, the actor effects represent stability over time and the partner effects represent the influence between the dyad members.

An application specific to the APIM is the study of similarity effects in dyads (e.g., does spouse similarity in attitudes predict spouses' satisfaction?) and buffering and enhancing effects (e.g., is in couples the positive effect of one's own agreeableness on one's own satisfaction enhanced by having an agreeable partner?). An advantage of the APIM over other dyadic models is that it can be analyzed using regression analysis, MLM, or structural equation modeling (SEM) when members are distinguishable. Indistinguishable members require the use of either MLM or SEM.

The dyadic growth curve model is a natural extension of the growth curve model for individual data and allows researchers to study change at the level of the dyad members. Figure 2 shows the dyadic GCM for members *A* and *B* for a single variable *Y* measured at three time points. In this model, a person's score is a function of the intercept *I*, the linear slope *S*, and a time-specific error *e*. The intercept is where dyad members start on average at the point that is fixed to zero, whereas the slope is the average rate of linear change.

Figure 2 The Dyadic Growth Curve Model for member A and member B and three equally-spaced time points



Both the intercept and the slope factors have a variance, and as indicated by the double-headed arrows, these factors are correlated both within and between members. With distinguishable members, there is an intercept and slope for each type of member, whereas with indistinguishable members, there is only one intercept and one slope. There is also a correlation between the errors of the two dyad members' scores measured at the same time. The dyadic growth curve model can be estimated by both SEM and MLM.

Another important model in dyadic research is the common fate model, which assumes that each member of a dyad is affected by a shared (common) influence, such as the living environment in couples or the working conditions in coworker dyads. The dyad members’ scores serve as indicators of a latent variable reflecting the shared influence. This model has been designed to assess associations at the dyadic level and recently extended to model systematic change at the dyadic (group) level. Because the common fate model is a latent variable model, it requires the use of SEM or multilevel SEM.

Thomas Ledermann and Robert A. Ackerman

See also Analysis of Variance; Path Analysis; Pearson Correlation Coefficient; Social Network Analysis; Structural Equation Modeling

Further Readings

Card, N. A., Selig, J. P., & Little, T. (2011). Modeling dyadic and interdependent data in the developmental and behavioral sciences. New York, NY: Routledge.

Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. Psychological Bulletin, 118, 430–439. doi:10.1037/0033–2909.118.3.430

Kashy, D. A., Donnellan, M. B., Burt, S. A., & McGue, M. (2008). Growth curve models for indistinguishable dyads using multilevel modeling and structural equation modeling: The case of adolescent twins’ conflict with their mothers. Developmental Psychology, 44, 316–329. doi:10.1037/0012–1649.44.2.316

Kenny, D. A., & Cook, W. L. (1999). Partner effects in relationship research: Conceptual issues, analytic difficulties, and illustrations. Personal Relationships, 6, 433–448. doi:10.1111/j.1475–6811.1999.tb00202.x

Ledermann, T., & Kenny, D. A. (2017). Analyzing dyadic data with multilevel modeling versus structural equation modeling: A tale of two methods. Journal of Family Psychology, 31(4), 442–452.

Roland H. Good Roland H. Good Good, Roland H.

Dynamic Indicators of Basic Early Literacy Skills Dynamic indicators of basic early literacy skills

555

558

Dynamic Indicators of Basic Early Literacy Skills

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is a set of standardized measures across kindergarten through sixth grade that can be used to assess students' early reading skills and are sensitive to student growth. Roland H. Good III and Ruth A. Kaminski are the primary researchers who created the original set of measures. The revised version is titled DIBELS Next. The DIBELS measures and early research are based on Stan Deno's work in the area of curriculum-based measures at the University of Minnesota. Like curriculum-based measures, DIBELS assessments are short, repeatable measures with many alternate forms that can be completed in 10–15 minutes and are valid and reliable. The DIBELS measures extend curriculum-based measures to kindergarten and early first grade and provide generic passages that are not curriculum-specific. The skills assessed by DIBELS change at each grade level and time of year across a developmental continuum that reflects the expected acquisition of phonemic awareness, basic phonics and the alphabetic principle, accurate and fluent reading of connected text, and oral and silent reading comprehension skills. Each general skill indicator is predictive of future reading success and provides a target for intervention or instruction. The purpose of the DIBELS measures is to inform educational decisions within an outcomes-driven model (ODM). This entry reviews the basics of DIBELS Next, lists its continuum of skills, shows how it can be used in an ODM, and lists additional DIBELS measures.

Basics About DIBELS Next

DIBELS Next, the revised version of the DIBELS assessments, is a tracking system for educators. The vision of DIBELS is to create a road map for teachers and districts that tracks how students move from being nonreaders to becoming skilled readers. The goal of DIBELS is to enable teachers to know, at any given time, their students' current reading skills and where their skills need to be for adequate progress. They also have information on the likely level of support needed for the student to be successful.

The DIBELS assessments are designed to be quick and efficient because every minute spent assessing a student is a minute in which instruction or intervention is not occurring. Therefore, the DIBELS measures are not intended to provide comprehensive, in-depth information on every component of proficient reading. Instead, they are designed to be indicators or critical components. An example of a common indicator is a mile marker on a highway that marks distance and indicates past distance traveled and further distance needed to reach the intended destination. Looking at the mile marker takes little time, but it signals progress toward the goal and can be a predictor of how long it will take to reach the destination.

Similar to the mile marker, DIBELS measures are quick indicators that give a "snapshot" of current reading skills within a developmental sequence of related reading skills. The point at which a student is making minimally acceptable progress is called a benchmark goal. Benchmark goals are based on research that examined the longitudinal predicative validity of a score at a particular point in time. The changing benchmarks across grades provide a continuum that links student performance on earlier skills with later skills. When a benchmark is attained at the designated time, the student is likely (generally 80–90% likelihood) to achieve the next benchmark at that designated time. Within each grade level, benchmarks are set for individual DIBELS components as well as for the DIBELS Next Composite score. The DIBELS Next Composite consists of all DIBELS components and represents overall reading proficiency, including reading at an adequate rate, with a high degree of accuracy, and for meaning.

Critical to any assessment is validity and reliability. Validity means that the assessment is measuring the essential early literacy skills that it is intended to measure. Content validity indicates that the content of the assessment adequately represents critical skill areas. In 2000, the National Reading Panel produced a report that synthesized reading research and identified the essential reading and early literacy skills as phonemic awareness, alphabetic principal and phonics, fluent reading of connected text, reading comprehension, and fluent reading of connected text, reading comprehension, and vocabulary/language skills. The DIBELS Next measures were designed specifically to be linked to these critical early literacy skills. Criterion-related validity is the extent to which a student's performance on a criterion measure can be estimated from that student's performance on the assessment being validated. DIBELS Next was compared to the Group Reading Assessment and Diagnostic Evaluation. As reported in the DIBELS Next technical manual, in first through sixth grades, all validity coefficients were in the strong range for both predictive and concurrent criterion-related validities (.55 to .91) and in the moderate-to-strong range for kindergarten (.40 to .70).

DIBELS Next reports three different types of reliability: alternate form, test–retest, and interrater. Alternate-form reliability indicates the extent to which different forms of the same measure are related. Test–retest reliability specifies the degree to which students' results are stable when the same test form is administered twice within a short interval. Interrater reliability indicates the extent to which different assessors collect and score data in the same manner. A minimum reliability of .80 is required for making screening decisions, and a minimum reliability of .90 is required for important educational decisions concerning an individual student. For the DIBELS Next composite score, alternate-form reliability ranged from .66 to .98, test–retest reliability ranged from .81 to .94, and interrater reliability ranged from .97 to .99.

DIBELS Next Continuum of Skills

The skills assessed in DIBELS change over time, as students develop from nonreaders to skilled readers. In early kindergarten, DIBELS Next includes First Sound Fluency, which measures early phonemic awareness, and Letter Naming Fluency, which measures alphabet knowledge. In later kindergarten and early first grade, DIBELS Next includes Letter Naming Fluency, Phoneme Segmentation Fluency (phonemic awareness), and Nonsense Word Fluency (alphabetic principal and basic phonics). In the later part of first grade, Oral Reading Fluency with Retell (accurate fluent reading of connected text and reading comprehension) is added. At the second-grade level, DIBELS Next includes Nonsense Word Fluency and Oral Reading Fluency with Retell, and in third through sixth grades, DIBELS Next includes Oral Reading Fluency with Retell and a maze measure called Daze (fluency and comprehension). Benchmark assessments are collected for all students at the beginning, middle, and end of the academic year. Progress monitoring assessments are available to more frequently assess targeted students who are identified as likely to need additional support.

DIBELS Next in an ODM

An educational decision model provides a set of steps for assessment to inform instruction to improve student outcomes. ODM is a data-driven decision-making model that was first developed by Ruth Kaminski and Roland Good. The ODM is based on the earlier problem-solving model, but with an emphasis on early intervention and prevention. DIBELS was developed to inform educational decisions within the ODM and can be used to inform educational decisions within similar models. The ODM is similar to response-to-intervention, response-to-instruction, multitiered systems of support, and problem-solving models. These models share common features that include (a) providing generally effective reading instruction for all, (b) universal screening to identify students who may be at risk for reading difficulties, (c) targeting of specific students and skills for additional instructional support, (d) frequent monitoring of students' progress during intervention and instruction, (e) modifying instruction and intervention based on student progress, and (e) schools and districts examining their overall, system-wide effectiveness in implementing a system of instructional supports.

Within the ODM, teachers first screen students with the grade-level DIBELS benchmark assessment. Teachers then identify students who may need additional instructional support by evaluating students' current level of academic skills with the level of skills that predict attaining future benchmark goals. During this step, it is helpful to have comprehensive reports that summarize class-wide performance and provide a recommended level of instructional support. DIBELS Next reports of students' skills acquisition and likely need for instructional support are available through Dynamic Measurement Group's DIBELSnet, Voyager Sopris Learning VPORT, and Amplify mCLASS. For students who may need additional support, the next step is to validate that need for support. Alternate forms of all DIBELS assessments are available to retest the student on a different day or under different conditions to be reasonably confident in educational decisions. If there are discrepancies in student performance, the DIBELS assessment is validated to ensure confidence in the accuracy of data.

In the next step, teachers use DIBELS data to plan and implement support. If a student achieves a score below the benchmark on the DIBELS composite score, the student is likely to need additional instructional support to attain subsequent goals. DIBELS is unique in offering Pathways of progress to assist educators in establishing individual student learning goals. Pathways of progress are based on student progress percentiles and are designed to assist educators in (a) setting an ambitious, meaningful, attainable goal, (b) creating an aim line for individual, grade-level progress monitoring, and (c) evaluating progress for individual students. Pathways of progress provide a normative reference that is based on the reading progress of one student relative to other students with similar initial skills.

After instructional support has been implemented for a student, the next step in the ODM is to evaluate and modify the support as needed for the student to make adequate progress. There are at least 20 alternate forms of each DIBELS assessment, each taking only minutes to administer, to enable teachers to monitor student growth, and modify support as needed. As a final step in the ODM, educators review the procedures and outcomes for the school or district as a system to evaluate system-wide effectiveness in providing instructional supports.

Additional DIBELS Measures

In addition to K–6 DIBELS Next assessments, Dynamic Measurement Group has developed additional DIBELS measures. They include Preschool Early Literacy Indicator by Kaminski and Mary I. Abbott, Math (K–6th) by Courtney Wheeler and Good, Content Area Reading Indicators by Abbott and Good, and In-Depth Assessment of Literacy (known as DEEP) by Kelly A. Powell-Smith, Kaminski, and Good. The theoretical underpinnings and procedures that were used to develop DIBELS Next were used during the research and development of each of these products.

Roland H. Good

See also Benchmark; Curriculum-Based Measurement; Outcomes; Problem Solving; Progress Monitoring; Response to Intervention

Further Readings

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. Exceptional Children, 52, 219–232.

Dewey, E. N., Powell-Smith, K. A., Good, R. H., & Kaminski, R. A. (2015). DIBELS Next technical adequacy brief. Eugene, OR: Dynamic Measurement Group. Retrieved from http://dibels.org/papers/DIBELSNextTechnicalAdequacy.pdf

Good, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Petersen, K., Powell-Smith, K., & Wallin, J. (2010). Dynamic indicators of basic early literacy skills next. Longmont, CO: Sopris. Retrieved from http://dibels.org/

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2013). DIBELS Next Technical Manual. Eugene, OR: Dynamic Measurement Group. Retrieved from http://DIBELS.org/next

Good, R. H., Powell-Smith, K. A., & Dewey, E. N. (2013). DIBELS pathways of progress: Setting ambitious, meaningful, and attainable goals in grade level material. Eugene, OR: Dynamic Measurement Group. Retrieved from http://dibels.org/papers/Pathways_Handouts_PCRC2013.pdf

Kaminski, R. A., Cummings, K. D., Powell-Smith, K. A., & Good, R. H. (2008). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology V (pp. 1181–1204). Bethesda, MD: National Association of School Psychologists.

National Reading Panel. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups. Bethesda, MD: National Institute of Child Health and Human Development.

Kathleen H. Nielsen Kathleen H. Nielsen Nielsen, Kathleen H.

Dysgraphia

Dysgraphia

558

558

Dysgraphia

Dysgraphia is a brain-based written language disability defined by difficulty in finding letters in memory, retrieving them, and writing them legibly and automatically. Some think handwriting is just a motor skill, but it also depends on orthographic coding, serial finger movements, and integrating orthographic coding with serial finger movements. Orthographic coding is seeing letters alone or in written words in the “mind’s eye.”

Dysgraphia occurs in individuals whose cognitive abilities are in the normal range and even above average. Dysgraphia is a disorder in letter-level writing, which in turn may affect the fluency and quality of written work at the word, sentence, and paragraph levels. It may present as inconsistency in letter formation, inconsistent use of uppercase and lowercase letters, difficulty organizing written work on the page, and inability to put together coherent written sentences and text. Writing is a laborious process for someone with dysgraphia that can cause fatigue and discomfort of the hand. Dysgraphia can occur alone or in conjunction with dyslexia or other learning disabilities. It emerges in early childhood, but educators do not always screen and intervene to prevent it or reduce its severity or diagnose it.

Brain imaging studies show differences in the structural and functional connectivity of the brains of children with developmental dysgraphia compared to those who are typical written language learners. Effective instruction teaches a plan for consistent serial stroke production, coding letters into memory and retrieving them from memory, and transfer to spelling and composing. Individuals often need accommodations such as more time to complete written assignments and using technology tools, but they also need explicit instruction in assignments and using technology tools, but they also need explicit instruction in using technology tools including touch typing.

Kathleen H. Nielsen

See also Dyslexia; Learning Disabilities; Literacy; Special Education Identification; Special Education Law

Further Readings

Chung, P., & Dilip, P. (2015). Dysgraphia. International Journal of Child and Adolescent Health, 8, 27–36.

Crouch, A. L., & Jakubecy, J. J. (2007). Dysgraphia: How it affects a student’s performance and what can be done about it. Exceptional Children Plus, 3, 1–13.

James, K., Jao, J. R., & Berninger, V. (2015). The development of multi-leveled writing systems of the brain: Brain lessons for writing instruction. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), Handbook of writing research (pp. 116–129). New York, NY: Guilford.

Richards, T. L, Grabowksi, T., Askren, K., Boord, P., Yagle, K., Mestre, Z., & Berninger, V. (2015). Contrasting brain patterns of writing-related DTI parameters, fMRI connectivity, and DTI-fMRI connectivity correlations in children with and without dysgraphia or dyslexia. Neuroimage Clinical. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26106566

Kathleen H. Nielsen Kathleen H. Nielsen Nielsen, Kathleen H.

Dyslexia

Dyslexia

558

560

Dyslexia

Dyslexia is a brain-based language-learning disability affecting 15%–20% of the population. Reading and spelling skills are below population mean and expected levels based on verbal reasoning. Research has shown a genetic basis for dyslexia with multiple associated genes and often familial history. Differences in brain patterns between normal readers and those with dyslexia have been shown in structural and functional magnetic resonance imaging and studies examining the connectivity of brain regions.

Historically, this learning disability identified in the 1800s was called *word blindness*. Characterized by poor underlying phonological processing of heard and spoken words and orthographic processing of read and written words, dyslexia interferes with an individual’s ability to convert written words into spoken words (decoding) and spoken words into written words (spelling). Executive functions and working memory related to language are related deficits.

Most individuals show average or better expressive language and reading comprehension abilities despite difficulties with word reading and spelling. Dyslexia may present differently in each affected individual, although common characteristics are trouble with decoding unfamiliar words, which in turn affects rate and fluency of reading, and spelling, which in turn affects rate and fluency of composing. Some also have handwriting problems. The severity of these characteristics can vary among individuals.

Although dyslexia is thought of as a childhood disorder, it can present throughout the life span. Students may need accommodations and interventions from elementary school through college and beyond. Spelling issues tend to persist throughout the life span.

Dyslexia can be treated through explicit, systematic teaching at all levels of language: subword, word, and syntactic and text levels. Phonological, orthographic, and morphological awareness needs to be addressed. Research has shown brain changes and normalization of reading and/or spelling in children with dyslexia who undergo specialized reading and/or writing intervention. Individuals with dyslexia can exhibit different strengths and weaknesses in reading and spelling skills and often need individualized intervention to remediate their deficits. Some are twice exceptional, and their superior cognitive capabilities may mask their dyslexia unless they are carefully assessed.

Kathleen H. Nielsen

See also Dysgraphia; Giftedness; Learning Disabilities; Literacy; Special Education Identification; Special Education Law

Further Readings

Berninger, V., Nielsen, K., Abbott, R., Wijsman, E., & Raskind, W. (2008). Writing problems in developmental dyslexia: Under-recognized and under-treated. Journal of School Psychology, 46, 1–21.

Berninger, V., & Richards, T. (2010). Inter-relationships among behavioral markers, genes, brain, and treatment in dyslexia and dysgraphia. Future Neurology, 5, 597–617.

Brkanac, Z., Chapman, C., Igo, R., Matsushita, M., Nielsen, K., Berninger, V., Wijsman, E., & Raskind, W. (2008). Genome scan of a nonword repetition phenotype in families with dyslexia: Evidence for multiple loci. Behavioral Genetics, 38, 462–475.

Shaywitz, S., Morris, R., & Shaywitz, B. (2008). The education of dyslexic children from childhood to young adulthood. Annual Review of Psychology, 59, 451–476.

E

Boaz Shulruf Boaz Shulruf Shulruf, Boaz

Ebel Method

Ebel method

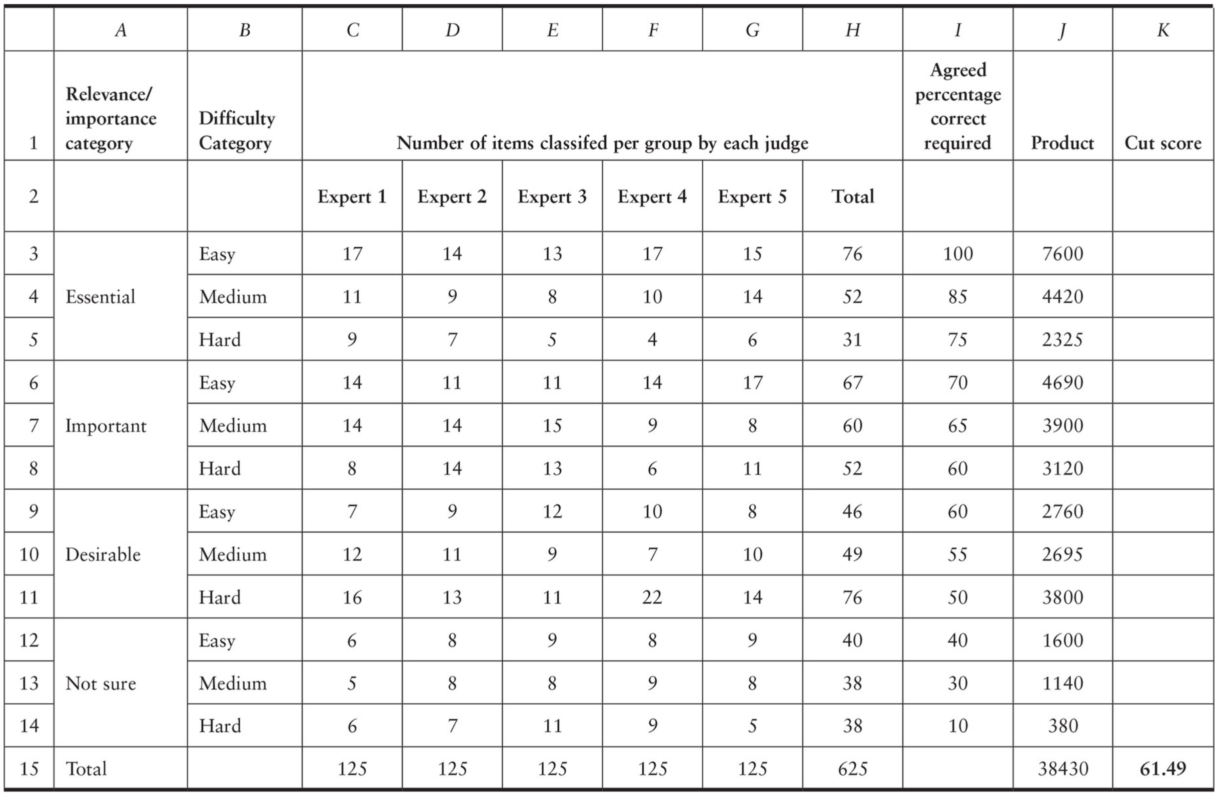
561

563

Ebel Method

The Ebel method is a standard setting method normally used to determine a cut score for multiple-choice question types of tests. The Ebel method has been used for setting standards for examinations within the fields of higher education and medical and health professions and for applicant selection decision-making.

The Ebel method involves a panel of experts who classify each item by two criteria: (1) level of difficulty (e.g., easy, medium, hard) and (2) relevance or importance (e.g., essential, important, desirable, unsure). Then, the panel reaches a consensus regarding the expected percentage of items that should be answered correctly for each group of items, classified by both difficulty and relevance/importance. To determine the cut score for the test, the total number of items in each group is multiplied by the required percentage of correct answers; then the sum of all the groups is divided by the number of items multiplied by the number of panelists. The hypothetical example shown in Table 1 demonstrates how that works.



	A	B	C	D	E	F	G	H	I	J	K
1	Relevance/ importance category	Difficulty Category	Number of items classifed per group by each judge						Agreed percentage correct required	Product	Cut score
2			Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Total			
3	Essential	Easy	17	14	13	17	15	76	100	7600	
4		Medium	11	9	8	10	14	52	85	4420	
5		Hard	9	7	5	4	6	31	75	2325	
6	Important	Easy	14	11	11	14	17	67	70	4690	
7		Medium	14	14	15	9	8	60	65	3900	
8		Hard	8	14	13	6	11	52	60	3120	
9	Desirable	Easy	7	9	12	10	8	46	60	2760	
10		Medium	12	11	9	7	10	49	55	2695	
11		Hard	16	13	11	22	14	76	50	3800	
12	Not sure	Easy	6	8	9	8	9	40	40	1600	
13		Medium	5	8	8	9	8	38	30	1140	
14		Hard	6	7	11	9	5	38	10	380	
15	Total		125	125	125	125	125	625		38430	**61.49**

In Table 1, a panel of five experts applies the Ebel method to an examination consisting of 125 items.

1. Expert 1 classifies each item to one of the 12 groups; the number of items classified to each group for Expert 1 is shown in column C;
2. Expert 2 does the same for column D and so forth with all other experts. The sum of items in each group across all experts is placed in column H.
3. The agreed percentage correct required per group is placed in column I. Column J is the product of multiplication of H × I.
4. The total number of items classified is the sum of column H (H15).
5. The total sum of product J is J15.
6. The calculated cut score is J15/H15, that is, 38,430/625 = 61.49.

Modifications

A few modifications have been suggested to the Ebel method. For example, instead of reaching a consensus on the percentage correct required, it is possible to ask each expert about his or her preferred percentage correct required and then calculate the average across all experts. Alternatively, the percentage correct required could be determined by policy rather than by experts’ advice. Another modification is to classify the degree of difficulty according to the probability of items responded to correctly, for example, <0.5 hard, 0.5–0.7 medium, and >0.7 easy. Ernest Skakun and Samuel Kling suggested adding taxonomy criteria to the Ebel method.

Strengths and Weaknesses

Strengths

The Ebel method requires experts to estimate the difficulty and the relevance of items. This replicates the process that the item/examiner writers would normally go through—trying to identify items that are relevant and at the right pitch. As such, the method adds an additional perspective to the same process of examination design. For the experts, there is no need to imagine the hypothetical borderline examinee, which may vary significantly across experts. They all relate to the same group of examinees and the same curriculum, hence the reference for the judgment is similar to all.

The unique feature in Ebel methods is the judgment of relevance/importance. This additional classification distinguishes essential and nonessential competencies. Consequently, although the estimated cut score is for the entire examination, it is influenced by the relevance of the items. Other standard setting methods do not consider the relevance/importance of the test items, which at times may create a gap between the difficulty of an examination and its relevance to the curriculum. A meta-analysis that compared a number of standard setting methods with the Angoff method identified only a small difference between the cut scores derived from Ebel and Angoff methods.

Weaknesses

The Ebel method, although not requiring sophisticated psychometric skills, is a complex process for both experts and those who manage the process. It is a complex concept to convey to students and therefore may not be well received. The method has also been criticized because of the difficulty for experts to use bidimensional judgment (difficulty and importance/relevance). The most problematic critique however is related to the rationale underlying the classification of “not sure” or “questionable.” Since the category of acceptable is the next in order, it means that the lowest classification is not acceptable. As the next in order, it means that the lowest classification is not acceptable. As such, items classified in this category should not be included in the examination if they are not acceptable. This is an inherent flaw of the Ebel method of which to date no plausible solution has been suggested in the literature.

A more fundamental critique has been made by Gene Glass who argued that the idea of minimal competence is a bad logic and even worse psychology. Glass demonstrated that when comparing Nedelsky and Ebel methods (both use the minimal competence concept), approximately 95% of the examinees would pass the test if the Nedelsky criterion was used, whereas only 50% would pass the Ebel cutoff.

Boaz Shulruf

See also Angoff Method; Classification; Cut Scores; Psychometrics; Standard Setting; Tests

Further Readings

Bontempo, B., Marks, C., & Karabatsos, G. (1998). A meta-analytic assessment of empirical differences in standard setting procedures. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cantor, J. (1989). A validation of Ebel’s method for performance standard setting through its application with comparison approaches to a selected criterion-referenced test. Educational and Psychological Measurement, 49(3), 709–721. doi:10.1177/001316448904900326

Cizek, G., & Bunch, M. (2007). The Ebel method. In G. Cizek & M. Bunch (Eds.), Standard setting (pp. 75–81): SAGE Publications.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall.

Glass, G. (1978). Standards and criteria. Journal of Educational Measurement, 15(4), 237–261.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17(3), 229–235. doi:10.1111/j.1745–3984.1980.tb00830.x

Pat J. Gehrke Pat J. Gehrke Gehrke, Pat J.

Ecological Validity Ecological validity

563

565

Ecological Validity

Ecological validity is the degree of correspondence between the research conditions and the phenomenon being studied as it occurs naturally or outside of the research setting. For example, if one is studying how students solve simple arithmetic problems, the ecological validity of the study depends on how closely the research design corresponds to the conditions in which students encounter and solve such problems in their own lives. Weak ecological validity in the design and conduct of any research may be the result of overreliance on standardized experimental procedures (such as standardized test procedures), an inadequate definition of the phenomenon being studied, misunderstanding the phenomenon’s natural occurrence outside the research setting, a lack of sufficient resources, or the erroneous assumption that context does not affect behavior. This entry discusses why ecological validity is important in research, the development of the concept of ecological validity, and the conditions that are necessary for a study to be said to have strong ecological validity.

Strong ecological validity is a fundamental requirement for research to be meaningful or applicable to conditions outside of contrived research settings. Regardless of the cause, studies with weak ecological validity cannot be generalized to any actually existing phenomenon regardless of their external or internal validity. While sometimes confused with external validity, ecological validity is an independent criterion for good research. External validity measures usually do not demonstrate correspondence to naturally existing phenomenon because they only test correspondence to other research settings. High external validity and even meta-analyses of research results usually show only consistency within contrived research settings.

As with all other forms of validity, ecological validity cannot be answered with a simple “yes” or “no,” but only in degrees. Although it is a fundamental standard simple "yes" or "no," but only in degrees. Although it is a fundamental standard for good research, ecological validity is always a goal we strive to attain and never something a study definitively has or lacks. Instead, we should discuss the ecological validity of research along a spectrum, such as from very strong to very weak.

The standard of ecological validity is generally credited to the work of Kurt Lewin and Egon Brunswik in the 1940s. Both were influential psychologists, Lewin being one of the founders of social psychology. Lewin's theory of ecological validity is the one closest to how we use the term today, while Brunswik's concept of ecological validity was particular to visual processing and is generally only used by Brunswikians.

The modern principles of ecological validity were given shape in the 1970s by Urie Bronfenbrenner and Ulric Neisser. Bronfenbrenner was a developmental psychologist with significant influence on research in education and assessment, while Neisser is considered one of the founders of cognitive psychology. Both were dismayed by how the rise in laboratory experiments in psychology had created a disconnect between the conditions under which a phenomenon was studied and how the phenomenon actually occurred outside the laboratory.

Bronfenbrenner laid out 20 propositions that describe how research in developmental psychology and education should be conducted to ensure ecological validity. The first three of his propositions became the foundation for nearly every theory of ecological validity that followed: 1. The research should be conducted in settings that actually occur in the ecology for purposes other than research or that might occur if practices or policies in that ecology were changed. Bronfenbrenner believed that researchers should try to consider nearly every possible element of the setting and ecology, including roles, physical space, time, activities, and perceptions.

2. The research should keep distortions of the setting to a minimum. While nearly any kind of research risks reactivity or an observer effect, the research design must strive to preserve the integrity of the setting as much as possible.

3. The research design should account for how the larger social and cultural contexts of the participants may be relevant to the ecological validity of the research and setting.

Added to these is a fourth principle of phenomenological validity, which refers to the consistency of the research with how the participants define the situation to the consistency of the research with how the participants define the situation. While less consistently considered by later researchers, this fourth criterion sometimes reveals disparities between the researchers' and participants' understanding of the phenomenon. Such disparities can create research findings that do significant violence to both the phenomenon studied and the self-concept of the participants.

Ecological validity does not, however, require abandoning experiments or even laboratories. Although some have claimed that ecological validity demands that researchers leave laboratory settings and dedicate themselves to field research, Bronfenbrenner was clear that properly designed laboratory experiments can have strong ecological validity. Depending on the research question and the phenomenon being studied, it may even be that a laboratory setting is the ideal space.

Bronfenbrenner cites Stanley Milgram's experiment on obedience to authority figures as a case where an experimental design in a laboratory setting offered strong ecological validity. Additionally, even field research in naturalistic settings can offer weak ecological validity if it distorts its settings in ways that significantly diminish their integrity or alter the social and cultural contexts of the participants. For example, when assessing students' communication competencies, the presence of a camera may heighten anxiety and increase prosocial behavior, producing results that are not reflective of their behaviors in their natural setting.

Much also depends on the phenomenon being studied. If one purports to study participants' knowledge and opinions of a scientific theory, then interference with their usual channels of information and methods of collectively or individually coming to an opinion will dramatically weaken the ecological validity of the research. Alternatively, if one wishes to study how a technique for presenting information might produce a difference in participant's knowledge and opinions, then such interference will be required but should be strictly limited to the bare minimum disturbance in the setting necessary to produce relevant results.

Pat J. Gehrke

See also Design-Based Research; Generalizability; Observer Effect; Reactive Arrangements; Representativeness

Further Readings

Bronfenbrenner, U. (1976). The experimental ecology of education. Educational Researcher, 5, 5–15.

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. Cambridge, MA: Harvard University Press.

Cicourel, A. V. (2007). A personal retrospective view of ecological validity. Text … Talk, 27, 735–752.

Cole, M. (1996). Cultural psychology: A once and future discipline. Cambridge, MA: Harvard University Press.

Lewin, K. (1943). Defining the “field at a given time.” Psychological Review, 50, 292–310.

Gregory Arief D. Liem Gregory Arief D. Liem Liem, Gregory Arief D.

Ser Hong Tan Ser Hong Tan Tan, Ser Hong

Educational Psychology Educational psychology

565

568

Educational Psychology

Educational psychology is the branch of psychology that is centrally concerned with how students can learn effectively. This entry discusses the early development of the field of educational psychology, research areas within educational psychology, comparisons with other psychology disciplines, and the research methods used in educational psychology.

Philosophical Influences

Educational inquiry has its roots in ancient philosophical writings. Plato postulated that individuals are endowed with knowledge from birth, which can be improved by further learning in life. Aristotle brought up the idea that related concepts are more easily understood and remembered. This forms the basis of learning principles elaborated by later educational psychologists in the cognitive paradigm. Additionally, John Locke conceptualized *tabula rasa* to describe the blank slate of mind individuals have at birth. Locke also founded empiricism, which states that experiences with external stimuli are necessary for knowledge to be acquired internally. This lays the grounds to test and establish the validity of knowledge, which is a critical cornerstone in experimental studies in educational psychology.

Educational Leaders’ Perspectives

Early theorizing by educational leaders also contributed to knowledge on how students learn before the establishment of educational psychology. Notable key figures in the 18th and 19th centuries include Jean-Jacques Rousseau, Johann figures in the 18th and 19th centuries include Jean-Jacques Rousseau, Johann Heinrich Pestalozzi, and Johann Friedrich Herbart.

Like Locke, Rousseau believed that experience is instrumental to learning. Rousseau argued strongly that children should be allowed to learn and explore knowledge on their own rather than coming under teachers' directive instructions. Furthermore, Rousseau believed that teachers can tap into learners' idiosyncratic talents and learning styles to promote effective learning.

A similar liberal and dynamic view of education is echoed by Pestalozzi, whose ideas on instructional practices are aligned with current conceptions of student-centered education. Specifically, Pestalozzi spoke against the regurgitation of knowledge but advocated that education should serve one's personal growth and societal advancement. Pestalozzi also wrote about the ideal school climate where warmth, camaraderie, and acceptance prevail instead of fear. In his school, Pestalozzi applied Rousseau's ideas such as leveraging and developing an individual student's talents. This application of ideas and concepts to actual practice came to be known later as applied educational psychology.

Although Rousseau and Pestalozzi largely based their conceptions on their own observations and moral reasoning, Herbart took a more scientific approach to education. Herbart introduced the concept of apperception as a basic psychological process to understand learning. Accordingly, apperception explains interest where existing strong impressions in memory make it more favorable for related new ideas to be assimilated. Learning is therefore explained in terms of a coalition of ideas following cognitive processing. Apperception can also be taken as a motivational theory insofar as it accounts for interest. Motivation continues to be a core pillar in the modern study of educational psychology. Scholars following up on Herbart's works proposed the concept of schema, defined as an existing framework of events and information in the mind, which is one of the key cognitive constructs used to study learning.

Pioneering Educational Psychologists

Three notable psychologists stand out in the early development of the field of educational psychology, namely, G. Stanley Hall, William James, and Edward L. Thorndike. In the late 1800s, Hall was interested in ways to educate students and teachers and believed in studying science in natural environments. Hall started the child study movement, which is a forerunner of educational psychology both theoretically through genetic psychology and methodologically as in the use of questionnaires to collect data in real-world settings. The name *educational psychology* was later introduced into the professional lexicon and became an independent field of study, replacing child study.

Around the same time, James, who was also one of the founding fathers of general psychology, started to give lectures to teachers on psychology in 1892 after publishing his psychology textbook in 1890. These lectures helped to address the practical issues that teachers face. James also advocated very strongly the reciprocal relationship between psychology and teaching in that both areas must be aligned with each other.

In the early 20th century, Thorndike was a key figure in the behaviorist paradigm. Specifically, his learning theory called the law of effect describes how rewards promote behaviors while punishments diminish behaviors. Thorndike extended his influential learning theories to education by applying his concepts of transfer in learning to develop pedagogical practices. Another important area of contribution by Thorndike was his development of measures to assess students' learning in reading, writing, and arithmetic. Being an experimental psychologist, Thorndike shifted the study of educational psychology from field settings to the laboratory, which is not in line with the earlier ideas of Hall and James. Nonetheless, Thorndike's works emphasized heavily the scientific method of study where measurements continue to play a central role in establishing the validity of empirical studies in educational psychology research.

Research Paradigmsin Educational Psychology

In contemporary literature, educational psychology draws together the scientific study of psychology and the applied field of teaching. The multidisciplinary field is also susceptible to the real-world events and general movements in mainstream psychology. During World War II in the 1940s, educational psychologists helped in the hiring and training of military personnel. This led to an interest in training adults, such as teachers, and a shift to work on applied interventions from basic research.

The paradigm shift in mainstream psychology from the behaviorist approach to the cognitivist approach in the 1960s brought about a spike in studies on cognitive representations and processes in educational psychology. Behaviorism is a paradigm in psychology that studies how environmental stimuli cause lasting behavioral changes, hence leading to learning. Two main theories in behaviorism are classical conditioning proposed by Ivan Pavlov and operant conditioning proposed by B. F. Skinner and Thorndike. Although classical conditioning describes learning as arising from affective and physiological reactions to stimuli, operant conditioning describes learning as a result of rewards and punishments for behaviors. The cognitivist approach, in contrast, focuses on internal representations such as networks of thoughts and knowledge. Examples of topics adopting the cognitivist approach in learning include the study of memory, language development, and metacognition, which is the conscious monitoring of cognitive processes.

A schematic mapping of the diverse topics studied in modern educational psychology is best represented with the student in the middle of an ecological network. Starting from inside out, educational psychology examines students' abilities and disabilities in learning. Major areas of student-centric inquiry include the development of students' cognitive processes, motivation, and achievement.

Next, educational psychology reviews the elements present in students' learning environment. Proximal elements include teachers' motivation, classroom instructional practices, and motivational setting. On the other hand, distal elements refer to the broader institutional contexts and pedagogical practices such as assessment and feedback processes. Students are also embedded in the broader sociocultural context where cross-cultural differences often provide meaningful insights into the way students think, learn, and are motivated. These lines of inquiry come under the broader sociocultural perspective of psychology where the emphasis is placed on how social interactions affect learning.

Comparing Educational Psychology With Other Psychology Disciplines

Although educational psychology stands as an independent field of study, it overlaps with theories and methodologies from other psychology disciplines. One closely related discipline is developmental psychology. Stage theories such as Erikson's psychosocial theory from developmental psychology allow researchers and educators to understand learners' developmental trajectories. Educational psychology takes a special interest in developmental psychology theories, such as Jean Piaget's cognitive developmental theory, which have theories, such as Jean Piaget's cognitive developmental theory, which have implications for students' learning.

Cognitive psychology is another specialization whose theories on information processing, language development, and metacognition help to inform instructional practices. Leveraging the knowledge gained from cognitive psychology, educational psychology examines how the cognitive mechanisms and processes affect students' learning while embedded in the broader classroom and educational contexts.

Theories from personality psychology are also utilized to understand individual differences in stable traits and motivation in the learning context. Through social psychology, researchers study social phenomena and the influence on behavior of social factors, such as peer groups, family dynamics, and classroom environments. An application of social psychology theories to educational psychology is, for example, the study of attributions and the influence of contextual cues to explain school success and failure.

The emphasis in social psychology theories is the variability of cognitions and behaviors across different situations and contexts. In contrast, theories in personality psychology view individuals as having unique identities and traits that are stable across situations. Educational psychology reconciles the differences between social and personality psychology by adopting the interactionist perspective, in which behavior results from the interplay of a person's traits and the specific situation.

Importantly, educational psychology is not simply the application of theories in other psychology disciplines or basic research to education. It is confronted with specific issues and problems in schools related to students' learning and instructional practices. Theories and research methodologies unique to the field are used to address these issues in research and practice. The closely intertwined relationship between basic research and applied practices means that both areas inform each other in their developments. This makes educational psychology stand out from the other disciplines. However, tension may arise when findings proved to be effective and beneficial in basic research do not translate easily into pragmatic applied practices.

Research MethodsUsed in Educational Psychology

Both qualitative and quantitative research methods are used in educational

Both qualitative and quantitative research methods are used in educational psychology research. Although less commonly used, qualitative research methods and case studies offer rich information into exploratory areas of interest. For example, observational study documents and describes learning behaviors in natural settings. Quantitative methodology is widely used in educational psychology research, especially questionnaire and correlational studies. Cross-sectional studies have gradually given way to more sophisticated methodologies such as longitudinal studies as well as advanced statistical techniques such as structural equation modeling.

The experimental methodology is also frequently used in educational psychology research to investigate factors that affect learning. One of the key methods in such studies is randomized controlled trials in which participants are randomly assigned into the treatment and control groups. Although the experimental methodology is also used in cognitive and social psychology research, the main difference is that the experimental methodology in educational psychology research has to fulfill both ecological validity and generalizability requirements. This is because findings from educational psychology research, ideally, should inform and guide applied practices.

A cross between qualitative and quantitative research results in the mixed-methods design. Accordingly, mixed-methods design is instrumental to investigate complicated issues in educational psychology research by employing a more extensive and rigorous methodology. The field of educational psychology is also credited for developing the quasi-experimental study methodology as well as instruments used to assess students' learning specifically.

Educational psychology research is constantly evolving to match the dynamic education landscape. Reliability and validity, the two pillars of scientific inquiry, no longer depend only on statistics. Interpretations of data and findings are also crucial to draw reliable and valid conclusions from research. The challenge herein is for educational psychology research to draw from diverse theories outside the discipline, if necessary, and to use appropriate methodological and statistical tools to effectively answer the research questions of interest.

Gregory Arief D. Liem and Ser Hong Tan

See also Behaviorism; Cognitive Development, Theory of; Educational Research, History of; Experimental Method; Learning Theories; Mixed Methods

Research; Motivation; Nonexperimental Designs; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Berliner, D. C. (2006). Educational psychology: Searching for essence throughout a century of influence. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (2nd ed., pp. 3–27). Mahwah, NJ: Erlbaum.

Bredo, E. (2016). Philosophical perspectives on mind, nature, and educational psychology. In L. Corno & E. M. Anderman (Eds.), Handbook of educational psychology (3rd ed., pp. 3–15). New York, NY: Simon … Schuster Macmillan.

Calfee, R. C., & Berliner, D. C. (1996). Introduction to a dynamic and relevant educational psychology. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of educational psychology (1st ed., pp. 1–11). New York, NY: Simon … Schuster Macmillan.

Hilgard, E. R. (1996). History of educational psychology. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of educational psychology (1st ed., pp. 990–1004). New York, NY: Simon … Schuster Macmillan.

Liem, G. A. D., & Martin, A. J. (2013). Latent variable modeling in educational psychology: Insights from a motivation and engagement research program. In M. Khine (Ed.), Applications of structural equation modeling in educational research and practice (pp. 187–216). Rotterdam, the Netherlands: Sense Publisher.

Mertens, D. M. (2015). Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods (4th ed.). Thousand Oaks, CA: Sage.

Penuel, W. R., & Frank, K. A. (2016). Modes of inquiry in educational psychology and learning sciences research. In L. Corno & E. M. Anderman (Eds.), Handbook of educational psychology (3rd ed., pp. 16–28). New York, NY: Simon … Schuster Macmillan.

Craig A. Mertler Craig A. Mertler Mertler, Craig A.

Educational Research, History of Educational research, history of

569

573

Educational Research, History of

The history of educational research reflects an almost 200-year journey that began with the recognition in the mid-1800s that education is a science. Educational research methodology has been dominated from the start by a quantitative experimental approach. This domination continues, although the last 50 years have seen a rise in and acceptance of postpositivistic and qualitative research approaches. In the United States, federal funding and policy have supported important educational research and driven the increased use of standardized tests in schools.

This entry highlights major events in the field of educational research, primarily in the United States, to provide a summary of its progression over the nearly two centuries since formalized data collection and dissemination first began. It is not meant to represent an exhaustive history of educational research but rather provides a sense of the nature of educational research as it evolved over time.

Education as Worthyof Scientific Investigation

In the mid-19th century, educators Horace Mann and Henry Barnard were pioneers in educational data collection and the dissemination of educational literature. They suggested that school supervision and planning should be influenced by systematic data collection to examine and describe the function of education in a democracy and to develop scholarly literature to make available to educators new ideas related to education that were emerging in other countries. In 1855, Barnard founded the *American Journal of Education* and served as its editor for more than 25 years. In 1867, the U.S. Congress passed “An Act to establish a Department of Education,” with the purpose of collecting information on schools and teaching methods to help states establish effective school systems.

The child study movement, an important precursor to the field of educational psychology, began with the publication of psychologist G. Stanley Hall's essay "The Contents of Children's Minds" in an 1883 issue of *The Princeton Review*. The essay resulted from Hall's survey study of 64 kindergarten teachers and hundreds of children. His findings influenced the reform of schools and the training of teachers. Hall encouraged belief in educational progress through the scientific study of the child.

By the late 19th century, there was increasing interest in and focus on scientific exploration and investigation, controlled experimentation, and rational reform. There was a substantial increase in the use of surveys as a prime method of determining directions for needed reforms in education. They became a routine feature of school management, with teams of professors and experienced school administrators from other communities coming together to review local school systems.

In 1895, Joseph Mayer Rice developed standardized spelling tests and administered them to approximately 16,000 students. The purpose was to examine the relationship between spelling instruction and actual performance. Rice is often credited with being the founder of empirical scholarship in education. A year later, John Dewey founded the Laboratory School at the University of Chicago. The school was an attempt to explore practical techniques and test hypotheses that others could use in practice, but to do so from a psychological perspective and not a behavioral one, which had long influenced educational research up to that point.

In 1904, Dewey's colleague, psychologist Edward Thorndike, published *An Introduction to the Theory of Mental and Social Measurements*. The two shared similar approaches and beliefs in terms of the study of education. More so than Dewey, however, Thorndike had a preference for the production of statistics and precise measurements that could be analyzed as an approach to educational research. This preference for quantification became widely accepted across academia in the United States and abroad and helped educational research become perceived as a legitimate science.

In 1916, the American Educational Research Association, originally known as the National Association of Directors of Educational Research, was founded in Washington, D.C., as a professional organization representing educational researchers in the United States and around the world. Its purpose is to advance knowledge about education and promote the use of research to improve education and contribute to the public good. In 1919, the American Educational Research Association began publication of *The Journal of Educational Research*.

Educational Measurement

In 1905, 10 years after Rice developed his standardized spelling tests, Alfred Binet’s article, “New Methods for the Diagnosis of the Intellectual Level of Subnormals” was published in France. The article described his collaborative work with Theodore Simon in the development of a measurement instrument that would identify students with mental retardation. The Binet-Simon Scale, a standardized intelligence test, became an effective means of measuring intelligence and introduced the still widely popular definition of measurable *intelligence* as that ability that predicts success in school. The concept of standardization also extended to high school credits, as the Carnegie Foundation for the Advancement of Teaching, founded in the United States in 1905, encouraged the adoption of a standard unit equating “seat time” to high school credits.

By 1916, Lewis M. Terman and Stanford University graduate students completed an American (and English language) version of the Binet-Simon Scale. The Stanford Revision of the Binet-Simon Scale, now known as the Stanford-Binet Intelligence Scales, became a widely used individual intelligence test, and along with it, the concept of the IQ was born. Around this time, the United States entered World War I, and there was a need for screening and classifying the intellectual ability of its recruits. Robert Yerkes, an Army officer and then president of the American Psychological Association, became chairman of the Committee on Psychological Examination of Recruits. The committee, which included Terman, developed a group intelligence test, which unlike earlier tests did not need to be given one-on-one. Yerkes and his team of psychologists designed the Army Alpha and β tests, which had little impact on the war but did lay the groundwork for future standardized tests.

In 1939, David Wechsler developed the Wechsler Adult Intelligence Scale. It introduced the concept of the deviation IQ, which calculates IQ scores based on how far subjects’ scores differ (or deviate) from the mean score of others who are the same age. The Wechsler Intelligence Scale for Children is still widely used in U.S. schools to help identify students needing special education services.

Influenced by the Army Alpha group intelligence test, standardized tests for college admission appeared in the first half of the 20th century. In 1926, SAT was first administered, and in 1949, the GRE General Test was developed. In 1959, the ACT was first administered as a college readiness assessment. Unlike the SAT, it did not focus on cognitive reasoning but rather on information that is taught in schools.

Research Design and Methodology

In the early 1900s, Charles Judd arrived at the University of Chicago. Judd brought a more rigorous and scientific approach to the study of education. He was a proponent of the scientific method and worked to integrate it into the practice of educational research. Judd had a preference for quantitative data collection and analysis, with a focus on psychology. Around the same time, Thorndike’s book, *Educational Psychology: The Psychology of Learning*, was published. In it, Thorndike described his theory that human learning involves connections between stimuli and responses. These ideas, which contradicted much of traditional psychology, came to greatly influence American educational practice.

In 1916, Lucy Sprague Mitchell founded the Bureau of Educational Experiments in New York City, with the purpose of studying child development and children’s learning. A year later, the Iowa Child Welfare Research Station opened at the University of Iowa. Its focus was to serve as a demonstration center and a laboratory for the study of “normal” children.

Among the early important studies in child development was the work of John B. Watson and his assistant Rosalie Rayner, who conducted experiments with children in the 1920s using classical conditioning. Watson and Rayner’s work—often referred to as the Little Albert Study—showed that children could be conditioned to fear stimuli to which they had previously been unafraid. This study could not be conducted today because of ethical safeguards currently in place.

In 1921, another important study began when Terman initiated his longitudinal study of gifted children. At first, the study was to span 10 years, but it was later extended to study these same children as adults. Terman held the belief that, by identifying the most gifted at a young age, society could ensure the flow of talent to leadership positions. Data collection continued for more than 75 years.

The use of inferential statistics and hypothesis testing, introduced in the context of agricultural science, became the primary method of analyzing data in the 1920s and 1930s. In 1921, R. A. Fisher first published his applications of the analysis of variance to crop variation data. In 1925, a subsequent book by Fisher, *Statistical Methods for Research Workers*, was published and would later become one of educational research's most influential books. In this book, which refined statistical methods, Fisher first put forth the arbitrarily set—and popularly used—*p* value of .05 sciences. Ten years later, the implications of significance testing on research design were made clear with Fisher's *The Design of Experiments*, that includes the "lady tasting tea," now a famous design of a statistical randomized experiment which uses Fisher's exact test and is the original exposition of Fisher's notion of a null hypothesis. In 1963, Donald T. Campbell and Julian C. Stanley described and defined *experimental* and *quasi-experimental* designs for research and threats to the validity of results. This framework became the standard for quantitative educational research.

In the 1970s, important new quantitative research designs and approaches ware developed. In a 1976 presidential address at the American Educational Research Association, Gene Glass coined the term *meta-analysis* and explained it as essentially "the analysis of analyses." After years of applications in medical research, it is now applied to educational research with increasing frequency. In 1977, Lee Cronbach and Richard Snow posited their theory of aptitude-treatment interaction, the concept that some instructional strategies (treatments) are either more or less effective for individuals based on their inherent aptitudes and specific abilities.

Alternative paradigms for educational research emerged starting in the 1950s and 1960s. Action research was introduced as an approach to research that focuses on change, but at a more local (i.e., not large-scale) level. This approach attracted attention in schools, as educators seeking change in schools set up research projects in local schools under the guidance of local university professors. Then, a postpositivist movement in educational research began to take shape. Although positivistic approaches to conducting research in education continue to be favored by many social scientists, other epistemological continue to be favored by many social scientists, other epistemological approaches have been introduced or reemerged as strong, viable alternatives.

The late 20th century was a time of vigorous debate—both inside and outside of academia—about the virtues of various theoretical perspectives about knowledge, science, and methodologies. These debates continue to play a very important role in the continuing development of educational research as a field. In 1985, Yvonna Lincoln and Egon Guba published their influential book, titled *Naturalistic Inquiry*. They confronted a basic premise that all questions can be answered by employing empirical and testable research techniques and maintain that there are scientific facts that existing paradigms cannot explain, arguing against traditional positivistic inquiry.

Educational Reforms

In the 1920s, the focus of education began to change to an emphasis on social control and efficiency, and disagreement among educational research scholars as to the purpose of education began to grow. The population of the United States was rapidly increasing and the demographic makeup changing markedly, due to immigrants from around the globe as well as the migration of African Americans from rural areas and Southern states to urban areas in the Northeast and Midwest. The demographics of student bodies began to diversify rapidly. The arrival of new immigrants coincided with the U.S. Army’s “testing movement” that emerged during World War I. Sociology researchers at the University of Chicago began to study racial differences in test scores, as did Otto Klineberg at Columbia University.

The Eight-Year Study (also known as the Thirty-School Study) was undertaken by the Progressive Education Association in 1930. In this study, 30 high schools redesigned their curriculum while initiating innovative practices in student testing, program assessment, student guidance, curriculum design, and staff development. The purpose was to determine whether subject matter requirements as prerequisites for college admission were necessary and justified. The students in the experimental schools performed as well in college as did their counterparts. The experimental schools were stimulated to develop new programs that proved to be better and more effective for young people.

The late 1930s through the mid-1950s were a period of a *pragmatic action* orientation in education. This period initially began to witness a decline in educational research, due in part to a gradual separation of the previously collaborative relationship that had existed between pragmatically oriented educators and more traditionally oriented academicians. This observed decline was also due to limited availability of resources related to the Great Depression and to World War II. The latter half of this period of time saw educational research resurface and flourish. Growth in schools of education across the country continued to rise. In addition, more and more academic journals with a focus on educational issues emerged as a mechanism to disseminate new knowledge related to educational issues.

Federal Involvementin Educational Reform and Research

In the early 1950s, the National Science Foundation led an investigation of the nature and status of science education in schools in the United States and determined there was a gross inadequacy of instructional materials available to teachers. As a result, the Cooperative Research Act, passed by Congress in 1954, authorized the Department of Education to enter into financial agreements with colleges, universities, and state educational agencies for research, surveys, and demonstrations in the field of education. Through this law, the federal government took a much more active role in advancing and funding research on education within academia.

In 1954, the National Science Foundation began providing support for activities aimed at the improvement of mathematics and science instruction in elementary and secondary schools. The decade that followed saw passage of significant legislation and creation of initiatives supporting educational research and providing a means for the dissemination of new educational knowledge, including the National Defense Act of 1958 and the establishment of the Educational Resources Information Center in 1964.

During the 1970s, the federal government introduced an evidence-based movement in educational research. Prior to this time, research in education focused largely on resource allocation, student access, and curriculum and paid relatively little attention to actual results of educational research studies. Assessment and evaluation of educational programs, in order to determine whether these programs are worth the money being spent on them, took on greater importance.

In 1983, the National Commission on Excellence in Education released its report titled *A Nation at Risk*. It called for sweeping reforms in public education and teacher training. The resulting national debate over school reform culminated with the controversial No Child Left Behind Act signed into law by President George W. Bush on January 8, 2002. The law mandated high-stakes student testing, held schools accountable for student achievement levels, and provided penalties for schools that did not make adequate yearly progress toward meeting the goals of No Child Left Behind Act. The science of educational measurement flourished. Although Congress gave states more power over academic standards when it replaced No Child Left Behind Act in 2015 with the Every Student Succeeds Act, the law still requires annual statewide testing and reporting of results.

Craig A. Mertler

See also Educational Psychology; Educational Researchers, Training of; Learning Theories

Further Readings

Comp, D. (2009, July 13). A brief history of research on education [Blog post]. Retrieved from http://ihec-djc.blogspot.com/2009/07/brief-history-of-research-on-education.html

Johanningmeier, E. V., & Richardson, T. (2008). Educational research, the national agenda, and educational reform: A history. In K. Riley (Series Ed.), Studies in the history of education. Charlotte, NC: Information Age.

Knox, H. (1971). A history of educational research in the United States. Retrieved from ERIC database (ED088800).

Lagemann, E. C. (2000). An elusive science: The troubling history of education research. Chicago, IL: University of Chicago Press.

Sass, E. (2016, October 6). American educational history: A hypertext timeline. Retrieved from http://www.eds-resources.com/educationhistorytimeline.html

Margaret Kristin Merga Margaret Kristin Merga Merga, Margaret Kristin

Educational Researchers, Training of Educational researchers, training of

Educational Researchers, Training of

The training of educational researchers refers to the process by which individuals acquire the skills and knowledge required to effectively and ethically conduct research in the field of education. Training in educational research may commonly begin during pursuit of a graduate degree, though it may begin earlier at the undergraduate level. The training of educational researchers involves an ongoing process of learning beyond completion of formal academic training, extending across the career of educational researchers. As such, even highly established educational researchers may view themselves as students of educational research, as they may be committed life-long learners in this field, constantly acquiring a greater depth and breadth of skills and knowledge in this area. This entry discusses the importance of adequate training of educational researchers, areas of initial training in how to conduct research, and training specific to conducting educational research.

Adequate training of educational researchers is vital due to the contribution that educational research ultimately plays in informing both teaching practice and policy in schools, colleges, and universities. Contemporary educational institutions are increasingly urged to adopt evidence-based strategies in order to ensure instructional pedagogy and content is compliant with current understandings of best practice. This evidence is principally derived from the output of educational researchers. In order for educational institutions to have access to the highest quality evidence, educational researchers need to have undertaken sufficient training in their field, so that their work adheres to the highest research standards.

Training of educational researchers may be conducted in many forms. For example, individuals may attend seminars or dyadic peer support sessions or complete online modules. In contemporary times, training opportunities are complete online modules. In contemporary times, training opportunities are becoming increasingly multimodal in response to newer forms of technology. While traditionally a principal graduate school supervisor held primary responsibility for the training of educational researchers, this responsibility has been increasingly shared by larger supervisory teams, graduate research schools, and training groups within institutions.

Areas of Initial Training

Initial training of educational researchers may involve the traditional components of research. Students of educational research learn to review the literature so that their research is responsive to and builds upon extant findings in their field. They receive instruction in adopting appropriate theoretical and conceptual frameworks as lenses for their research to make explicit the beliefs and assumptions that underpin their research. Students learn how to choose and implement sound research methods, adhering to rigorous technical and ethical requirements.

Students need to attain practical skills for undertaking research data collection and acquire broad familiarity with a range of data analysis methods. In addition, a graduate degree should be viewed as an opportunity for academic writing apprenticeship, with students supported to learn to write effectively using an appropriate authorial voice. Students also need to understand the importance of acknowledging the limitations of their research findings and identifying key areas for future research.

Responsiveness to contemporary academic culture also places emphasis on communicative competence, which increasingly extends beyond the traditional thesis. Students are expected to attend to dissemination of findings through peer-reviewed publication earlier in the research journey and share their findings beyond the research community through an active commitment to research translation.

Students also need to learn how to be responsive to, and how to give, critical and constructive feedback within the peer-review context, and how to communicate their findings effectively across a range of writing and speaking styles. For example, in addition to writing peer-reviewed journal articles, researchers may need to learn to write conference papers, press releases, and plain English reports for participating schools. They may need to learn to deliver their findings orally to audiences outside of academia, such as parents from linguistically and culturally diverse backgrounds at school seminars.

As relatively few graduate students ultimately become professional academic researchers, there has been an increasing focus in recent times on providing research skills that can be beneficial in professions beyond academia. Many of these communicative competence skills are transferrable, so students who do not remain in academia may draw upon these communicative skills in seeking industry-based employment.

In addition to these communicative competencies, research training may involve fostering beneficial social and life skills. For example, research students may benefit from training in areas such as resilience and stress management in order to deal with the challenges of the research environment.

Training to ConductResearch in Educational Contexts

Beyond these somewhat generic research skills is a subset of skills that are to some extent unique to the educational research context. Educational researchers often conduct data collection in schools, colleges, and universities, and they must be responsive to the contexts, cultures, and expectations of these institutions, without compromising the integrity of their research. Educational researchers must also have a sound understanding of the ethical protocols and processes relevant to collecting data from minors, as defined by their own institutional and regulatory bodies, such as government departments of education, and those imposed within the school or other educational institution. This often involves seeking ethics approval from multiple sources prior to exposing students to the research in any form. Ethics permissions sought from participants may also be multilayered in the case of minors, with approvals required from both the minors and their legal guardians. Conducting research in educational contexts often involves working with marginalized communities and individuals, and as such, educational researchers need to have had current diversity training to ensure that their research design, tools, and methods do not further exacerbate existing inequities.

Margaret Kristin Merga

See also American Educational Research Association; American Evaluation Association; American Psychological Association; Educational Psychology; Educational Research, History of; Learning Theories

Further Readings

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. Learning and Instruction, 7(2), 161–186.

Esbensen, F., Melde, C., Taylor, T. J., & Peterson, D. (2008). Active parental consent in school-based research. Evaluation Review, 32(4), 335–362.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. Educational Researcher, 31(8), 4–14.

Labaree, D. F. (2003). The peculiar problems of preparing educational researchers. Educational Researcher, 32(4), 13–22.

Merga, M. (2015). Thesis by publication in education: An autoethnographic perspective for educational researchers. Issues in Educational Research, 25(3), 291–308.

Paul, J. L., & Marfo, K. (2001). Preparation of educational researchers in philosophical foundations of inquiry. Review of Educational Research, 71(4), 525–547.

Irvin R. Katz Irvin R. Katz Katz, Irvin R.

Jeffrey P. Johnson Jeffrey P. Johnson Johnson, Jeffrey P.

Educational Testing Service Educational testing service

Educational Testing Service

The Educational Testing Service (ETS) is a nonprofit educational measurement organization based in Princeton, New Jersey. Founded in 1947, ETS develops, administers, and scores more than 60 million tests in more than 190 countries each year. ETS's portfolio, consisting of tests it owns as well as some it develops under contract, serves purposes such as:

- admissions (e.g., the GRE General Test, the College Board's SAT, ETS's Test of English as a Foreign Language programs for English learners),
- course equivalency (e.g., the College Board's Advanced Placement program, ETS's high school equivalency test),
- licensure or certification (e.g., ETS's Praxis teacher certification exams),
- school accountability (e.g., state K–12 assessment programs), and
- public policy (e.g., National Assessment of Educational Progress, Programme for International Student Assessment).

This entry describes ETS's history, its research and development work, and its contributions to educational measurement and assessment training.

History

During and after World War II, some prominent higher education officials advocated for coordinating educational testing research and development under a unified organization. These officials included Harvard University president James Bryant Conant and his assistant dean of admissions, Henry Chauncey, who later became ETS's first president.

Three nonprofit organizations—the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board—combined their testing programs to create ETS. In 1947, New York state granted ETS a charter as a nonprofit organization. ETS operates under U.S. laws governing tax-exempt organizations that serve a public benefit.

Among the programs for which the new organization assumed research and development responsibilities were American Council on Education’s National Teacher Examinations (NTE), the Carnegie Foundation’s Graduate Record Examination (now the GRE General Test), and College Entrance Examination Board’s SAT. The NTE was the predecessor to ETS’s Praxis Series, which many states now use in teacher certification. The College Board remained the SAT’s sponsor and retained control of its design and use—but ETS began writing test questions and administering the test under contract. Although its initial test portfolio focused on admissions or certification in U.S. postsecondary education, ETS’s current assessments and research now serve needs at all education levels, internationally, and in the workplace.

Research and Development

As of early 2016, ETS had more than 3,300 employees. At least 1,200 work in ETS Research … Development, which is responsible for assessment development, operational statistical analysis, and research. Several thousand additional employees work at ETS’s wholly owned subsidiaries.

Research Programs

ETS states its mission as “to advance quality and equity in education by providing fair and valid assessments, research and related services” (ETS, Who We Are, n.d., n.p.). ETS research aims to support existing assessments; develop knowledge and capabilities for future assessments, related products, and the field more generally; and inform public policy related to all levels of education.

In its first few decades, ETS established a legacy of research and development contributions to the field in areas such as classical test theory, item response theory, test validity theory, score equating and scaling methods, large-scale survey assessment methods, differential item functioning, and computer adaptive testing. ETS researchers contributed beyond psychometrics as well, investigating topics such as the structure of abilities, early childhood education, and response styles.

By the early 21st century, ETS had intensified research in the cognitive and learning sciences to support demands for more complex, technology-rich assessments and the assessment of new constructs such as workforce readiness and digital literacy. This research included efforts to better understand the nature of proficiency in specific domains and to apply a systematic approach known as evidence-centered design to create assessments in those domains, doing so in a way intended to be accessible to all test takers.

For new and existing tests, ETS conducts validity research related to, for example:

design and development of assessment systems, in terms of intended score use;

fairness, or whether scores have the same meaning for all test takers;

score interpretation, including how scores are reported to support decision making; and

intended and unintended consequences of assessment use.

Assessment Design,Development, and Scoring

ETS assessment developers, psychometricians, and researchers collaborate to design and develop tests that are intended to be fair, yield reliable results, and be valid for their intended purposes. Assessment developers follow procedures that include, but are not limited to, activities such as:

creating specifications that define a test’s target population and the knowledge and skills it is designed to measure;

collaborating with external stakeholder committees and with ETS areas responsible for administering tests on paper or on computer;

establishing score scales and procedures for scoring and reporting results;

maintaining program quality through activities such as pretesting content, equating results to previous administrations, performing postadministration item analyses for quality control purposes, and conducting validity research; and

undergoing regular internal audits of a program’s alignment to the *ETS*

Standards for Quality and Fairness, which are intended to be consistent with the *Standards for Educational and Psychological Testing*.

Internships and Education

ETS’s summer research internships support graduate students who collaborate on-site with ETS researchers. Interns are usually studying a field closely related to ETS’s Research … Development work, such as psychometrics, cognitive psychology, learning sciences, assessment of English language learners, or natural language processing. ETS also sponsors a limited number of postdoctoral fellows in these fields. In addition, ETS’s Global Institute offers seminars and training on topics in assessment for nongovernmental organizations, ministries of education, and professionals working in education and measurement worldwide.

Irvin R. Katz and Jeffrey P. Johnson

See also Admissions Tests; Classical Test Theory; Item Response Theory; SAT; *Standards for Educational and Psychological Testing*

Further Readings

American Educational Research Association, American Psychological Association, … National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bennett, R. E. (2005). What does it mean to be a nonprofit educational measurement organization in the 21st century? Retrieved from http://www.ets.org/Media/Research/pdf/Nonprofit.pdf

Educational Testing Service. (n.d.). About ETS: Who we are. Retrieved from https://www.ets.org/about/who/

Educational Testing Service. (n.d.). ETS’s research legacy in statistics and psychometrics. Retrieved from http://www.ets.org/research/topics/statistics_psychometrics/legacy/

Educational Testing Service. (n.d.). How ETS approaches testing: Purpose of standardized tests. Retrieved from http://www.ets.org/understanding_testing/purpose

Educational Testing Service. (2015). ETS standards for quality and fairness. Retrieved from http://www.ets.org/s/about/pdf/standards.pdf

Elliot, N. (2014). Henry Chauncey: An American life. New York, NY: Peter Lang.

Mislevy, R. J., Almond, R. F., & Lukas, J. F. (2003). A brief introduction to evidence-centered design (ETS Research Report No. RR-03–16). Princeton, NJ: ETS.

Stricker, L. J. (2013). ETS research on cognitive, personality, and social psychology: I (ETS Report No. SPC-13–01). Princeton, NJ: ETS.

Catherine O. Fritz Catherine O. Fritz Fritz, Catherine O.

Peter E. Morris Peter E. Morris Morris, Peter E.

Effect Size

Effect size

576

578

Effect Size

Research seeks to identify effects in the sense of a relationship in the data. Effects are usefully described in terms of their *size* and their *likelihood of being observed in further samples from the same population*. Effect sizes are independent of sample size, unlike tests of statistical significance. Although tests of significance usually focus on whether an observed statistical value is likely to be greater than zero in the population from which a sample was chosen, effect sizes are a summary of the observed relationship in sample data.

Effect sizes are interpreted in the light of their potential importance—even a small effect is important if it may save or markedly improve lives. In general, though, larger effects have more impact and so are seen as more important. Effect sizes are also used when determining the number of participants required for follow-on research with adequate power. There are effect size statistics for all types of data and effects. This entry describes a few of the more commonly used effect size statistics.

The effect size describes the effect observed in a sample; it also provides an estimate of the effect size in the population from which the sample was drawn. If the researcher intends to draw inferences about the population, then the confidence interval (CI) for that effect size is also required; it indicates the likely range within which the actual population effect size falls. The CI for the effect size statistic becomes narrower—providing a more precise estimate—with larger samples. Using an effect size statistic along with appropriate CIs (e.g., 95% CIs) provides a clear, simple, decision-making procedure that can complement inferential statistical tests or can be used in their place. If the CI does not include zero, then it is quite likely that there is an effect in the population and the size of that effect is likely to be within the CI.

When examining differences between groups, the simplest, and perhaps most meaningful, effect size statistic is the difference between the means for the groups. For example, if the reading age of one group of children is greater than that of an age-matched group, the difference in mean reading ages is a useful description of the size of the effect.

But much research compares groups using measures unique to a particular study. For example, differences in reaction times on a task or information recalled from studied material can be more useful when reported in standardized units. Even when reporting simple effect sizes, it is usually good practice to report standardized effect sizes as well.

The *d* statistic is a widely used standardized effect size; it is quite easily calculated by subtracting one mean (M_1) from the other (M_2) and dividing by an appropriate standard deviation (*SD*):

$$d = \frac{(M_1 - M_2)}{SD}$$

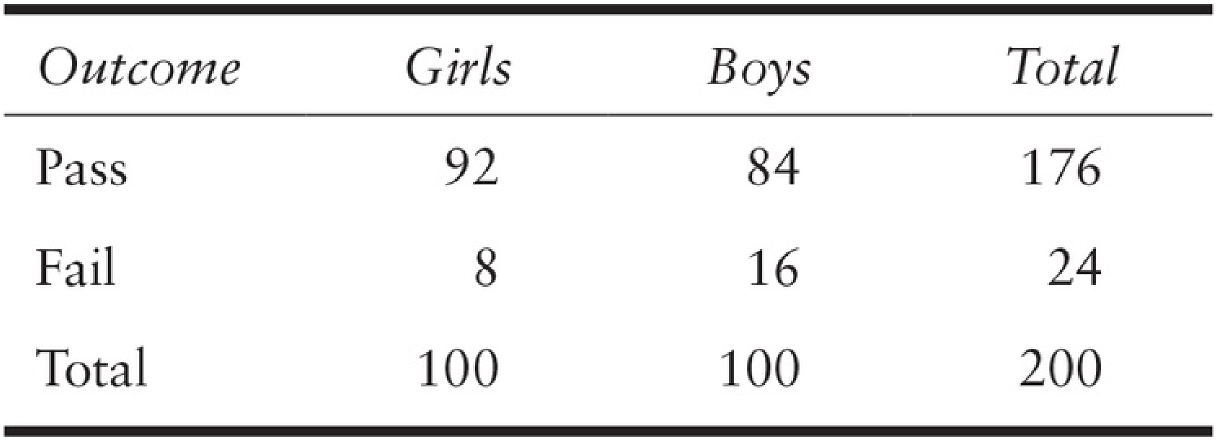
Where the *SDs* of the two means differ substantially there are simple formulae for weighting the two contributing *SDs*.

An alternative to *d* is the point biserial correlation, *r* or r_{pb}, which can be calculated by coding the two groups as 1 and 2, respectively, and correlating these codes with the data. One strength of the *r* statistic is that it has a familiar meaning and range, from 0 to 1. Furthermore, r^2 describes the proportion of variability in the data that is related to group membership. Meta-analyses often use *r* when combining the observed effect sizes across several studies. There are simple formulae to convert between *d* and *r* as well as to calculate either from the *t* statistic.

Where an effect involves more than two levels of a variable (i.e., when three or more groups are compared), then a commonly reported effect size is eta squared (η^2). η^2 is similar to r^2, describing the proportion of the total variability in a data set that is associated with an effect. A closely related statistic, often reported for multifactor designs, is partial , which describes the proportion of variability associated with an effect, after excluding the effects of other variables.

For multiple regression, the R^2 and R^2 change statistics directly provide effect size measures that are equivalent to η^2 and except that the latter capture nonlinear as well as linear relationships. Downloadable resources can calculate CIs for η^2 and R^2 statistics.

The effect size statistics described so far are for normally distributed, continuous data; there are also effect size statistics for data that are ranked and analyzed by tests such as Mann–Whitney or Wilcoxon signed ranks, as well as for categorical data. For the ranking tests, a *z* value can be calculated from the observed *U* or *T* statistic and then can be converted to the *r* effect size statistic. (Divide *z* by the square root of the sample size.) For categorical data with a 2 × 2 design (e.g., two categories and two outcomes such as girls/boys and pass/fail), there are several effect size statistics; commonly used ones include: phi coefficient (φ or *r*φ), risk ratio, and odds ratio. The phi coefficient is easy to interpret because it is similar to a correlation; it can be calculated from chi-square: , where *N* is the total number of cases. The risk ratio is a simple ratio of two proportions, such as the proportion of girls who fail divided by the proportion of boys who fail; it describes the difference in failure rates, as shown using fictional data in Table 1.



Outcome	*Girls*	*Boys*	*Total*
Pass	92	84	176
Fail	8	16	24
Total	100	100	200

Note: Chi-square for these data = 21.76.

For the example, in Table 1, r_{φ} = (.08/.16) = .5; girls are half as likely to fail as are boys. The odds ratio is similar; it is the ratio of the odds for the two groups. From Table 1, the odds that a girl fails are 8/92 = .09; the odds that a boy fails is . The odds ratio for girls/boys is .09/.19 = .47; the odds of a girl failing is slightly less than half the odds of a boy failing.

Catherine O. Fritz and Peter E. Morris

See also Chi-Square Test; Eta Squared; Inferential Statistics; Meta-Analysis; Multiple Linear Regression; Normal Distribution; Odds Ratio; Phi Correlation Coefficient; Power

Further Readings

Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York, NY: Routledge.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. Journal of Experimental Psychology: General, 141, 2–18. doi:10.1037/a0024338

Grissom, R. J., & Kim, J. J. (2012). Effect sizes for research: Univariate and multivariate applications (2nd ed.). New York, NY: Routledge.

Lakens, D. (2014, June 7). Calculating confidence intervals for Cohen’s *d* and eta-squared using SPSS, R, and Stata. Retrieved from http://daniellakens.blogspot.co.uk/2014/06/calculating-confidence-intervals-for.html

Richard E. Boyatzis Richard E. Boyatzis Boyatzis, Richard E.

Emotional Intelligence Emotional intelligence

578

580

Emotional Intelligence

The concept of emotional intelligence (EI) is widely perceived as relevant to the workplace, home, and educational settings. Meanwhile, the scholarly work on EI has been challenged because of disagreements about many definitions and measures. At the most basic level, EI is the intelligent use of one’s emotions. Another definition is the ability to manage one’s own emotions and the emotions of others when interacting with them. The latter is sometimes separated from EI and called social intelligence (SI). This entry discusses the different levels at which EI has been studied, measures of EI, and research into efforts to help people develop EI.

First introduced to psychological research in academic journals by Peter Salovey and John Mayer in 1990 and then popularized by Daniel Goleman in 1995, the understanding and development of EI has become a major theme in education from K–12 to higher education, in organizational and leadership development, and throughout all professions from medicine and dentistry to engineering.

Multiple Levels of EI

An overview of the research suggests that EI exists at multiple levels within a person. These are (a) the physiological level, in which EI is seen as based on an individual’s preferred neural networks or hormonal dispositions; (b) trait level; (c) self-schema or self-perception level; and (d) the behavioral level.

At the trait level, EI may be seen as a type of intelligence. A trait measure looks at EI as the ability to be aware of and manage one’s emotions and those of others. In 2005, Neal Ashkanasy and Catherine Daus categorized this approach to measurement as Stream 1 when this ability was measured directly and Stream

2 when it was measured through self-assessment built on the ability model. Mayer, Salovey, and David Caruso claimed that their approach to EI is a type of intelligence related to general intelligence (*g*). Their measure of this trait level is the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT).

At the self-perception level, EI is a set of characteristics that are best assessed by the person. Self-perception of EI would likely be different from but related to personality traits more than to general cognitive ability, referred to in psychology as *g*. Although measures such as the EQ-i, designed by Reuven Bar-On, have been shown to predict job performance in various meta-analyses, they also show a consistent and strong correlation with personality traits in these studies. These measures have been labeled mixed models by Mayer. Ashkanasy and Daus referred to the measures as Stream 2 if they are based on the MSCEIT model and Stream 3 if they are based on other models of EI.

The behavioral level of EI requires observation of behavioral expression of EI. Known as behavioral EI, it offers a closer link to job and life outcomes. Notably, it has been shown to predict job performance above and beyond *g* and personality.

Although some scholars have claimed that the competencies that are characteristic of EI are not a type of intelligence, others have claimed they can be considered part of naturally, neurologically driven capability and, therefore, are a form of intelligence. Behavioral measures of EI would include coding of audiotapes of critical incidents from work, as first documented by Richard Boyatzis in 1982 or videotapes of simulations or multisource observations as explained by Boyatzis in 2009.

The behavioral level is based on David McClelland’s concept of competency from his 1951 book *Personality*. Boyatzis has described this level as involving *observed, specific behaviors* organized around a single intent. The behavior and actions can be thought of as manifestations of the intent appropriate to the specific situations or times. EI at this level “is an ability to recognize, understand, and use emotional information about oneself that leads to or causes effective or superior performance” (cited in Emmerling … Boyatzis, 2012, p. 8). SI at this level is “the ability to recognize, understand, and use emotional information about others that leads to or causes effective or superior performance” (cited in Emmerling … Boyatzis, 2012, p. 8).

The physiological level of EI is found in hormonal systems and neural networks.

EI characteristics such as emotional self-control and internal reflections appear to be related to the task-positive network in the brain. Tony Jack and his colleagues have shown in fMRI studies that when people are engaged in dealing with social situations, a different network is activated and it is quite similar to the default mode network. These two neural circuits are different, which implies that emotional and SI are different. This difference is supported by endocrine studies, suggesting that the exercise of EI is more closely associated with the sympathetic nervous system and that of SI with the parasympathetic nervous system.

In addition to the different levels, there are three other distinctions that are important to consider in defining what EI is and how best to measure it. The first is whether it is a single construct or a composite set of many abilities. There is pressure from those applying the concept to arrive at a single construct that would enable people to classify others as relatively high or low in EI. This invites attributions of a halo effect. Alternatively, some believe that EI is a composite set of abilities or, at the behavioral level, competencies. From this perspective, it is overly simplistic to call it one “thing.” A second major issue is the degree to which it is an “intelligence” and closely related to *g*.

Mayer and Salovey argued that any definition of and measurement of EI should be closely related to *g*. Others, including Cary Cherniss and Boyatzis, have referred to it as an intelligence because EI and SI can be related to distinct neural networks, so that they can be considered a result of specific cognitive functioning. In the latter approach, EI and SI would be somewhat related to *g* but not closely. The MSCEIT measure, conceptualized as a performance trait or ability, does appear more closely related to measures of *g* than other measures of EI.

A third issue of conflict is whether EI is merely another way to describe personality. The level of self-perception or self-schema would suggest that how individuals view their own EI would be related to their own personality traits. Research has shown that the EQ-i, mostly used as a self-assessment measure in research and applications, is more closely related to personality trait measures than other approaches to EI. The three most widely used measures of EI are the MSCEIT, the EQ-i, and the Emotional and Social Competence Inventory. The Emotional and Social Competence Inventory was designed by Goleman, Boyatzis, and Hay Group, a global management consulting firm, to assess the emotional and social competencies thought to characterize outstanding leaders.

Developing EI

The Consortium for Research on Emotional Intelligence in Organizations identified only 15 programs in the scientific literature that improved EI in adults. The few published studies of more than one of these competencies show an overall improvement in EI of about 10%. MBA programs have even less impact unless they are trying to develop EI. Researchers found that graduating students from two highly ranked MBA programs showed only 2% improvement in EI. Students from four other high-ranking MBA programs showed a gain of 4% in self-awareness and self-management abilities but a *decrease* of 3% in social awareness and relationship management.

Longitudinal studies at the Weatherhead School of Management of Case Western Reserve University have shown that people can develop the behavioral level of EI (and SI). They reported behavioral improvements of 60–70% during the 1–2 years of the full-time MBA program, 55–65% improvement during the 3–5 years of the part-time MBA program, and then leveling off at about 50% improvement during the 5–7 years after entering into the part-time MBA. A longitudinal study of four cohorts of the Professional Fellows Program, the executive program at WSOM, showed that the 45-to 55-year-olds improved on 67% of the EI competencies. Other research has suggested that few 4-year undergraduate programs significantly help their students develop EI, unless it is a specific objective.

Research on EI development in 5-to 18-year-olds is followed closely by the Collaborative for Academic, Social, and Emotional Learning at the University of Illinois at Chicago. A meta-analysis of studies of programs instructing students on how to recognize and handle their emotions found students had better grades; were less out of school; showed less negative, antisocial behavior; and even performed substantially better on standardized tests.

EI is used in human resource management and development systems in organizations and in education to help develop the whole person. Assessment and feedback on EI, especially with the behavioral approach, can help individuals see how their EI appears to others.

Richard E. Boyatzis

See also *g* Theory of Intelligence; Intelligence Quotient; Intelligence Tests;

Multiple Intelligences, Theory of; Socio-Emotional Learning

Further Readings

Bar-On, R. (1997). Bar-on emotional quotient inventory: Technical manual. Toronto, Canada: Multi-Health Systems.

Boyatzis, R. E. (2009). A behavioral approach to emotional intelligence. Journal of Management Development, 28(9), 749–770.

Boyatzis, R. E., Gaskin, J., & Wei, H. (2015). Emotional and social intelligence and behavior. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts (pp. 243–262). New York, NY: Springer.

Boyatzis, R. E., & Goleman, D. (1996). Emotional competency inventory. Boston, MA: The Hay Group.

Boyatzis, R. E., & Goleman, D. (1999). Emotional competency inventory. Boston, MA: The Hay Group.

Boyatzis, R. E., & Goleman, D. (2002). Emotional competency inventory. Boston, MA: The Hay Group.

Boyatzis, R. E., & Goleman, D. (2007) Emotional competency inventory. Boston, MA: The Hay Group.

Emmerling, R. J., & Boyatzis, R. E. (2012). Emotional and social intelligence competencies: Cross cultural implications. Cross Cultural Management, 19(1), 4–18.

Goleman, D. (1995). Emotional intelligence. New York, NY: Bantam Books.

Mayer, J. D., Salovey, P., & Caruso, D. R. (1999). Emotional intelligence meets traditional standards for an intelligence. Intelligence, 2, 267–298.

Mayer J. D., Salovey P., Caruso D. R., & Sitarenios G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. Emotion, 3, 97–105.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. Imagination, Cognition and Personality, 9, 185–211.

Websites

Collaborative for Academic, Social, and Emotional Learning: www.casel.org

David Fetterman David Fetterman Fetterman, David

Empowerment Evaluation

Empowerment evaluation

580

584

Empowerment Evaluation

Empowerment evaluation involves the use of evaluation concepts, techniques, and findings to foster improvement and self-determination within a program or organization. In 2005, Abraham Wandersman and colleagues described it as an approach that “aims to increase the likelihood that programs will achieve results by increasing the capacity of program stakeholders to plan, implement, and evaluate their own programs” (p. 27). In essence, empowerment evaluation is a tool to help people produce desired outcomes and reach their goals. This entry discusses the conceptual building blocks of empowerment evaluation and the role of empowerment evaluators.

Conceptual Building Blocks

Empowerment evaluation’s conceptual building blocks include theories, principles, concepts, and steps. Together, they link theory to practice.

Theories

An exploration into the theories guiding empowerment evaluation will help to illuminate the integral relationship between method and use in empowerment evaluation. The pertinent theories guiding empowerment evaluation are empowerment theory, self-determination theory, evaluation capacity building, process use, and theories of use and action. *Empowerment theory* is divided into processes and outcomes. This theory has implications for the role of the empowerment evaluator or facilitator, which can differ from that of a traditional evaluator. *Self-determination* is one of the foundational concepts underlying empowerment theory and it helps to detail specific mechanisms or behaviors that enable the actualization of empowerment. *Process use* represents much of the rationale or logic underlying empowerment evaluation in practice because it cultivates ownership by placing the approach in community and staff members' hands. Finally, the alignment of *theories of use and action* explain how empowerment evaluation helps people produce desired results.

Principles

The theoretical foundations of empowerment evaluation lead to specific principles required to inform quality practice. Empowerment evaluation principles provide a sense of direction and purposefulness throughout an evaluation. Empowerment evaluation is guided by 10 specific principles as identified by David Fetterman and Wandersman in 2005. They include:

1. Improvement—Empowerment evaluation is designed to help people improve program performance; it is designed to help people build on their successes and reevaluate areas meriting attention.
2. Community ownership—Empowerment evaluation values and facilitates community control; use and sustainability are dependent on a sense of ownership.
3. Inclusion—Empowerment evaluation invites involvement, participation, and diversity; contributions come from all levels and walks of life.
4. Democratic participation—Participation and decision making should be open and fair.
5. Social justice—Evaluation can and should be used to address social inequities in society.
6. Community knowledge—Empowerment evaluation respects and values community knowledge.
7. Evidence-based strategies—Empowerment evaluation respects and uses the knowledge base of scholars (in conjunction with community knowledge).
8. Capacity building—Empowerment evaluation is designed to enhance stakeholders' ability to conduct evaluation and to improve program planning and implementation.
9. Organizational learning—Data should be used to evaluate new practices, inform decision making, and implement program practices; empowerment evaluation is used to help organizations learn from their experience

(building on successes, learning from mistakes, and making midcourse corrections).

10. Accountability—Empowerment evaluation is focused on outcomes and accountability; empowerment evaluations function within the context of existing policies, standards, and measures of accountability; did the program or initiative accomplish its objectives?

Empowerment evaluation principles help evaluators and community members make decisions that are in alignment with the larger purpose or goals associated with capacity building and self-determination. The principles of inclusion, for example, remind evaluators and community leaders to include rather than exclude members of the community, even though fiscal, logistic, and personality factors might suggest otherwise. The capacity building principle reminds the evaluator to provide community members with the opportunity to collect their own data, even though it might initially be faster and easier for the evaluator to collect the same information. The accountability principle guides community members to hold one another accountable. It also situates the evaluation within the context of external requirements and credible results or outcomes.

Concepts

Empowerment evaluation concepts provide a more instrumental view of how to implement the approach. Key concepts are critical friends, cultures of evidence, cycles of reflection and action, communities of learners, and reflective practitioners. A critical friend is an evaluator who facilitates the process and steps of empowerment evaluation. They believe in the purpose of the program but provide constructive feedback designed to promote improvement. A critical friend helps to raise many of the difficult questions and, as appropriate, tells the hard truths in a diplomatic fashion. They help to ensure the evaluation remains organized, rigorous, and honest.

Empowerment evaluators help cultivate a culture of evidence by asking people why they believe what they believe. They are asked for evidence or documentation at every stage, so that it becomes normal and expected to have data to support one’s opinions and views. Cycles of reflection and action involve ongoing phases of analysis, decision making, and implementation (based on evaluation findings). It is a cyclical process. Programs are dynamic, not static, and require continual feedback as they change and evolve. Empowerment evaluation is successful when it is institutionalized and becomes a normal part of the planning and management of the program.

Empowerment evaluation is driven by a group process that involves a community of learners. The group learns from each other, serving as their own peer review group, critical friend, resource, and norming mechanism. Individual members of the group hold each other accountable concerning progress toward stated goals. Finally, empowerment evaluations help create reflective practitioners. Reflective practitioners use data to inform their decisions and actions concerning their own daily activities. This produces a self-aware and self-actualized individual who has the capacity to apply this worldview to all aspects of the individual’s life. As individuals develop and enhance their own capacity, they improve the quality of the group’s exchange, deliberation, and action plans.

Steps

There are many ways in which to implement an empowerment evaluation. In fact, empowerment evaluation has accumulated a warehouse of useful tools. The three-step approach and the 10-step Getting To Outcomes (GTO) approach to empowerment evaluation are the most popular tools in the collection.

Three-Step Approach

The three-step approach includes helping a group (1) establish its mission, (2) take stock of its current status, and (3) plan for the future. The popularity of this particular approach is in part a result of its simplicity, effectiveness, and transparency.

Mission

The group comes to a consensus concerning its mission or values. This gives group members a shared vision of what’s important to them and where they want to go. It anchors the group in common values.

Taking Stock

After coming to a consensus about the mission, the group evaluates its efforts. First, the empowerment evaluator helps members of the group generate a list of the most important activities required to accomplish organizational or the most important activities required to accomplish organizational or programmatic goals. The empowerment evaluator helps participants prioritize this list. The top 10 activities are selected. They represent the heart of part two of taking stock: rating.

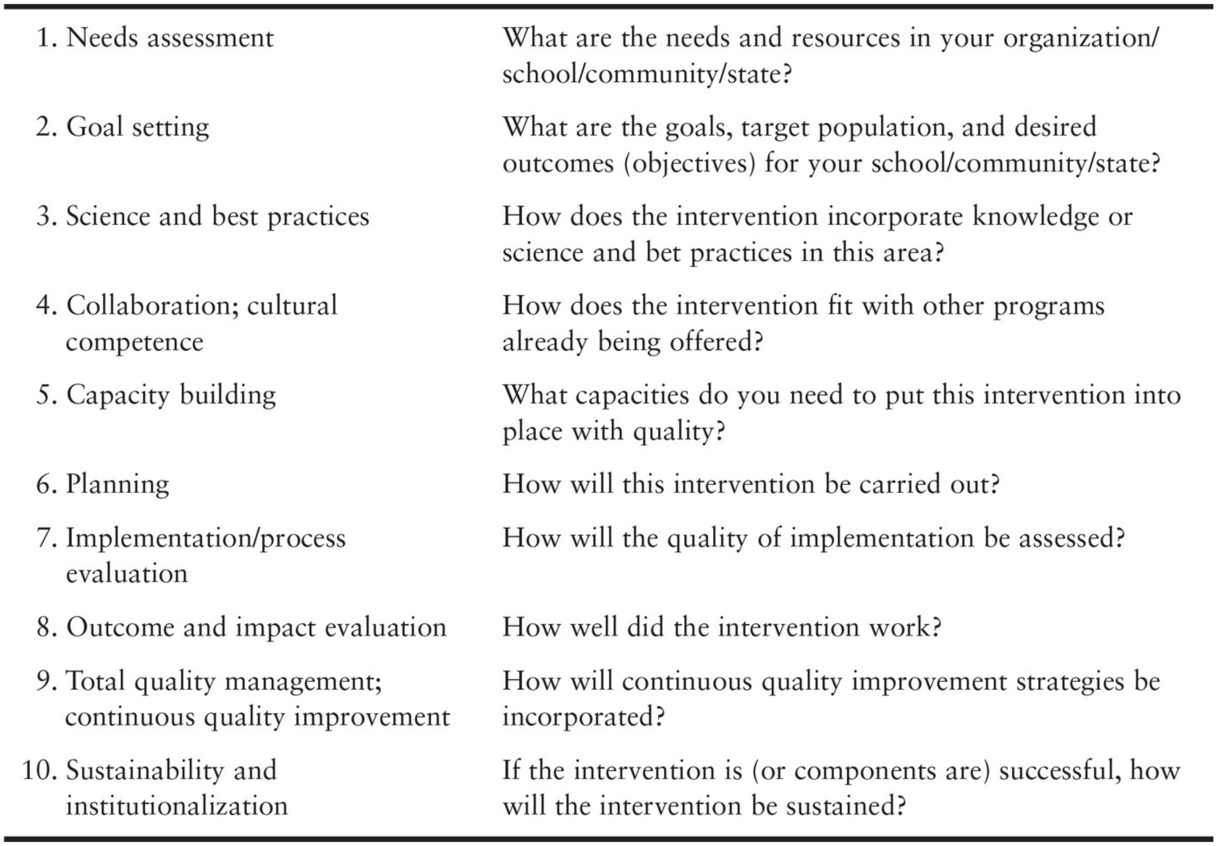
The empowerment evaluator asks participants in the group to rate how well they are doing concerning each of the activities selected, using a 1 (low) to 10 (high) scale. This provides the group with the basis for a meaningful dialogue, as group members discuss the reasons for their ratings. In addition to clarifying issues, evidence is used to support viewpoints and “sacred cows” are surfaced and examined. Moreover, the process of specifying the reason or evidence for a rating provides the group with a more efficient and focused manner of identifying what needs to be done next, during the planning for the future step of the process.

Planning for the Future

Many evaluations conclude at the taking stock phase. However, taking stock is a baseline and a launching off point for the rest of the empowerment evaluation. After rating and discussing programmatic activities, it is important to do something about the findings. It is time to plan for the future. This step involves generating goals, strategies, and credible evidence to determine whether the strategies are being implemented and whether they are effective. The goals are directly related to the activities selected in the taking stock step.

Ten-Step Approach

Table 1 presents the 10-step empowerment evaluation approach known as GTO. The questions are based on those discussed in *Getting To Outcomes 2004: Promoting Accountability Through Methods and Tools for Planning, Implementation, and Evaluation*, by Matthew Chinman, Pamela Imm, and Abraham Wandersman.



1. Needs assessment	What are the needs and resources in your organization/school/community/state?
2. Goal setting	What are the goals, target population, and desired outcomes (objectives) for your school/community/state?
3. Science and best practices	How does the intervention incorporate knowledge or science and bet practices in this area?
4. Collaboration; cultural competence	How does the intervention fit with other programs already being offered?
5. Capacity building	What capacities do you need to put this intervention into place with quality?
6. Planning	How will this intervention be carried out?
7. Implementation/process evaluation	How will the quality of implementation be assessed?
8. Outcome and impact evaluation	How well did the intervention work?
9. Total quality management; continuous quality improvement	How will continuous quality improvement strategies be incorporated?
10. Sustainability and institutionalization	If the intervention is (or components are) successful, how will the intervention be sustained?

Note: Adapted from Table 2.1, p. 36, in Fetterman, D. M. (2015).

Conventional and innovative evaluation tools are used to monitor the strategies used in the evaluation, including online surveys, focus groups, and interviews. In addition, program-specific metrics are developed, using baselines, benchmarks or milestones, and goals. For example, an empowerment evaluation for a tobacco prevention program in Arkansas established:

1. Baselines (the number of people using tobacco in the community)
2. Goals (the number of people expected to stop using tobacco by the end of the year as a result of the program)
3. Benchmarks or milestones (the number of people expected to stop using tobacco each month or quarter as a result of the program)
4. Actual performance (the actual number of people who were helped to stop using tobacco each month or quarter throughout the year)

These metrics are used to help a community monitor program implementation efforts and enable program staff and community members to make midcourse corrections and substitute ineffective strategies for potentially more effective ones as needed. These data are also invaluable when the group conducts a second taking stock exercise (3–6 months later) to determine whether it is making progress toward its desired goals and objectives. Additional metrics enable community members to compare, for example, their baseline assessments with their benchmarks/milestones or expected points of progress, as well as their goals. In addition, empowerment evaluations are using many other tools, including photo journaling, online surveys, virtual conferencing formats, blogs, shared web documents and sites, infographics and data visualization, and creative youth self-assessments.

Role of Evaluator

Empowerment evaluators differ from many traditional evaluators. Empowerment evaluators are not in charge. The people they work with are in charge of the direction and execution of the evaluation. Empowerment evaluators are critical friends or coaches. They believe in the merits of a particular type of program but they pose the difficult questions (in a diplomatic fashion).

Empowerment evaluators are trained evaluators with considerable expertise. However, they listen and rely on the group's knowledge and understanding of their local situation. The critical friend is much like a financial advisor or personal health trainer. Important attributes of a critical friend are creating an environment conducive to dialogue and discussion; providing or requesting data to inform decision making; facilitating rather than leading; and being open to ideas, inclusive, and willing to learn.

David Fetterman

Note: Adapted from Fetterman, D. M. (2015). Empowerment evaluation theories, principles, concepts, and steps, in D. M. Fetterman, S. J. Kaftarian, … A. Wandersman, editors, *Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability*, Second edition (pp. 20–42). Thousand Oaks, CA: Sage.

See also Collaborative Evaluation; Participatory Evaluation; Program Evaluation; Stakeholders

Further Readings

Chinman, M., Imm, P., & Wandersman, A. (2004). Getting to outcomes 2004:

Promoting accountability through methods and tools for planning, implementation, and evaluation. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/technical_reports/TR101/

Dunst, C. J., Trivette, C. M., & LaPointe, N. (1992). Toward clarification of the meaning and key elements of empowerment. Family Science Review, 5(1…2), 111–130.

Fetterman, D. M. (2001). Foundations of empowerment evaluation. Thousand Oaks, CA: Sage.

Fetterman, D. M. (2005). Empowerment evaluation: From the digital divide to academic distress. In D. M. Fetterman & A. Wandersman (Eds.), Empowerment evaluation principles in practice. New York, NY: Guilford.

Fetterman, D. M. (2015). Empowerment evaluation theories, principles, concepts, and steps. In D. M. Fetterman, S. J. Kaftarian, & A. Wandersman (Eds.), Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability (2nd ed., pp. 20–42). Thousand Oaks, CA: Sage.

Fetterman, D. M., & Bowman, C. (2002). Experiential education and empowerment evaluation: Mars rover educational program case example. Journal of Experiential Education, 25(2), 286–295.

Fetterman, D. M., & Wandersman, A. (2005). Empowerment evaluation principles in practice. New York, NY: Guilford.

Wandersman, A., Snell-Johns, J., Lentz, B., Fetterman, D. M., Keener, D. C., Livet, M., & Flaspohler, P. (2005). The principles of empowerment evaluation. In D. M. Fetterman & A. Wandersman. Empowerment evaluation principles in practice. New York, NY: Guilford.

Zimmerman, M. A. (2000). Empowerment theory: Psychological, organizational, and community levels of analysis. In J. Rappaport & E. Seldman (Eds.), Handbook of community psychology (pp. 2–45). New York, NY: Kluwer Academic/Plenum.

Alison L. Bailey Alison L. Bailey Bailey, Alison L.

English Language Proficiency Assessment English language proficiency assessment

584

589

English Language Proficiency Assessment

Millions of individuals worldwide participate in the assessment of their comprehension and production of English. These learners range from young students with non-English-speaking backgrounds enrolled in English-speaking schooling systems to adult learners who are enrolled in English-as-a-second or -foreign language courses. Additionally, assessment of adults may measure English for specific purposes such as higher education, occupational or skilled migration, and citizenship eligibility. This entry focuses on the purposes that English language proficiency (ELP) assessment serves, commonly used ELP assessments and the test takers they serve, and several challenges facing ELP assessment development and measurement.

Purposes of ELP Assessment

Large-scale direct assessment of ELP takes a census of the language skills and knowledge of English learners and is commonly used for accountability and enrollment eligibility purposes. Accountability uses determine initial classification and program placement, monitor language progress, and reclassify learners as English proficient. Program funding is frequently attached to the abilities of schooling systems to demonstrate learner proficiency. Enrollment eligibility uses of ELP assessment determine whether learners have attained a sufficient level of English proficiency to be admitted to a course of study that has English as the language of instruction. Each function is described in further detail in this section.

Classification and Program Placement for English

Support/Instruction

ELP assessments used for classification must at minimum be able to discriminate between test takers who are already proficient in English and those who are not to provide appropriate language support services or placement into English language courses at commensurate difficulty levels. For example, under the No Child Left Behind Act of 2001 and the legislation that replaced it, the Every Student Succeeds Act of 2015, all students whose home language backgrounds suggest an influence of a language other than English must be initially screened using a state-approved ELP assessment based on English proficiency standards in four language domains: listening and reading (i.e., receptive capabilities) and speaking and writing (i.e., expressive capabilities). In some states, this evaluation is conducted with a full-form annual summative ELP assessment although most use a short-form ELP screener for initial classification purposes.

Monitoring Progress in English

ELP assessment is also used to measure progress in the acquisition of English language skills and knowledge. For accountability reporting purposes, this is typically conducted with a large-scale summative ELP assessment of the four language domains to capture growth in oral and written proficiencies. Scale scores are converted to between three and six levels of increasing accomplishment with performance descriptors outlining skills in English at each proficiency level.

Progress can also be measured by other forms of assessment such as classroom-based assessment, including dynamic, portfolio, and performance assessment approaches that involve learners in role play, oral presentations, writing activities, and long-term collaborative projects. Schools increasingly use formative assessment or learning-oriented assessment approaches that do not involve “scoring” students’ English proficiency but rather involve self-monitoring by learners to achieve language goals and observation by teachers to find the “best fit” along a language learning progression to provide feedback to both the student and teacher for modifying learning and instruction.

Reclassification to English Proficient

ELP assessment also plays a role in the reclassification of English learners to English proficient status (i.e., no longer in need of English support services). However, in elementary and secondary schooling contexts at least, such decision making frequently includes how successfully a learner is also achieving academically in language arts/reading and even mathematics. This situation can lead to complex models of performance standards and decision rules for reclassification. At their simplest, models utilize equal weighting of the four language domains in a composite score, while a more complex model might weight the literacy domains more highly and/or utilize an aggregation of ELP and academic achievement assessments. This leads to a greater number of standards to be met to exit than to originally enter a system of language support services.

Enrollment Eligibility

ELP assessment can function to identify students who have reached a level of English proficiency that enables them to access content taught in English. Typically, this use of ELP assessment is made at English-speaking institutions of higher education for making enrollment decisions with students from non-English-speaking countries or non-English-speaking background students. English proficiency status may also be used for enrollment eligibility with advanced placement or college preparatory courses in secondary schools.

Commonly Used ELP Assessments and Target ELP Test Takers

For adult English learners, two commercial ELP assessments dominate. In Australia, Canada, New Zealand, and the United Kingdom, the International English Language Testing System (Cambridge English Language Assessment) is designed to assess the necessary English proficiency for pursuing skilled immigration to, or higher education and work in, English-speaking countries. In the United States, the Test of English as a Foreign Language (Educational Testing Service) is primarily designed to measure the English language demands of “everyday academic settings” but is also used to assess whether test takers possess the proficiency necessary for skilled immigration and work purposes.

For school-age learners, the suite of tests in the Cambridge English: Young Learners series and the Junior and Primary versions of the Test of English as a

Foreign Language focus on the EFL needs of school-age learners. Many young learners from a non-English-speaking background being raised in English-speaking countries are acquiring English as a second or additional language. For example, approximately 5 million school-age students in the United States alone are identified as English learners and will be increasingly assessed with “next generation” ELP assessments that have been newly designed to determine progress in and attainment of English proficiency that matches the more challenging English language demands of new college and career-ready standards for academic content.

Most U.S. states have moved from commercial “off-the-shelf” assessments to comply with federal legislation by adopting the ELP standards-based assessments of a state consortium (e.g., the multistate WIDA or ELPA21 consortia) or by developing their own assessment (e.g., the English Language Proficiency Assessments for California). However, commercial ELP assessments such as LAS Links (DRC-CTB) are still used for some accountability and placement purposes.

ELP Assessment Development and Measurement Challenges

This discussion of assessment development and measurement challenges focuses on operationalizing the ELP construct, addressing intersections with academic content, setting the standard for English proficient, and extending target language uses (TLU) of ELP assessments with consequences for assessment development.

The ELP Construct

Language researchers are far from unanimous in deciding what aspects of the ELP construct should and can be represented on summative ELP assessments. The field has largely adopted Lyle Bachman’s assessment use argument (i.e., linking assessment performance to assessment use) as an interpretative framework for evaluating technical adequacy. ELP assessment intended for school-age learners has largely been designed to measure language constructs inherent in ELP standards that are used to guide language instruction and assessment. As mentioned, in the United States, the most recent incarnation of many language standards were written to match the language demands of the college and career ready standards. This was assisted by such initiatives as the development of the *Framework for English Language Proficiency Development Standards Corresponding to the Common Core State Standards and the Next Generation Science Standards* (English Language Proficiency Development Framework), which was published in 2012.

The English Language Proficiency Development Framework highlights an emphasis in academic standards on communication of academic concepts and collaborative learning contexts. Psychometric analyses of prior assessments suggest a unitary ELP construct (i.e., general English proficiency) but with new assessments built around academic language demands, future analyses may reveal additional dimensions. It is questionable, however, whether a large-scale summative form of ELP assessment alone can adequately measure these far more contextualized aspects of proficiency even with technology-enhanced items, hence the need to promote classroom-based ELP assessment. The latter offers multiple measures in multiple contexts to supplement summative inferences about learner performance.

With an almost exclusive emphasis placed on ELP assessment for educational accountability with young learners, items measuring interpersonal and intrapersonal nonacademic uses of language have been curtailed. In contrast, adult ELP assessment has a broader range of TLU (i.e., education, work, skilled immigration, citizenship, and even personal growth through language learning). Test developers have addressed measurement of relevant language constructs by using corpus linguistics (e.g., databases of authentic college lectures, textbooks, leisure reading materials) in the selection of TLU for adult ELP assessment. School-age corpora for assessment purposes are only just beginning to emerge and offer new challenges in ELP assessment designed for school-age learners.

Intersections WithAcademic Content Knowledge

While the content of ELP assessments should not require academic knowledge on the part of test takers, the content of the assessment should still make the same language demands on the test taker as those found in educational contexts (if the TLU is chosen to assist educators with stated educational purposes such as classification or placement). Ideally listening and reading passages and oral language and writing prompts used to test comprehension and elicit language should not presuppose knowledge of subjects such as science and mathematics.

This concern with the relevancy of the construct is matched in the academic content assessment arena with a similar tension between content and language: A mathematics assessment turns into an ELP assessment when the linguistic complexity becomes a barrier to students demonstrating their mathematics knowledge (i.e., construct irrelevant).

Assessment researchers have also acknowledged the unavoidable reciprocal nature of language and academic content; measures of language proficiency must be about something if they are to yield useful information (i.e., not just tests of decontextualized parts of speech), and academic content knowledge is most often learned and displayed *through* the medium of language. Furthermore, content knowledge and language intersect in ways that make the separate assessment of each less meaningful with evidence that complexity of content influences language performance. Consequently, there are proposals to extract a language proficiency dimension from academic content assessments, treating the single assessment as a multidimensional measure and yielding, for example, both a mathematics and a "language of mathematics" score.

Setting the Standard for English Proficient

Establishing a norm for and providing evidence of English proficient status have proven elusive in ELP assessment. With proficiency most frequently expressed in levels along a continuum, there has been a movement away from attaining native-like skills and knowledge as demonstration of the highest proficiency level. For example, the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* developed by the Council of Europe does not invoke native-or near Epistemologies, Teacher, and Student native-speaker competencies in the performance descriptor of a proficient user who has mastery of a new language. Future research in this area needs to address fundamental questions such as *How much progress is reasonable to expect on an annual basis?* and *How is progress influenced by current level of proficiency, domain of language, instruction, and learner characteristics (e.g., age, motivation)?*

Analytical methods such as decision consistency and logistic regression techniques have recently been explored to make empirically guided reclassification decisions for school-age English learners, including examining the convergence of ELP performance and academic performance. Such efforts by Gary Cook and others can help determine at the levels of proficiency wherein by Gary Cook and others can help determine at the levels of proficiency wherein English no longer restricts student performance on concurrent academic assessments. There are additional concerns involving equivalency across different jurisdictions (i.e., states) in how (re)classification models/decision rules are applied. Patty Carroll and colleagues have pointed out that differences in weighting language domains or aggregating two or more assessments lead to undesirable differences in setting the standard for English proficiency.

Extending ELP TLU andConsequences for ELP Assessment

There have been efforts to extend ELP assessment use to native speakers of English, both adults and children who are considered standard English learners. Such speakers commonly use a variety of English that differs from the socially dominant variety and consequently, it is argued, may benefit from formal English language instruction, especially English for academic purposes. The questions are whether the ELP construct is the same for native speakers as for non-native speakers and whether existing ELP assessments can effectively assist with standard English learners instructional placement, progress monitoring, and determining English proficiency. Moreover, the use of English as a lingua franca beyond the boundaries of English-speaking countries also gives rise to questions about the ELP construct. For example, *Is the ELP construct sufficiently comprehensive in its currently operationalized form to measure the communicative knowledge and skills that lingua franca interactions include?*

Extensions of ELP assessment use to native speakers or to countries where English is not the primary language can disrupt the traditional alignment, in the adult English-as-a-second or -foreign language arena especially, between ELP proficiency frameworks such as the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*; the content of ELP curricular materials, including instructional programs and textbooks; and ELP assessment. This closed system may have led to a specific conceptualization of the ELP construct—one that can be readily assessed by a large-scale summative assessment but not one that captures the broad interactive contexts of learners' lived experiences.

Alison L. Bailey

See also Common Core State Standards; Every Student Succeeds Act; Literacy;

No Child Left Behind Act; Proficiency Levels in Language; Written Language Assessment

Further Readings

Bachman, L. F. (2014). Ongoing challenges in language assessment. *The companion to language assessment*. In A. J. Kunnan (Ed.), The companion to language assessment. Hoboken, NJ: Wiley Blackwell.

Bailey, A. L., & Carroll, P. (2015). Assessment of English language learners in the era of new academic content standards. Review of Research in Education, 39, 253–294.

Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. TESOL Quarterly, 48(3), 480–506.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., & Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus (TOEFL Monograph Series No. MS-25). Princeton, NJ: Educational Testing Service.

Carroll, P., & Bailey, A. L. (2016). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. Language Testing, 33(1), 23–52.

Chase, K. B., & Johnston, J. R. (2013). Testing local: Small-scale language sample databases for ESL assessment. Canadian Journal of Speech-Language Pathology … Audiology, 37(1), 42–57.

Council of Chief State School Officers. (2012). Framework for English Language Proficiency Development Standards corresponding to the Common Core State Standards and the Next Generation Science Standards. Washington, DC: Author.

Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the U.S. Language Assessment Quarterly, 11(4), 432–457.

Linquanti, R., Cook, H. G., Bailey, A. L., & MacDonald, R. (2016). Moving toward a more common definition of English learner: Collected guidance for states and multistate assessment consortia. Washington, DC: Council of Chief State School Officers.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. Language Testing, 24(4), 489–515.

Gu, L., Lockwood, J., & Powers, D. E. (2015). Evaluating the TOEFL Junior® Standard Test as a measure of progress for young English language learners. ETS Research Report Series, 2, 1–13.

McNamara, T. (2012). English as a lingua franca: The challenge for language testing. Journal of English as a Lingua Franca, 1(1), 199–202.

Murray, N. (2013). Widening participation and English language proficiency: A convergence with implications for assessment practices in higher education. Studies in Higher Education, 38(2), 299–311.

Wolf, M. K., & Faulkner-Bond, M. (2016). Validating English language proficiency assessment uses for English learners: Academic language proficiency and content assessment performance. Educational Measurement: Issues and Practice, 35(2), 6–18.

William A. Sandoval William A. Sandoval Sandoval, William A.

Epistemologies, Teacher, And Student Epistemologies, teacher, and student

Epistemologies, Teacher, And Student

“Personal epistemology” refers to the beliefs that people hold about knowledge and knowing; the psychological study on this topic started with the seminal work of William Perry in 1970. It seemed plausible that the ideas students have about the nature of knowledge and how one comes to know something influence the learning strategies that they use, while the epistemologies of teachers seemed likely to influence how they teach. This entry further defines personal epistemology and discusses research connecting students’ and teachers’ epistemological beliefs with teaching and learning.

Personal epistemology researchers concentrate on two broad dimensions of epistemology: (1) the nature of knowledge and (2) the nature of knowing. The nature of knowledge is conceptualized in terms of beliefs about its simplicity (simple vs. complex) and about its credibility status (certain vs. tentative). The nature of knowing is conceptualized in terms of the source of knowledge (internal or external to the knower) and the means of justification (authority vs. evidentiary standards). Personal epistemology researchers generally argue that a developmental progression exists across the life span, wherein individuals start from an absolut-ist stance that sees knowledge as simple, knowable with certainty, as having its source in the world, and justified by trusted authorities; later they move to holding an unmoored multiplism (or relativism) in which knowledge is regarded as uncertain, supposedly authoritative sources are untrustworthy, and all knowledge claims are equally justifiable. Later in development, this multiplism is resolved into an evaluative stance that concedes that knowledge is constructed and is not knowable with absolute certainty, but that nevertheless asserts that knowledge claims can be justified according to standards of reason and evidence.

Research has shown that there is some degree of association between

Research has shown that there is some degree of association between epistemological beliefs and learning strategies, school achievement, and course-taking patterns. There is a tendency for students who have adopted the evaluative stance to have higher achievement, to take more math and science courses, and to use deeper learning strategies. At the same time, however, it must be acknowledged that clear and direct associations between professed epistemic beliefs and students' learning in subjects such as science or math have been hard to come by, and studies have faced a range of problems of measurement and conceptualization.

Research on teacher epistemologies has largely focused on associations between epistemological beliefs and other kinds of beliefs about teaching or learning. Compared with studies of student epistemologies, research on teachers is limited. Within math and science specifically, the general finding is that the teachers across K–12 grade levels tend to have what researchers consider naive views of the epistemology of their subject specialties. As yet, relatively little work has been done to trace the influence of these views on teaching practices. There is some empirical suggestion that myriad concerns and in-the-moment judgments have a much stronger effect on instructional practices than epistemological beliefs.

Research on personal epistemologies has been hampered by a rather large variety of definitions of what counts as "epistemological." Models of epistemological development proliferate, so far with little effort to discriminate among them. Connections between relevant developmental milestones, such as attainment of the ability to engage in causal reasoning, or development of the child's theory of mind, are underexplored. Questions remain concerning how an individual's beliefs about knowledge and knowing are related to the individual's beliefs about learning and whether the latter should be considered part of a personal epistemology.

A persistent concern in personal epistemology research has been the reliance on general survey instruments that lack validation with other pos-sible assessments of epistemological belief. Research subjects are typically asked to state their level of agreement with general statements about knowledge or about knowing —but such assessments are far removed from people's actual efforts to construct or evaluate knowledge for themselves and presuppose that individual's epistemic beliefs are stable and available for explicit reflection. A related problem is that commonly used instruments often include items about topics that bear little relation to epistemology.

There is another significant issue, namely, that a wide variety of empirical evidence undermines claims that there is a simple developmental trajec-tory from absolutism to evaluativism. People can espouse apparently contradictory epistemologies at the same time, both within and across subject matter or judgment domains. The assignment of individuals to broad epistemological positions may reflect researchers' biases more than the actual beliefs of the people concerned. This has spawned a variety of theoretical models of epistemological development; these include developmental theories as described here and models that posit multidimensional, somewhat independent belief systems. At the moment, the field appears to be in ferment without a clear way of discriminating between competing models.

A related issue is that the dominant conceptualization of epistemological beliefs as described above seems simplistic both intuitively and philosophically. Intuitively, it is not hard to recognize that some knowledge is simple, such as knowing your own phone number, but other knowledge is complex, such as knowing the theory of natural selection. The epistemological status of the first is different from that of the second—and this makes it difficult for students to give a single, universally applicable account of the nature of their beliefs. Philosophically, epistemologists concern themselves with a much broader range of issues than is typical in personal epistemology research, and in particular emphasize the aims of knowing, and the role played by values in epistemological matters.

Models of epistemological development thus are making efforts to be more philosophically rigorous, and investigative methods are changing to enable comparisons between what researchers now distinguish as professed epistemologies (what people say that they believe about knowledge and knowing) and enacted epistemologies (what people do when they construct and evaluate knowledge themselves). This includes a shift away from assessments of beliefs toward the study of processes of epistemic cognition. This shift stems in part from research on learning in the disciplines, especially math and science. In science, for example, the evidence is quite clear that students' efforts to investigate scientific questions (enacted epistemologies) share much with professional scientific practice, while their professed epistemological beliefs about science seem hopelessly naive and immune to instruction. It remains an open question how the intuitive and apparently tacit ideas students apply to their own knowledge construction can be developed into explicit concep-tions of the epistemologies of particular disciplines of science, mathematics, and others.

William A. Sandoval

Note: Adapted from Sandoval, W. A. (2014). Epistemologies, teacher and student. In D. C. Phillips (Ed.), *Encyclopedia of educational theory and philosophy* (Vol. 1, pp. 284–286). Thousand Oaks, CA: Sage.

See also Metacognition; Objectivity; Positivism; Postpositivism

Further Readings

Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. Educational Psychologist, 46(3), 141–167.

Hofer, B. K., & Bendixen, L. D. (2012). Personal epistemology: Theory, research, and future directions. In K. R. Harris, S. Graham, & T. Urdan (Eds.), Theories, constructs, and critical issues: Vol. 1. APA educational psychology handbook (pp. 227–256). Washington, DC: American Psychological Association.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. Review of Educational Research, 67(1), 88–140.

Perry, W. G. Jr. (1970). Forms of intellectual and ethical development in the college years. New York, NY: Holt, Rinehart … Winston.

Peter M. Bentler Peter M. Bentler Bentler, Peter M.

EQS

EQS

590

593

EQS

EQS is a statistical software package distributed by Multivariate Software for producing and analyzing structural equation models. Two main mathematical systems exist for specifying linear structural equations models (SEM), the LISREL, and Bentler–Weeks (BW) conceptualizations. In the LISREL approach, measurement and simultaneous equations models are strictly separated. The former relates observed to latent variables, while the latter relates latent variables to each other. In the BW approach, implemented in EQS, any type of SEM model is simply a set of equations, hence the name EQS. This entry reviews how EQS handles equations and variances or covariances, details the model and output files of EQS, describes the Diagrammer and Build_EQS functions, explains how the BW model is set up in EQS and how statistical methods are utilized in EQS, and lists the available versions of EQS.

Equations and Variances–Covariances

The LISREL and BW models involve matrix equations and require matrix algebra. This is a challenge in many educational settings and is avoided in many LISREL-type implementations by the use of simple interfaces. Unfortunately, these interfaces can obscure the actual model being run. This difficulty is avoided in EQS, where any SEM model is specified via equations involving *V, F, E*, and *D* variables, and the exact model being run is always visible. *V*s represent observed data variables that are numbered as ordered in the data file: *V*1, *V*2, and so on. The rest are hypothetical generating variables—latent factors (*F*s), residuals in equations for *V*s (*E*s), and residuals in equations for *F*s, namely, Ds (for disturbance). If a model has three Fs, typically these are numbered $F1$, $F2$, and $F3$. The number in an E or D corresponds to its V or F. In EQS, the dependent variable in any equation—the left side of an equation—can only be a V or F variable. Illustrative equations are $V3 = {*}F1 + E3$ and $F5 = {*}F1 + {*}F2 + {*}V6 + D5$, where the “*” are unknown parameters. Notice that in the $F5$ equation, predictors $F1$ and $F2$ are latent factors, while $V6$ is an observed variable—such mixed predictors are simple in EQS but difficult in LISREL-type models. According to the BW model, the parameters of any SEM are the coefficients in its equations (the “*” in the examples) and the variances and covariances of independent variables, which are variables that never are dependent variables in any equation. Structured means models, such as growth curve models, also have the means of independent variables as parameters. The remaining discussion focuses on models without mean parameters.

Model and Output Files

Equations and variance/covariance specifications can be typed into a model file text file (called an eqs file) and then submitted to EQS for immediate processing. The output will be a text file (called an out file) with optimal parameter estimates as well as many types of statistics and indices indicating the adequacy of the model. Examples of output include case numbers that contribute to multivariate normality and a test on multivariate normality, information on the identification status of the parameters, standard errors of parameter estimates, model fit indices such as the comparative fit index and root-mean-square error of approximation, and standard and robust chi-square statistics for evaluating the model’s ability to explain the observed sample variances and covariances.

To illustrate an EQS setup, in the simple mediation model $V1 \rightarrow V2 \rightarrow V3$, there are two equations and the model file would contain this specification.

```
/EQUATIONS

V2 = *V1 + E2

V3 = *V2 + E3

/VARIANCES
```

*V*1=*; *E*2=*; *E*3=*

Just as any regression equation has a residual term, Residuals E2 and E3 are added to the equations even though residuals are not the focus of the model. There are no covariances here, so there is no /COVARIANCE section. There are five "*" terms listed, so there are five parameters to be estimated. A 3 × 3 sample covariance matrix of *V*1 − *V*3 would be required as input data. Such a matrix has 3(4)/2 = 6 different data elements, so the model has 6 − 5 = 1 degree of freedom. This tests the absence of a direct connection between *V*1 and *V*3.

Diagrammer

A proper path diagram carries exactly the same information as equations and variances–covariances of independent variables (many published path diagrams misrepresent their models by omitting key information such as residual variables), and hence an SEM beginner might prefer to specify a model visually rather than with equations. Diagrammer offers tools that allow one to put many *V, F, E,* and *D* variables on a page and to easily connect them with unidirectional or bidirectional arrows (i.e., to create a SEM model). Each *V* or *F* having a one-way arrow aiming at it is a dependent variable will get its own equation. Each such path is a coefficient akin to a β coefficient in regression; this is typically an unknown parameter. Bidirectional arrows among independent variables represent covariances, which are implicit in regression but essential in SEM. EQS checks that mistakes are not made during this process (e.g., that only independent variables are connected by two-way arrows). When the path diagram is completed, EQS starts the Build_EQS procedure. It allows one to immediately create a model file and run it or to specify additional details such as Lagrange multiplier tests.

Build_EQS

Models can also be built without Diagrammer. The *Specifications* dialog box allows specification of the data file, method of estimation, and options such as a multilevel, multiple group, or structured means model as well as options for specifying categorical variables, missing data, or deleting cases. Robust methods, including case-robust methods, can be specified. Statistical options are further discussed in the following paragraphs.

The *Build Equations* and *Build Variances/Covariances* dialog boxes allow specification of latent factors, if any, and show matrices in graphical form that represent equations for dependent variables and variances/covariances for independent variables. Simple operations like clicking on a cell entry can indicate a free parameter, and click-and-drag can be used to specify a set of free parameters. When these two dialogs are completed, the model is created, and it can be immediately run and the output examined.

Of course, more complicated choices and options can be made with additional dialog boxes, if desired. For example, equality and inequality constraints on parameter values can be specified. Lagrange multiplier tests can be specified to evaluate the plausibility of restrictions on fixed parameters and on equality constraints (such as cross-group constraints). Wald tests can be set up to evaluate the necessity of sets of free parameters. Various printing options, such as that of indirect and total effects, can be chosen, and technical aspects, such as control of the convergence criterion, specified. Specialized output and data-saving options are available. A simulation methodology with a variety of options can be specified. Various choices for computing reliability coefficients and performing an exploratory factor analysis can be specified. The computer’s memory allocation also can be modified.

BW Model

Although an EQS user can set up, run, and evaluate models without knowing the technical details of the BW model, model specifications are set up internally via the BW matrix equation $\eta=\beta\eta+\gamma\xi$. All the *V*s and *F*s that are dependent variables (i.e., have one-way arrows aiming at them and are on the left-side of equations) are placed into the vector of dependent variables η. The other *V*s and *F*s, and all *E*s and *D*s, are independent variables and are placed into the vector ξ. Coefficients or paths from η to η variables are placed into b, and those from ξ to η are placed into γ. The variances and covariances of independent variables are placed into covariance matrix Φ. The three BW parameter matrices allow all types of SEM, including path analysis, confirmatory factor analysis, simultaneous equation systems, LISREL models, and a variety of models that require special treatment with dummy variables in other programs.

The BW model is more general than ordinary regression because it allows one to have latent variables (*F*s) as dependent variables as well as predictors and furthermore permits dependent variables to predict other dependent variables if desired. The BW model uniquely clarifies when a model contains latent variables. The BW matrix equation can equivalently be written as $\nu=B\nu+\Gamma\xi$, where $\nu'=(\eta'|\xi')'$ contains all variables in the model, the observed variables $x=G\nu$ are selected from all variables by a matrix Gwith known 0,1 elements, Bis a 2 × 2 supermatrix containing β and zero matrices elsewhere, and Γ is a 2 × 1 supermatrix containing γ and an identity matrix. Then the covariance structure of BW is given by $\Sigma=G(I-B)^{-1}\Gamma\Phi\Gamma'(I-B)^{-1'}\, G'$, where $\sum$ is the covariance matrix of the *V*s. A model is a latent variable model only if the rank of Φ exceeds that of $\sum$ (i.e., if the dimensionality of the independent variables exceeds the dimensionality of the observed variables).

Statistical Methods

For data that are multivariate normally distributed, the sample covariance matrix is a sufficient statistic and SEM reduces to covariance structure analysis. For this situation, EQS allows three estimation methods: least squares, generalized least squares, and reweighted least squares, which is equivalent to maximum likelihood. Of course, real data are rarely normal, and when that is so, these methods yield test statistics and standard errors that are misleading. EQS provides a number of methods that work under violation of normality, all of which require raw data input. The most widely known and accepted are the Satorra–Bentler scaled and adjusted test statistics that are obtained by simply adding the word “ROBUST” after the chosen method (e.g., METHOD=ML,ROBUST). When this is done, a series of additional tests—the several residual-based test statistics developed by Ke-Hai Yuan and Peter Bentler—are also computed and printed.

When data are nonnormal and the sample size is huge, the asymptotically distribution-free method is optimal. It is called arbitrary distribution generalized least square in EQS. An important feature of EQS is to provide Yuan–Bentler corrections to asymptotically distribution-free statistics to deal with smaller samples sizes. EQS also uniquely provides methods to handle special situations where the form of nonnormality is known to be either elliptical or with heterogeneous kurtoses. In such cases, the provided methods are far more efficient than the asymptotically distribution-free method.

EQS identifies cases that are outliers and that contribute to multivariate kurtosis. In addition to outlier removal, EQS allows the user to smoothly downweight

In addition to outlier removal, EQS allows the user to smoothly downweight cases that are outlying with its case-robust specification. This allows modeling of weighted means and covariances.

In the case of one-group models, as in factor analysis, it may be desirable to model the correlation matrix rather than the covariance matrix. EQS provides correlation structure methods for normal and nonnormal data that parallel those in covariance structure analysis.

Categorical variables are handled with SEM for polychoric/polyserial correlations as well as odds ratio-based estimates of polychoric correlations. The methodology for continuous variables is all adapted to ordinal variables, as are corrections to nonnormality, to yield correct inference (e.g., ME=LS,ROBUST).

Versions

EQS is available from Multivariate Software under individual, class, or university-wide licenses for Windows, Mac, and Linux operating systems. While the three versions process eqs files identically and provide the same output, only the Windows program has the Diagrammer and Build EQS procedures as well as a wide variety of non-SEM graphical and statistical modules. Free trial versions are available. The programs are accompanied by a manual and a supplement that describe how to use the program and interpret its output. They also document the technical basis and sources for its statistics.

Peter M. Bentler

See also LISREL; Structural Equation Modeling

Further Readings

Bentler, P. M. (2008). EQS 6 structural equations program manual. Temple City, CA: Multivariate Software.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. Psychometrika, 45, 289–308.

Bentler, P. M., & Wu, E. J. (2015). Supplement to EQS 6.3 for Windows user’s guide. Temple City, CA: Multivariate Software.

Jöreskog, K. G., & Sörbom, D. (1984). LISREL VI user’s guide. Mooresville, IN: Scientific Software International.

Won-Chan Lee Won-Chan Lee Lee, Won-Chan

Hyung Jin Kim Hyung Jin Kim Kim, Hyung Jin

Equating

Equating

593

598

Equating

The term *equating* refers to a statistical process used to establish comparable scores on alternate test forms built to the same test specifications. A test form is a collection of items or tasks intended to measure examinees' performance on a set of predefined domains of a test. Most large-scale testing programs have multiple alternate test forms, each of which is developed according to the same specifications. These forms are administered at different times so that test users have some flexibility in selecting a test date. Such flexibility, however, requires that reported scores on the alternate forms administered at different dates be comparable and should not offer (dis)advantages to examinees depending on which date they took the test.

Alternate forms of a test are typically constructed to be as similar as possible in content and statistical characteristics including difficulty levels. However, no matter how carefully forms are constructed, there will be some differences in difficulty between alternate forms. As a result of equating, scores on alternate forms of the same test have the same meaning and can be used interchangeably. This entry begins with the basic concept of equating; provides an overview of data collection designs, equating methods, and smoothing techniques; and concludes with some cautionary remarks on the accuracy of equating.

Basic Concept of Equating

The goal of equating is to find an equating relationship that transforms the scores on Form X (i.e., a new form) to the scale of Form Y (i.e., an old or base form). It is assumed here that some process has been used to establish a raw-to-scale score transformation for the base form Y.

Typically, the first step in equating is to determine a transformation function based on raw scores (e.g., number-correct scores) for both forms. Then, the equated raw scores on Form X are converted to scale scores on Form Y. These scale scores are the scores reported to examinees. When this process is performed successfully, the reported scale scores have the same meaning regardless of which form was administered. For example, a scale score of 500 on Form X indicates the same level of performance as a scale score of 500 on Form Y.

Equating Designs

For equating to be successful, differences attributable to forms must be separated from differences in the examinee groups taking the forms. Accomplishing this requires using a data collection design in which the sets of data for Form X and Form Y have some link between them—either common items or common (or similar) persons. The three most commonly used designs are the random groups design, the single group design, and the common-item nonequivalent groups (CINEG) design, sometimes called the nonequivalent anchor test design.

Random Groups Design

In the random groups design, the groups taking Form X and Form Y are randomly equivalent so that the differences between the scores on the two groups can be viewed as a direct indication of differences in difficulty between the two forms. A spiraling process is often employed to randomly assign forms. For example, Form X and Form Y can be distributed to examinees in an alternating order (e.g., Form X to the first examinee, Form Y to the second examinee, and Form X to the third examinee).

Single Group Design

In the single group design, the same examinees are administered both Form X and Form Y so that the differences in the scores on the two forms are attributable solely to form differences. Because examinees take two forms in one administration, the order of form administration can influence examinee performance. To reduce such unintended order effects, the order of administration of the forms is often counterbalanced. For example, half of the examinees are administered Form X first and Form Y second, whereas the other half are administered Form Y first and Form X second.

CINEG Design

In the CINEG design, two forms are administered to different groups of examinees that likely differ in proficiency level. Therefore, the difference in group performance on the two forms reflects both differences in the proficiency levels of the groups as well as differences in difficulty between the forms. In order to separate the group differences from the form differences, this design uses a set of common items that appear on both Form X and Form Y. The scores on the common-item set are used to estimate differences in proficiency levels between the groups. This estimate, along with certain statistical assumptions, makes it possible to establish an equating relationship between the full-length forms. It is particularly important that the set of common items be representative of the full-length forms with respect to both content and statistical specifications. A set of common items is called either internal or external depending on its contribution to the total score on the test.

Equating Methods forthe Random Groups Design

The equating methods discussed here are based on the use of observed scores (as opposed to true scores) and can be applied to both the random groups and single group designs. An observed score equating method produces, after equating, score distributions on Form X and Form Y that have the same first few moments. Some equating methods that involve true scores are considered in later sections.

Mean Equating

The simplest method is mean equating, which is based on a strong assumption that Form X and Form Y differ in difficulty by a constant amount along the score scale. In mean equating, the scores on Form X are adjusted so that the mean of the transformed scores on Form X are the same as the mean of the Form

Y scores. The transformation function for mean equating is given by $M_y(x)=x-\mu_X+\mu_Y$, where μ indicates the mean and $M_Y(x)$ refers to the mean equating function that converts score x on Form X to the scale of Form Y.

Linear Equating

In linear equating, the differences in difficulty between Form X and Form Y are allowed to vary along the score scale. The goal is to find an equating function for Form X scores that has the same mean and standard deviation as for the Form Y scores. The linear equating function is given by $L_Y(x)=(\sigma_Y/\sigma_X)(x-\mu_X)+\mu_Y$, where σ is the standard deviation and $L_Y(x)$ is score x on Form X converted to the Form Y scale using linear equating.

Equipercentile Equating

Equipercentile equating is a nonlinear method in which the scores on Form X are adjusted so that the cumulative distributions of the scores on Form X and Form Y are as similar as possible. In other words, in equipercentile equating, the transformed Form X scores should have the same percentile ranks as the Form Y scores. Note that for continuous variables, a percentile rank can be found directly based on cumulative distribution functions. However, because observed scores are typically discrete, a continuization process must be used. For example, the percentile rank for an integer score x traditionally is defined as the percentile rank at the midpoint of an interval $[x - 0.5, x + 0.5]$, assuming that examinees obtaining the score x are uniformly distributed over the interval. Another continuization method often considered is Gaussian kernel smoothing.

Equating Methods for the CINEG Design

The CINEG design involves two populations, for which one is administered Form X and the other is administered Form Y. Because an equating function is typically defined for a single population, some equating methods for the CINEG design use the concept of a synthetic population. To construct the synthetic population, the new and old populations (denoted by N and O), are weighted by w_N and w_O, respectively, such that $w_N + w_O = 1$ and $w_N, w_O \geq 0$. The choice of the weights is arbitrary, but its impact on equating results tends to be minor.

Linear Methods

Two linear methods that use synthetic populations are the Tucker and Levine observed score methods. For both methods, the linear equating function is the same as the one defined for the random groups design, which is given by $L_{Y:S}(x)=(\sigma_{Y:S}/\sigma_{X:S})(x-\mu_{X:S})+\mu_{Y:S}$, except that all terms are now defined for the synthetic population, S. Each term in this linear function can be expressed as a weighted sum of the statistic for the two populations, in which some statistics (e.g., the mean and variance of the scores on Form X in the old population) cannot be obtained directly from the data and thus must be estimated. Doing so requires rather strong statistical assumptions.

The Tucker and Levine observed score methods make different assumptions based on observable quantities to estimate the four unobservable quantities. For example, the Tucker method assumes that the regression of total scores X on common-item scores V is the same linear function for the new and old populations and that the conditional variance of X given V is the same for both populations. Similar statements hold for Y given V. By contrast, the Levine observed score method makes assumptions that pertain to true scores, and the principles of classical test theory are used to relate the true scores to observed scores.

A few methods exist that do not depend on a synthetic population. The Levine true score method is one such method. The Levine true score method makes the same assumptions as the Levine observed score method. The major difference is that the Levine observed score method relates observed scores on X and Y, whereas the Levine true score method relates true scores on the two forms. In practice, observed scores are used in place of the true scores in the Levine true score equating function, although there is no theoretical justification for doing so.

Another method that does not involve a synthetic population is chained linear equating. Chained linear equating is conducted through a chain of linear equating as follows: (a) Find a linear equating function, $L_{V:N}(x)$, that converts scores on X to the scale of scores on V using the Form X data; (b) find a linear equating function, $L_{Y:O}(v)$, that converts scores on V to the scale of scores on Y using the Form Y data; and (c) obtain an equating relationship to transform scores on X to the scale of Form Y by using $L_Y(x)=L_{Y:0}[L_{V:N}(x)]$.

Equipercentile Methods

Three nonlinear equating methods are presented here: frequency estimation, modified frequency estimation, and chained equipercentile. For frequency estimation equipercentile equating, the frequency distributions of *X* and *Y* are expressed for a synthetic population as $f_s(x)=w_N f_N+w_0 f_0(x)$ and $g_s(y)=w_N g_N(y)+w_0 g_0(y)$, where f and *g* refer to the frequency distributions for Form X and Form Y, respectively. From the data collected under the nonequivalent groups design, $f_N(x)$ and $g_O(y)$ can be computed directly; however, $f_O(x)$ and $g_N(y)$ cannot. These two unobservable frequency distributions are estimated under the assumption that, for Form X, the conditional distribution of *X* given *V* is the same in the new and old populations. The same assumption is made for Form Y. With these assumptions, the cumulative distributions and the percentile rank functions for *X* and *Y* are derived for the synthetic population. Equipercentile equating is then conducted. The modified frequency estimation method alters one of the assumptions in the original frequency estimation method; the net effect is that the modified frequency estimation method tends to reduce equating bias.

Chained equipercentile equating is the equipercentile analogue of chained linear equating. Chained equipercentile equating is conducted as follows: (a) find an equipercentile equating function from *X* to *V*, $P_{V:N}(x)$ using the Form X data; (b) find an equipercentile function from *V* to *Y*, $P_{Y:O}(v)$ using the Form Y data; and (c) put the Form X scores on the scale of Form Y by using $P_Y(x)=P_{Y:0}[P_{V:N}(x)]$.

Smoothing

When samples of examinees are used in place of populations to estimate equipercentile equating relationships, random sampling error always exists. To reduce random error, two types of smoothing are often considered: presmoothing and postsmoothing. In presmoothing, score distributions are smoothed prior to equating, whereas in postsmoothing, equipercentile equating relationships are smoothed directly after equating. Because smoothness is viewed as a characteristic of score distributions and equipercentile relationships in the population, smoothing can be considered when irregularity appears in score distributions or equipercentile relationships, with the expectation that smoothing improves equating accuracy. Smoothing is performed to reduce random sampling error; however, smoothing introduces bias to some degree. Thus, the primary goal of presmoothing is to decrease overall equating error by reducing a substantial amount of random error while allowing for as little bias as possible.

When a presmoothing method is applied to an observed frequency distribution, it is important to ensure that the fitted smoothed distribution does not depart too much from the observed distribution; otherwise, considerable bias might be introduced. To prevent the smoothed distribution from being too disparate from the observed distribution, presmoothing methods are often desired so that the first few moments of the smoothed distribution are the same as those of the observed score distribution—this is often called the moment preservation property. One of the most commonly used presmoothing methods is based on polynomial log-linear models, in which polynomial functions are fit to the log of the sample density. The log-linear presmoothing method requires the user to select a value for the model parameter, which determines the number of moments preserved.

Another family of presmoothing methods is generally referred to as strong true score models. A strong true score model assumes a distribution of true scores. For example, the four-parameter β binomial model assumes a four-parameter β distribution of true scores and a binomial distribution for conditional observed scores. Fitting a strong true model to data results in a smoothed observed score distribution, for which the first four moments will agree with those of the observed score distribution.

The equipercentile equating relationship produced based on presmoothed score distributions will almost always be smooth. However, an equipercentile equating relationship based on unsmoothed score distributions will likely demonstrate a jagged pattern, which indicates sampling error. Fitting a curve to the jagged equipercentile equating relationship is called a postsmoothing method. The most frequently used postsmoothing method is based on cubic spline interpolation. The cubic spline postsmoothing method defines a different cubic function between contiguous integer scores. There is a parameter in the model that controls the degree of smoothing, which is set by the user. The values for the parameter are typically between 0 and 1.

Item Response Theory (IRT) Methods

IRT refers to a collection of mathematical models that relate examinee latent abilities to the probability of scores on the responses (e.g., right or wrong for a multiple-choice item) to items in a test. Various IRT models that differ in their functional form have been used in many testing applications, including equating. Two general approaches to IRT equating are true score equating and observed score equating. The discussion of IRT equating in this entry focuses primarily on unidimensional IRT models, although some aspects of the methodologies could be extended to multidimensional IRT models.

Prior to conducting equating, ability and item parameter estimates for Form X and Form Y must be placed on the same IRT ability scale. In particular, when the two forms were administered to groups that differ in their ability levels (i.e., CINEG design), separate calibration of the two data sets will result in parameter estimates on the two forms that are not on the same scale. In such a case, a scale linking process is needed using a method such as the Stocking-Lord and Haebara methods to place the parameter estimates for the two forms on a common scale. Multigroup concurrent calibration and fixed parameter calibration are two other scale linking methods frequently used in practice.

True Score Equating

In IRT true score equating, a pair of true scores on Form X and Form Y is considered to be equivalent when they are associated with the same ability level θ. Test characteristic curves are used to determine an equating relationship between true scores on the two forms. True score equating begins with finding a θ value corresponding to each raw score on Form X. Then, a true score on Form Y associated with the θ value is determined. These steps are repeated for all raw-score points in Form X. IRT true score equating involves true scores, which cannot be known in practice, and the resultant true score equating relationship is applied to examinees' observed scores without theoretical justification.

Observed Score Equating

IRT-observed score equating involves estimating observed score distributions for both forms using estimated item and ability parameters. The estimated observed score distributions can be viewed as the expected score distributions when the model fits the data perfectly. Traditional equipercentile equating is then applied to the estimated observed score distributions to find an equating relationship. For the CINEG design, the ability distributions for the two groups placed on the same scale are used to produce a synthetic population for which the equating relationship is determined.

Accuracy of Equating

Equating is conducted for the purpose of improving the accuracy of test scores that are often used to make important decisions about test takers. Thus, it is important to make every effort to minimize equating error so that the equated scores are accurate and stable. It is particularly important that equating be applied to alternate forms of a test that are built to be as similar as possible in terms of content and statistical specifications; if not, the scores on the alternate forms cannot be used interchangeably no matter how well the scores are adjusted. Some practical considerations for controlling equating error include rigorous test form development, use of a large sample size, smoothing techniques, adequate selection and proper implementation of an equating design, and appropriate choice of an equating method.

Won-Chan Lee and Hyung Jin Kim

See also Classical Test Theory; Item Response Theory; Percentile Rank; Scales; Score Linking; Score Reporting; Vertical Scaling

Further Readings

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Dorans, N. J., Pommerich, M., & Holland, P. (Eds.). (2007). Linking and aligning scores and scales. New York, NY: Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187–220). Westport, CT: American Council on Education and Praeger.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking:

Methods and practices (3rd ed.). New York, NY: Springer.

Cynthia S. Darling-Fisher Cynthia S. Darling-Fisher Darling-Fisher, Cynthia S.

Erikson’s Stages of Psychosocial Development Erikson’s stages of psychosocial development

598

601

Erikson’s Stages of Psychosocial Development

Erik Erikson’s theory of the stages of psychosocial development, first presented in the 1950s and refined over the course of his life, is fundamental to the understanding of an individual’s personality development over the course of the life span. Erikson proposed that personality development was a process that evolved through the interaction between biological, psychological, social/cultural, and historical factors. Erikson described eight psychosocial “crises” (or conflicts) that individuals face over the life span as they interact with their environment. Erikson proposed that each crisis must be resolved before individuals are prepared to move to the next stage and that unresolved conflicts at one stage influence development at later stages. According to Erikson, the sequence of the stages represents successive development of the component parts of the psychosocial personality and is invariant across cultures. However, the ways in which different cultural groups meet the stage conflicts may vary. Erikson’s theory has provided the foundation for studies by researchers from multiple fields interested in the study of personality development, the interaction between environment and personal development, and how individuals adapt to or cope with a variety of life issues over the course of the life span. This entry describes each of the eight stages of Erikson’s theory of psychosocial development and then examines the relevance of Erikson’s theory to research.

Basic Principles: The EightStages of Psychosocial Development

Each stage is characterized by a central conflict arising from the interaction between personality, developmental, and social processes. These place specific demands on the individual that are necessary for growth and positive ego formation. Successful resolution leads to development of a particular strength and virtue for that stage, which in turn influences later attributes. Resolution of the stage determines the relative prominence of either positive or negative attributes. If the balance is toward the positive, it will help the individual meet later crises and provide a better opportunity for unimpaired psychosocial development. On the other hand, some expression of the negative component is to be expected and even necessary for healthy development. Erikson also proposed that the individual's ability to resolve each of the developmental conflicts was related to interaction with significant individuals at the different stages of development. He termed these relationships "the radius of salient significant relations" for each stage.

Dealing with each conflict at a particular stage of development provides the basis for progress to the next stage. As a person faces each challenge, the person assumes both increased vulnerability and increased potential, and a new strength emerges that contributes to further development. Erikson stated that all components of the personality are present in some form even before their emergence as a "crisis," and they remain systematically related to all the other components. Therefore, optimal development depends on the proper resolution of conflicts in the appropriate sequence and integration of newly added identity elements with those already in existence.

The stages and resulting strengths and virtues as described in Erikson's writings are presented in the following sections. Stages 1–4 address infancy and childhood, Stage 5 focuses on adolescence, and Stages 6–8 address adulthood.

1. Basic Trust Versus Mistrust(Strength: Drive; Virtue: Hope)

Development of a trusting relationship with a parent during infancy creates a sense of security and belonging. Lack of trust or mistrust can lead to insecurity, withdrawal, and lifelong relationship difficulties. A positive resolution provides the basis for a hopeful outlook on life.

2. Autonomy Versus Shame/Doubt (Strength: Self-control; Virtue: Will-power)

Success provides the child with self-assurance, pride, and self-control. Tantrums and displays of stubbornness are part of the struggle toward autonomy. Lack of environmental support for autonomy can lead to shame and doubt. Positive resolution leads to self-assurance and self-reliance.

3. Initiative Versus Guilt(Strength: Direction; Virtue: Purpose)

Initiative builds upon autonomy and ingenuity, giving the child confidence to complete new tasks, and is the beginning of the development of morality, insight, responsibility, and the ability to form new relationships. Failure to achieve this broadening of social life can negatively affect initiative and lead to feelings of guilt, anxiety, and hopelessness. Positive resolution leads to the ability to initiate ideas and set goals.

4. Industry Versus Inferiority(Strength: Method; Virtue: Confidence)

The school-age child learns mastery of major social skills (such as relating to peers), intellectual skills (such as reading, writing, and arithmetic), and completing tasks. Discouragement and inadequacy can lead to feelings of inferiority, reversion back to previous developmental stages, and school failure. Positive resolution leads to the ability to make and produce things and the confidence to seek and respond to challenges.

5. Identity Versus Role Confusion/Identity Diffusion (Strength: Devotion; Virtue: Fidelity)

Adolescence involves the transformation from childhood to adulthood. The adolescent is concerned with the views of others. Social relationships become important support structures while the adolescents also learn to find their place within them. The adolescent must apply the cognitive, decision making, and coping skills learned thus far to different everyday situations and determine the social values necessary to succeed in the adult world. Successful resolution leads to a strong and stable sense of self and of potential for achievement in life, often associated with choosing a career path. Failure can lead to recklessness, delinquency, withdrawal, and an unclear sense of identity.

6. Intimacy Versus Isolation(Strength: Affiliation; Virtue: Love)

As young adults become secure in their self-identity, they become ready to share themselves with others through tangible, committed relationships (e.g., successful marriage and enduring friendships). Through these relationships, ethical commitments often resulting in sacrifice and negotiation are developed. Successful resolution leads to a capacity to bond with and commit to others and the ability to give and receive love. Avoidance of interpersonal relationships due to a lack of intimacy leads to isolation and loneliness. Failure to achieve such important relationships hinders personality development and may prevent success in the subsequent developmental stage of generativity.

7. Generativity Versus Stagnation/Self-Absorption (Strength: Production; Virtue: Care)

The main conflict of this stage is to establish and guide future generations through acts of care for others. Child-rearing and the establishment of a family, as well as productive and creative personal achievements (such as teaching and mentoring) for those without children, are prototypical of this stage. A sense of stagnation or self-absorption results when generativity cannot occur, and there is regression back to a form of pseudo-intimacy or a withdrawal from life. This stage can last over 30 years and is described by Erikson as probably the most fulfilling stage and a time of renewal and continued growth.

8. Integrity Versus Despair(Strength: Renunciation; Virtue: Wisdom)

The conflict of the last stage of development is to come to terms with the life one has lived and the person one has become. Integrity is characterized by life review during which the individual integrates the positive and negative elements of the life lived, often acknowledging and resolving regrets. Wisdom is the prevailing strength of this stage. A difficulty with resolution of this stage is characterized by an unwillingness to accept the life lived, a fear of death, and wishing for more time, hopelessness, pessimism, and despair.

wishing for more time, hopelessness, pessimism, and despair.

Relevance of Erikson’s Stages to Research

Application of Erikson’s theory can help researchers in the design of research, choice of variables to study, and constructs to measure. Erikson’s depiction of the differing conflicts addressed in each stage, the significant relationships that provide support for resolution of that stage, and anticipated attributes (strengths/virtues) arising from successful resolution highlights key variables to consider. These variables may have a direct effect on the outcomes or may act as mediators. Awareness of these factors may help determine the measures selected for the study. Alternatively, measurement of key constructs of Erikson’s theory of personality development (e.g., resolution of specific stages such as identity formation in adolescence) is also a focus for research.

Erikson’s psychosocial stages have been used in a variety of ways in research. Erikson’s theory has been used as a foundation for research by highlighting the issues to be studied (e.g., parenting), populations of interest (e.g., retirees, newly married), and specific mediating variables that might impact the research (e.g., individuals’ developmental issues and career development). The epigenetic nature of the theory also suggests that researchers should consider developmental stage issues when conducting research. While Erikson did not prescribe specific age ranges for each stage, he proposed an invariant sequence for psychosocial development. He also acknowledged that culture and context influenced how individuals met stage-related challenges. Consequently, when designing a study, attention to the psychosocial issues of concern to the study population may be as salient as demographic characteristics such as age. For example, with research on freshmen community college students, awareness of developmental issues, in addition to age and gender, in this potentially heterogeneous population could have an important influence on the outcome. A 20-year-old “traditional” student may have differing issues from one who is a parent or is returning to school from military service. In a study examining the quality of adult relationships, a measure of developmental attributes such as trust and intimacy may be important in examining relationship problems. In research with school-age children, it could be important to consider development of self-control (an outcome of autonomy vs. shame) in a study of educational achievement or in development of classroom interventions. Adolescents’ sense of their identity and relationships with their peers and family may also be important to consider in research with teenagers.

Another approach is to study the conflicts of life span psychosocial development and the component relationships. Taking an Eriksonian perspective, the researcher could investigate Erikson's proposed conflicts for that stage, significant support networks important at that stage, and include personality strengths, along with other variables in the study. For example, a study of new parents might consider the individual parent's developmental status, spouse support, and satisfaction with parenthood, acknowledging that some of these factors could be mediating variables to control in the project.

A third approach addresses construction of instruments to measure Eriksonian constructs and further test Erikson's propositions related to the adult stages. Considerable research has examined components of identity and the achievement of identity status. Research has looked at generativity and its different expressions in women and men, and individuals with and those without children. Prior to his death, Erikson and his wife, Joan, started work on a ninth stage that considered specific developmental issues experienced in extreme later life (i.e., late 80s, 90s, and beyond) including continued life satisfaction. These concepts need further expansion. Measures have also been developed to assess overall stage resolution (e.g., the balance between positive and negative attributes or strengths developed such as hope, purpose, fidelity, and wisdom). Research in this area is warranted as well.

The breadth of Erikson's theory of psychosocial development across the life span is a strength for research but also creates challenges. Critics of Erikson's theory have described it as too linear and gender biased, focusing more on males than females. Some also say it does not give enough recognition to cultural differences and that the adult stages are too broad. However, research has supported Erikson's major premises, in particular, that successful resolution of earlier stages is a foundation for later developmental achievement (however, the sequencing may vary based on gender and culture).

In summary, Erikson's stages of psychosocial development continue to be relevant to today's researchers. Erikson's theory has been used in research on a wide variety of topics including career development, educational attainment, mentoring, retraining programs for the unemployed, coping with illness, and response to normal life transitions in adulthood like marriage, birth, divorce, and aging. Researchers from a wide range of disciplines (e.g., psychology, education, sociology, social work, nursing, medicine, kinesiology, religion, administration) continue to find Erikson's theory useful for examining the issues individuals face at each developmental stage, their implications for how individuals develop face at each developmental stage, their implications for how individuals develop and adapt successfully to life’s changes, and how individuals can live productively in their society across the life span. Taking into account psychosocial developmental issues, along with other variables, will assist the researcher in identifying variables that may have either direct or mediating effects on study outcomes and thus promote more successful research.

Cynthia S. Darling-Fisher

See also Adolescence; Childhood; Developmental Evaluation; Personality Assessment

Further Readings

Erikson, E. (1959). Identity and the life cycle: Selected papers (Psychological Issues, Monograph 1). New York, NY: International Universities Press.

Erikson, E. H. (1963). Childhood and society (2nd ed.). New York, NY: W.W. Norton.

Erikson, E. H. (1982). The life cycle completed: A review. New York, NY: W.W. Norton.

Erikson, E. H. (1998). The life cycle completed. Extended version with new chapters on the ninth stage by Joan M. Erikson. New York, NY: W.W. Norton.

Erikson, E. H., Erikson, J. M., & Kivnick, H. Q. (1986). Vital involvement in old age. New York, NY: W.W. Norton.

Error

See Conditional Standard Error of Measurement; Parameter Mean Squared Error; Parameter Random Error; Standard Error of Measurement; Type I Error; Type II Error; Type III Error

Joseph A. Rios Joseph A. Rios Rios, Joseph A.

Ting Wang Ting Wang Wang, Ting

Essay Items

Essay items

601

605

Essay Items

An essay item requires students to produce a written expression in answer to a question or in response to a prompt. Such an item requires students to (a) recall factual, conceptual, or procedural knowledge; (b) mentally organize this knowledge; and (c) interpret the knowledge and construct it into a logical, integrated response in clear and appropriate language. There are several rules for developing essay items: (a) restrict their use to assess high-level learning outcomes such as creating or evaluating, (b) construct them to measure the skills necessary to achieve the learning outcomes, (c) clearly phrase a question, and (d) indicate response page and time limits if possible.

Essay items differ from selected-response (multiple choice or true/false) items in three ways: (1) More complex learning outcomes, such as analysis, synthesis, and evaluation, can be assessed; (b) students can pick the information that they would like to include and decide how to organize the information; and (c) students are required to provide an answer without having seen it presented, which greatly reduces the possibility of guessing. Essay items in general provide an efficient measure (i.e., they are often easier and less time consuming to construct than selected-response items) of higher order cognitive skills; however, when compared to selected-response items, they require greater resources for scoring, and the scoring itself is more subjective in nature. The remainder of this entry describes procedures for scoring essay items, challenges in gathering validity evidence for scores obtained from essay items, and the advantages and disadvantages of essay items.
>

Procedures for Scoring

This section provides a brief description, as well as the advantages and disadvantages, of the two approaches used in essay scoring: human scoring and automated scoring.

Human Scoring

Human raters are used to evaluate an essay's quality by assigning a score associated with response characteristics that are outlined in a scoring rubric. A scoring rubric provides predefined descriptive scoring schemes that are developed by substantive experts to guide the analysis of students' written responses. The assumption of employing predefined scoring schemes is that the evaluation of written responses becomes less subjective and provides greater consistency in ratings.

In general, there are two types of rubrics that are commonly used in human scoring: holistic and analytic. Holistic rubrics provide the descriptions of abilities, skills, and proficiencies that examinees are expected to demonstrate at a particular score level. Analytic rubrics are more specific and break down the characteristics of each score into several components, allowing raters to itemize and define the strengths and weaknesses of the responses. Scoring with holistic rubrics generally takes less time than scoring with analytic rubrics, but analytic rubrics provide more detailed individual-level scoring criteria and feedback. Although analytic rubrics tend to provide more fine-grained feedback to students, one concern is that the scoring dimensions tend to be highly correlated with one another. For examples of holistic and analytic rubrics, readers can refer to "Designing Scoring Rubrics for Your Classroom" by Craig A. Mertler and *Educational Assessment of Students* by Anthony J. Nitko.

In high-stakes testing programs, implementation of human scoring involves three major steps: training, calibration, and operational scoring. The first step requires raters to carefully read the rubric to gain familiarity and to examine sample essays that correspond to each score level. Each sample essay represents various aspects of the rubric for the assigned score, which assists raters with better familiarizing themselves with the basics of the scoring scheme. Then, raters are assigned a set of prescored responses, with the score withheld from the rater, to evaluate individually.

If raters cannot reach a specific agreement level between the preassigned score and the rater-assigned score, recalibration is required. However, raters who meet the criterion for calibration (i.e., the predefined agreement level) move on to operational scoring. During operational scoring, scoring leaders are assigned to supervise and ensure the quality of scoring by individual raters. Quality control can be implemented by having a certain percentage of essays double scored to ensure scoring consistency. If there are scoring discrepancies, the scoring leader can provide feedback to individual raters as to how scoring can be improved.

Employing human scoring provides a number of advantages. First, human raters can utilize cognitive judgment skills to decode contextualized responses, make connections to their prior knowledge, and, based on their understanding of the content, make a judgment concerning the quality of the response. Second, human raters have the ability to distinguish the nonnormative, incorrect ideas from normative, correct ones. Lastly, they can make judgments about examinees' higher level writing skills as well as the factual correctness of claims made in the written response. Although human raters provide a number of advantages, they can be difficult to recruit, they may require extensive training, and they must be closely monitored to maintain score quality. This process can be expensive and time consuming and will still result in less-than-perfect scoring quality.

Automated Scoring

Automated scoring can be defined as the use of computer technology to evaluate and score written prose. This is generally done by aggregating construct-relevant text features that can be extracted from written responses and combined into a mathematical model that produces a score. Implementation of automated scoring involves three major steps: model building, model evaluation, and scoring.

In building a model, it is important to acknowledge that the computer technology used for automated scoring can neither read nor understand the content of an essay as a human would. As an example, a human's score could be influenced by the interaction of internal (i.e., latent) variables inherent within written prose, such as diction, grammar, and fluency. As the computer technology cannot directly evaluate these internal variables, strong correlates or proxies must be identified that can be automatically extracted from text responses. These proxies are largely identified through cross-disciplinary research in natural language processing.

Once automated scoring developers have programmed procedures to automatically extract these proxies from text responses, they must decide on how each proxy should contribute to the machine-generated score. This is done by combining the proxies into statistical models that weight each proxy differently and then, comparing the agreement between the machine-generated scores from each model to human ratings from a training set of responses. The model that provides the highest agreement is then evaluated for agreement on an independent set of human ratings.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer proposed several criteria to evaluate an automated scoring model that are informed by the test stakes, purpose, and population for which the automated scoring model will be operationally deployed. In general, if these criteria are not met, the proxy weights must be altered. However, it should be considered that there are some items that cannot be adequately scored solely by computer, and consequently altering the proxy weights will not resolve this issue. If it has been decided that an item can be scored via computer and the criteria are met, generally, the model can be implemented for operational scoring. This process is repeated for every given item. For a review of different types of automated scoring approaches, the reader is referred to “Automated Essay Scoring: A Literature Review” by Ian A. Blood.

Advantages of using automated scoring include its efficiency, absolute consistency, and instant score feedback. In general, automated scoring can provide more objective scoring of written responses as the machine is not influenced by external factors (e.g., fatigue), nor is it susceptible to common human-rater errors and biases.

Researchers have identified the disadvantages of using automated scoring. Mo Zhang argues that automated scoring requires expensive system development, maintenance, and enhancement. Furthermore, Donald E. Powers, Jill C. Burstein, Martin Chodorow, Mary E. Fowles, and Karen Kukich have demonstrated that examinees can “game” the system by providing textual features important in the scoring model (e.g., response length, sentence complexity, key vocabulary) and receive high machine-generated scores regardless of the response content. As a result, when employing automated scoring in high-stakes assessment, testing programs (e.g., GRE) require that written responses are scored by both automated and human procedures written responses are scored by both automated and human procedures.

Challenges in GatheringValidity and Reliability Evidence

Professional standards in educational and psychological testing (e.g., the *Standards for Educational and Psychological Testing* from the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) require evidence of inferences made from test scores derived from essay items for both human and automated scoring. The most frequently used standard for evaluating the validity of automated scoring is assessing its comparability with human ratings for the total sample and subgroups. In general, previous research has demonstrated that automated scores strongly correlate with human scores and, in some cases, correlate more highly than the correlation between human raters. However, Randy E. Bennett argues that agreement with human ratings should not be seen as validation of automatic scoring.

Previous research has shown that human ratings can be biased based on (a) examinee handwriting quality, (b) occasion (i.e., raters may differentially score the same essay at different time points), (c) examinee ethnicity, (d) examinee sex, (e) rater experience, (f) severity drift (i.e., raters consistently shifting their scores from one end of the scale to the other), and (g) central tendency (i.e., the tendency to award scores around the mean) to name a few. Consequently, if these rater biases go unchecked, errors will be propagated to the automated scoring system and could lead to differential human ratings. Therefore, documenting that raters are conducting scoring consistent with the construct and measurement goals is an important aspect of gathering validity evidence regardless of whether human ratings will be used alone, as the target for developing automated scoring algorithms, or in conjunction with machine scores.

There are several methods for evaluating rater effects, such as generalizability theory and latent trait models. However, even if there is sufficient support for the validity of human ratings, the agreement between human and automated scores is not perfect. As a result, there is a need to collect additional validity evidence that should be consistent across all subpopulations of examinees, such as construct relevance and coverage, high correlations with external measures of the construct, and discriminant correlations with measures that differ from the construct.

Conclusion

Essay items provide three major advantages over selected-response items as they (a) reduce measurement error due to random guessing, (b) eliminate unintended corrective feedback (e.g., if a wrong answer computed by students themselves is not listed in the options of a selected-response item, students can be unintentionally reminded of their mistakes), and (c) improve the construct validity of the test, as essay items require students to write down how they apply academic principles to solve the problem and communicate the findings. On the other hand, grading essay items can be quite expensive, time consuming, less objective, and accurate than selected-response items. Essay items also limit the ability to ask a large number of questions on a wide range of subject materials in a given time period.

Despite these disadvantages, the use of essay items does not seem to be disappearing any time soon. Many examinees perceive essay items to be a more fair assessment of ability than selected-response items, and many teachers prefer essay items as they are perceived to more authentically measure higher order cognitive ability. In addition, research has indicated that essay items are as reliable and effective in predicting academic success as selected-response items. For essay items to be as effective as possible, however, more research is needed to better understand: (a) rater cognitions to improve human scoring agreement and the standard to which automated scoring quality is judged and (b) writing cognition to create an improved basis for deriving dimensions and features for scoring.

Joseph A. Rios and Ting Wang

See also Admissions Tests; Generalizability Theory; Performance-Based Assessment; Rubrics

Further Readings

Bennett, R. E., & Ben-Simon, A. (2005). Toward theoretically meaningful automated essay scoring. Journal of Technology, Learning, and Assessment, 6, 1–47.

Blood, I. A. (2011). Automated essay scoring: A literature review. Working Papers in TESOL … Applied Linguistics, 11, 40–64.

Bridgeman, B. (in press). Can a two-question test be reliable and valid? Educational Measurement: Issues and Practice.

Criswell, J. R., & Criswell, S. J. (2004). Asking essay questions: Answering contemporary needs. Education, 124, 510–516.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them (ETS R … D Connections, No. 11). Princeton, NJ: Educational Testing Service.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research … Evaluation, 7(25). Retrieved July 2016, from http://pareonline.net/getvn.asp?v=7…n=25

Nitko, A. J. (2001). Educational assessment of students (3rd ed.). Upper Saddle River, NJ: Merrill.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated essay scoring (GRE Board Professional Rep. No. 98–08bP, ETS Research Rep. No. 01–03). Princeton, NJ: Educational Testing Service.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.

Zhang, M. (2013). Contrasting automated and human scoring of essays (Report No. R … D Connections, 21). Princeton, NJ: Educational Testing Service.

Kentaro Kato Kentaro Kato Kato, Kentaro

Estimation Bias

Estimation bias

605

606

Estimation Bias

Estimation bias, or simply bias, is a concept in statistical inference that relates to the accuracy of parameter estimation. The term *bias* was first introduced in the statistical context by English statistician Sir Arthur L. Bowley in 1897. This entry provides the formal definition of estimation bias along with the concept of error, its implications and uses in statistical inference, and relevance to other types of bias that may arise in the data collection process.

The Concept ofError in Statistical Inference

Suppose that we would like to estimate a population parameter θ (e.g., population mean). An estimator is any sample statistic (e.g., sample mean) that is used to estimate θ. Because is sample based, it does not perfectly agree with the true value of θ. The difference between the values of the estimator and the parameter, , is called the error of estimation.

In the sampling theory, statistical performance of an estimator is evaluated by the smallness of error in the long run, in which one assumes that an infinite number of random sampling from the population is possible and considers the average size of error in the repeated sampling. For any estimator , the following equation generally holds:

$$E(\hat{\theta}-\theta)^2 = E(\hat{\theta}-E(\hat{\theta}))^2 + (E(\hat{\theta})-\theta)^2,$$

where the operator *E* refers to taking the expectation of the subsequent variable with respect to the sampling distribution of . The left-hand side is called the mean squared error (MSE), which is the average squared error of estimation. In general, estimators with smaller MSE are preferred to others when there are competing estimators of the same population parameter. The first term on the right-hand side is the variance of the estimator, indicating the average squared deviation of the estimator from its expected value (i.e., the mean of in the long run). The second term on the right-hand side is the squared difference between the expected value of and the true parameter value. The difference in the parentheses is called the bias of estimator for parameter θ, that is:

$$\mathrm{Bias}\,(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta.$$

The equations imply that the total error of an estimator can be decomposed into the variance and bias components. The variance component represents the amount of random error, or precision, which is the unpredictable fluctuation due to sampling and cancels out in the long run. In contrast, the bias component represents the amount of systematic error, or accuracy, which remains constant throughout sampling occasions. In order for the MSE to be small, both variance and bias must also be small.

Implications and Uses in Statistics

The MSE equals the variance of estimator if and only if the bias is zero (i.e.,). Such an estimator is said to be unbiased (it is biased, otherwise). Unbiasedness is considered as one of the desirable properties for estimators. The sample mean, for example, is an unbiased estimator of the population mean if it is calculated from a random sample. Also, maximum likelihood estimators are asymptotically unbiased.

Suppose a random sample of size n (X_1, X_2, … , X_n) from the normal distribution with mean μ and variance σ^2. On the one hand, the sample mean is an unbiased estimator of the population mean μ because its expected value is . On the other hand, the sample variance is biased with respect to the population variance σ^2 because its expected value is $E(S^2) = (n - 1)\sigma^2/n$. In this case, the amount of bias is Bias $(S^2,\sigma^2)=E(S^2)-\sigma^2=-\sigma^2/n$. Thus, the sample variance tends to underestimate the population variance, although the bias diminishes as the sample size n becomes large. In order to obtain an unbiased estimator of the population variance, one can use the unbiased variance . Yet, it can be shown that the MSE of S^2 is smaller than that of ; S^2 tends to produce estimates closer to the true population variance than in the long run, even though those estimates tend to be smaller than the true variance.

There are many other examples of estimators that are biased but have smaller MSE than their unbiased counterparts by achieving small variance that more than makes up the presence of bias. This is called the variance-bias trade off and is often a topic in Bayesian inference, which in general results in biased estimators. Other properties being equal, unbiasedness can be useful in restricting the class of “good” estimators for a particular estimation purpose. However, it is not the sole criterion for choosing a best estimator.

Still, efforts are made to obtain better estimators by taking into account the unbiasedness. For example, the η^2 is a measure of effect size that represents the proportion of variance explained by the factor of interest to the total variation. The ω^2 is its bias-corrected version, although it is not completely unbiased. Use of an unbiased or less-biased estimator could matter when realized values of the estimator are averaged over multiple studies to obtain a single estimate of a population parameter as in meta-analysis.

Relevance to Other Types of Bias

Estimation bias is a purely statistical concept; it is theoretically derived from the statistical assumptions (i.e., the population model and the sampling scheme) and the choice of estimator. However, there are other sources that could cause systematic errors in real data. For example, selection bias occurs if the sample is not representative of the target population. Measurement bias takes place if one uses an ill-calibrated measurement instrument or scheme. Estimators that are supposedly unbiased could be invalid in these circumstances because the statistical assumptions from which the unbiasedness of those estimators is derived are likely violated.

It is crucial to use an appropriate data collection design and statistical modeling so that one can reduce or separate possible bias that arises in the data collection process. In the scoring of an essay task, for example, a rater may produce consistently higher scores than the true scores (i.e., measurement bias). Such bias cannot be identified if only scores from that single rater are analyzed. In this case, scores from multiple raters would be necessary to reveal systematic differences among raters.

Kentaro Kato

See also Bayesian Statistics; Distributions; Eta Squared; Generalizability Theory; Maximum Likelihood Estimation; Meta-Analysis; Normal Distribution; Variance

Further Readings

Bowley, A. L. (1897). Relations between the accuracy of an average and that of its constituent parts. Journal of the Royal Statistical Society, 60, 855–866.

Carlin, B. P., & Louis, T. A. (2009). Bayesian methods for data analysis (3rd ed.). Boca Raton, FL: Chapman … Hall.

Lindgren, B. W. (1993). Statistical theory (4th ed.). Boca Raton, FL: Chapman … Hall.

Catherine O. Fritz Catherine O. Fritz Fritz, Catherine O.

Peter E. Morris Peter E. Morris Morris, Peter E.

Eta Squared

Eta squared

606

607

Eta Squared

η^2 is a commonly used effect size estimate. It describes the proportion of the total variability in a data set that is associated with an effect. Its value is zero when there is no effect and 1.0 when the effect accounts for 100% of the total variability. η^2 is most often used in association with analysis of variance and can be calculated from the analysis of variance summary table.

η^2 = Sum of squares effect Sum of squares total.

η^2 can also be calculated from published *F* ratios, as long as all *F* ratios in the design are reported. Jacob Cohen provided general guidelines for what constitutes small ($\eta^2 = .01$), medium ($\eta^2 = .06$), and large ($\eta^2 = .14$) effect sizes in many areas of psychological research.

The symbol R^2 is sometimes used rather than η^2, to conform to the modern convention that Greek letters are reserved for population parameters, but its use can lead to confusion. R^2 is more commonly used with multiple regression. Like r^2 (or R^2), η^2 describes the proportion of variability in one variable (the dependent variable) that is related to another variable (the independent variable). Unlike r^2, η^2 accounts for both nonlinear and linear relationships. η^2 is obtainable from many statistical software packages.

With more than one independent variable (factor), partial is often reported rather than η^2. does not address the total variability in the data set; it excludes variability associated with factors and interactions other than the one under consideration. It describes the proportion of variability associated with an effect when variability associated with all other effects is excluded from consideration. It can be calculated from the analysis of variance summary table or from the published *F* ratio:

$$\eta_p^2 = \text{Sum of squares effect sum of squares effect} + \text{Sum of squares error.}$$

$$\eta_p^2 = df \text{ effect} \times F \text{ effect } df \text{ effect} \times F \text{ effect} + df \text{ error.}$$

The value of any effect size statistic is influenced by the design of the research, which can increase or decrease error variability. With complex factorial designs, η^2 and must be interpreted with care because each factor and interaction account for some of the variability present, increasing the value of and decreasing the value of η^2. Thus, η^2 and may not be comparable across studies. Generalized is a similar statistic, designed to facilitate comparisons across designs.

Based on the sample, η^2 and provide a point estimate of the population parameter. To identify likely values for the population effect size, it is essential to know the confidence interval. Unfortunately, these confidence intervals are not centered on the statistic, making calculation difficult. There are, however, downloadable utilities for determining these confidence intervals. Daniel Lakens’s blog, The 20% Statistician, is a good resource.

The η^2 family of effect size statistics are optimistically biased; that is, they overestimate the population effect size. This overestimate is arguably no more than that found with conventional correlations, but more realistic estimates can be obtained using the corresponding ω^2 statistics. Unfortunately, ω^2 cannot be accurately calculated for studies involving repeated measures.

Catherine O. Fritz and Peter E. Morris

See also Analysis of Variance; Effect Size; Interaction; Multiple Linear Regression

Further Readings

Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. Journal of Experimental Psychology: General, 141, 2–18. doi:10.1037/a0024338

Grissom, R. J., & Kim, J. J. (2012). Effect sizes for research: Univariate and multivariate applications (2nd ed.). New York, NY: Routledge.

Lakens, D. (2014, June 7). Calculating confidence intervals for Cohen’s *d* and eta-squared using SPSS, R, and Stata. Retrieved from http://daniellakens.blogspot.co.uk/2014/06/calculating-confidence-intervals-for.html

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

Ethical Issues in Educational Research Ethical issues in educational research

607

612

Ethical Issues in Educational Research

Ethics in educational research are multifaceted, contextual, emergent, and relational; considering ethics critically requires attention to the procedural and transactional as well as the relational and sociopolitical. Ethical and valid research necessitates that educational researchers understand and approach their roles with humility and carefully consider these issues.

This entry begins by discussing ethical issues that are embedded in research processes and relationships including researcher reflexivity, the expert–learner binary, the relational approach to research, research boundaries, and reciprocity. It then describes processes used so that research conforms to ethical standards, including institutional review boards (IRBs), ethics committees, and codes of ethics. The entry concludes by examining the concepts of informed consent and assent and looking at the differences between confidentiality and anonymity and the challenges to these forms of privacy in a digital world.

The Ethical Dimensionsof Researcher Reflexivity

Given the researcher’s central role in the development, implementation, and dissemination of educational research, it is an ethical imperative for researchers to consider the implications of their role throughout the research process. Critically engaging with and challenging interpretations, and the biases and societal or organizational structures that shape them, constitutes an especially important set of considerations in educational research. Addressing this ethical responsibility requires a reflexive approach to research that has at its core an equity stance.

When researchers acknowledge that they are socialized into specific ways of viewing the world, which include their ideologies, biases, belief systems, assumptions, and prejudices, they must not only acknowledge but must also actively monitor their subjectivities and how these influence and mediate their research. The following subsections provide specific areas of researcher reflexivity with particular ethical dimensions.

Pushing Against the “Expert–Learner Binary”

Sociopolitical contexts (macro and micro) influence research contexts and research dynamics and relationships. As a part of an ethical stance, educational researchers should consider how these larger forces manifest themselves in research goals, processes, and interactions and then specifically within and across data collection and analysis processes. Many contemporary researchers argue for an approach that situates research participants as “experts of their own experiences,” meaning that everyone involved brings wisdom and generates knowledge. While this may sound obvious, it is a departure from how research often happens.

This stance requires researchers to work to become sufficiently comfortable with the uncertainties that arise from engaging in processes that decenter codified knowledge (i.e., knowledge valued in the Western academy) and expertise by allowing time and space for mutual evaluations and conversations that strive for balance and equity in reconciling what are sometimes divergent expectations and interpretations (whether among different researchers or between researchers and research participants). Research is often built on an expert–learner binary—that is, an assumption that researchers are more knowledgeable and expert, and participants are viewed, and sometimes even view themselves, as passive recipients of this information and the research process overall rather than as knowledge generators in their own right. This, of course, is ethically problematic.

Relational Ethics: Taking a Relational Approach to Research

To push against forces that dehumanize, generalize, and “other” people, educational researchers can commit to an inquiry stance on research. This requires, among other things, taking a relational stance on research. This approach, which emerged from feminist research methodology, critically examines the relational dynamics between researchers and participants (and between participants) in relation to broader social forces. The relational aspects of the research process (and product in terms of how data are analyzed and participants are represented in written accounts), with its methodological attunement to and address of issues of power, identity, and the need to contextualize interaction and data, are at the heart of a relational approach to research.

This person-centered, societally contextualized approach to research places a primacy on the authenticity of the relationships between researchers and participants. It examines the roles, power structures, and language used to frame these relationships (such as *subjects*, *informants*, *interviewees*, and *participants*). Within this approach, the active, critical consideration of and thoughtful engagement in equitable relationships is at the heart of the research process and is seen as part of the methods as well as the findings.

Research Boundaries

The ethical dimensions of research boundaries are important to consider. Some texts talk about the need to avoid being “too friendly” with participants, claiming that it can create unrealistic expectations and can shape or bias data in ways that jeopardize validity; others argue against the researcher being too stoic and removed. These questions are contextual, and where a researcher falls on the friend-to-removed-observer continuum depends on a number of elements, including roles in a setting (e.g., principal and teacher, coworkers), social identities (e.g., elder and younger faculty member, member of dominant and/or nondominant social groups) and positionality in the research (e.g., practitioner researcher [aka “insider”], outside facilitator, coresearchers), and sets of wider considerations such as national, community, and/or organizational power structures and dynamics. Educational researchers should consider, with intention, the methodological implications of how they position themselves in the educational spaces and relationships at the center of their work.

Reciprocity

Reciprocity, or giving back to participants in “exchange” for their time and insights, and incentivizing participation in research are not neutral; rather, they have significant ethical implications. At times, the ways researchers incentivize or reward participants are problematic and potentially inappropriate. These acts therefore have the potential to blur the boundaries and roles of researchers in ways that can negatively impact the research.

There are times when compensation of a few sorts can be appropriate such as providing a meal or snack during interviews or focus groups, compensating participants for travel costs, providing child care if needed to attend data collection sessions, or giving small tokens of appreciation such as an inexpensive pen from one's organization. Even with the best of intentions, some acts of reciprocity lead to unintended negative consequences that can harm the participants and/or the relationships as well as affect the data in negative ways. To be clear, these are ethical issues because they can negatively impact participants and can undermine equity and relational boundaries in a host of ways.

IRBs, Ethics Committees, and Codes of Ethics

Formalized guidelines for ethical research conduct stem from a legacy of problematic research projects in a historically unregulated milieu that caused considerable harm to individuals and groups. Specifically, discussions of the need for ethical regulation, such as through IRBs at universities and ethics committees at organizations, are the outgrowth of research in the medical realm as well as other kinds of experimental research that were impositional, intrusive, and even abusive and that preyed upon historically marginalized and vulnerable populations.

Universities have formed IRBs, which are centralized committees of faculty and staff, and various organizations, including school districts and human service agencies, have developed ethics review boards. These appointed committees are responsible for reviewing research proposals and overseeing ongoing research projects to ensure what is referred to as *beneficence*. Beneficence means that researchers should always have the welfare of participants in mind and should not cause harm to research participants in any way. At their most effective, IRBs can help to point out possible issues in proposed and ongoing research that help further focused thinking about how to safeguard against harm to participants.

The IRB process is often site-specific. Although the same federal guidelines are used across university settings, the local implementation varies considerably. As a necessary early step in most contemporary research, researchers must work to understand what is required of them if they are based at a college or university, if their research will be performed within another institution that may have additional guidelines and approval processes, or if they are working within any group or community that may have norms about research (this includes online communities).

IRBs make distinctions about populations that are considered “vulnerable” (these groups include pregnant women, prisoners, and children), so that special safeguards can be put into place to protect these groups (in the case of pregnant women, this was determined in relation to medical research that could harm a fetus). It is important to note that there are special ethical considerations with respect to particular vulnerable populations and that this term can be defined and interpreted differently depending on the research topic and context. From an ethical perspective, vulnerable populations could include, for example, historically marginalized or otherwise underrepresented or underserved groups and groups that are linguistically different from the norm. It is important to think carefully about these attributions because there is a fine line between understanding the special interests, situations, and needs of groups and projecting need or deficit onto certain groups in ways that reinscribe deficit orientations and structural discrimination.

Informed Consent and Assent

Participant consent in educational research can refer to situations in which researcher(s) seek to (a) access settings and groups to which they are outsiders, (b) obtain data or documents not publicly accessible, or (c) elicit information or data from research participants through interviews, focus groups, questionnaires, observation, writing, and other means. Informed consent should not only be considered transactional but also be thought of as a particular kind of attention to meaningful dialogue with participants about the research and their involvement in it.

Informing participants means that researchers provide information about what is being asked of participants, including demands on their time; what participation will entail; potential risks that could occur; how the data will be handled; who will have access to the data; how the final write-up will be disseminated; the purposes, goals, and methods of the research; who supports or funds the research; and any potential benefits. The process of informing participants can vary by study and context and, in some cases, comes in the form of explaining an informed consent form, which includes an overview of the research goals, a statement about the voluntary nature of the research, the benefits and risks, and a list of the requirements for participation, as well as contact information should they have additional questions or concerns.

What is of primary importance is that participants be informed in ways that are respectful, accessible, and transparent. This means that the form is written and/or communicated in a way that is accessible to participants so that they know exactly what they are agreeing to in terms of time commitment and logistics. It should be clearly communicated to participants that the research is voluntary and that they can refuse to answer any question or withdraw at any time without fear of upsetting the researcher or harming themselves in any way. In addition, consent forms can allow participants to decide if and how they will be represented in a final report and allow them to state if they do or do not consent to be audio or video recorded.

There can be contextual challenges to informed consent. For example, in some populations, formally writing down consent is considered unsafe. There are also times when it can seem culturally inappropriate to ask for formal written agreement. Some of these challenges are contextual and some are cultural. In any case, researchers must understand that informed consent is a concept and process born out of Western institutions, and therefore their cross-cultural, cross-national use should be carefully considered.

Assent refers to the process of providing informed consent to minors. Although assent is not legally mandated or binding, some argue that the ethic of informed consent should apply, in a relational sense, to minors as well. This involves giving minors assent forms that explain the research in developmentally appropriate ways so they understand what they will be involved in or so that they have an opportunity to decline participation even if their parents or guardians have given legal permission.

Confidentiality and Anonymity

Although related, there are important differences between confidentiality and anonymity. *Confidentiality* is related to an individual’s privacy and entails decisions about how and what data related to participants will be disseminated. Confidentiality might mean that pseudonyms will be used and/or other identifying facts will be changed or not disclosed. An example would be not including identifying information such as participants’ names, unique attributes, or job titles in a final report.

Anonymity means that there would be no way for anyone to identify an individual within a sample of participants because data and resulting reports are aggregated rather than individually contextualized or displayed (e.g., “Of the 100 people interviewed, 32 stated that they believe that the professional development practices supported instruction”). Depending on the type of research, anonymity is normally only promised in studies with larger samples.

A multitude of issues should be considered related to confidentiality and anonymity assurances that are made to participants. If these assurances are made, there must be a deliberate plan. Using pseudonyms throughout the research process and not just at the end of a study is one way to help safeguard participants’ identities. To have anonymous data, identifying information should be removed from all study materials, including transcripts and/or coding sheets so that responses cannot be connected to individuals.

Privacy in a Digital World

There are additional challenges to anonymity and confidentiality and to overall participant privacy brought on by the pervasiveness of social media and new technologies, including various forms of digital photography and video, audio and video recordings online, and virtual materials from the Internet, including blogs, chat room discussions, and other publicly accessible data on individuals that could undermine other deidentification processes. Additionally, while data storage and management have always been researchers’ concerns, new cloud technology and transcription services, as well as the incredible mobility of data through e-mail as well as on laptops, smartphones, and electronic storage devices, create a new set of serious ethical issues for researchers to consider related to data security.

Data management and security in an age of technology is a central and ongoing concern in the protection of anonymity and/or confidentiality. Careful consideration of all possible ways that data security can be breached must be considered and strategized at the outset of a study. It must then be attended to throughout the course of the research and even once studies have been concluded. Issues of consent, privacy, and transparency are central to these debates and call up contrasting notions of ethics within and beyond these areas. There are no clear answers to these issues, but the goal is to think about them proactively and to set up specific systems for a study so that the ways to deal with these issues are clear at the outset.

Sharon M. Ravitch

See also American Educational Research Association; American Psychological Association; Assent; Belmont Report; Confidentiality; Ethical Issues in Evaluation; Feminist Evaluation; Human Subjects Protections; Human Subjects Research: Definition of; Informed Consent; Institutional Review Boards; Interviewer Bias; Power; Qualitative Research Methods; Selection Bias; Trustworthiness; Validity

Further Readings

Alvesson, M., & Sköldberg, K. (2009). Reflexive methodology: New vistas for qualitative research. Los Angeles, CA: Sage.

American Educational Research Association. (2011). Code of ethics: American Educational Research Association. Educational Researcher, 40(3), 145–156.

American Psychological Association. (2010). Ethical principles of psychologists and code of conduct (including 2010 amendments). Retrieved from http://www.apa.org/ethics/code/

Cochran-Smith, M., & Lytle, S. L. (2009). Inquiry as stance: Practitioner research for the next generation. New York, NY: Teachers College Press.

Hammersley, M., & Traianou, A. (2012). Ethics in qualitative research: Controversies and contexts. Los Angeles, CA: Sage.

Jacoby, S., & Gonzales, P. (1991). The constitution of expert-novice in scientific discourse. Issues in Applied Linguistics, 2(2). Retrieved from http://escholarship.org/uc/item/3fd7z5k4

Ogden, R. (2008a). Anonymity. In L. M. Given (Ed.), The SAGE encyclopedia of qualitative research methods (Vol. 1, pp. 17–18). Thousand Oaks, CA: Sage.

Ogden, R. (2008b). Confidentiality. In L. M. Given (Ed.), The SAGE encyclopedia of qualitative research methods (Vol. 1, pp. 111–112). Thousand Oaks, CA: Sage.

Ravitch, S. M., & Carl, N. M. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: Sage.

Smith, L. (1999). Decolonizing methodologies: Research and indigenous peoples. London, UK: Zed Books.

Valencia, R. R. (2010). Dismantling contemporary deficit thinking: Educational thought and practice. New York, NY: Routledge.

Zeni, J. (Ed.). (2001). Ethical issues in practitioner research. New York, NY: Teachers College Press.

George Julnes George Julnes Julnes, George

Ethical Issues in Evaluation Ethical issues in evaluation

612

616

Ethical Issues in Evaluation

Ethics in evaluation are focused on what it means for evaluators to “do the right thing.” Although there is considerable controversy about what “the right thing” means, in philosophy as well as practice, there is general agreement that ethical challenges are common in all phases of the evaluation process, from initial contracting to the reporting and use of the findings. This entry discusses various approaches to ethics and their implications for evaluation, ethical challenges in evaluation tasks, two sets of guidelines for conducting good evaluations, and emerging perspectives on ethics in evaluations.

Approaches to Ethics

The nature of ethical behavior has been debated for millennia. From these debates, we recognize that there are different branches of the field of ethics (metaethics, normative ethics, and applied ethics) and different criteria (virtue ethics, deontological ethics, and consequentialist ethics), all having implications for ethics in evaluation today.

Metaethics

Metaethics addresses fundamental questions about moral claims, including questions about whether it is even possible to have actual knowledge about ethics, about what is right and wrong.

Most evaluators would agree that claims about ethical behavior are often not just subjective preferences (e.g., that certain conflicts of interest are unethical) and so accept what is called cognitivism, contrasted with noncognitivism as the view that all judgments about ethics are a matter of personal feelings. However, most that all judgments about ethics are a matter of personal feelings. However, most would also be of skeptical view, called centralism, that some universal principle (e.g., “moderation in all things”) can appropriately define moral behavior in all contexts, instead looking to apply the various ethical principles that best fit specific contexts (noncentralism; e.g., in the specific evaluation situation, it was wrong to exclude particular stakeholders).

Normative Ethics

Normative ethics presumes that it is possible to have standards of ethics that are prescriptive in distinguishing right from wrong (not just descriptive accounts of the standards that people do use), with debates on the appropriateness of the three major positions of virtue ethics, deontological ethics, and consequentialism.

Virtue Ethics

Virtue ethics focus on the quality of an individual’s character, which requires some understanding of *why* a particular action was taken. For example, someone who shared internal documents that revealed illegal government behavior could be viewed as a virtuous whistleblower if the motive was to safeguard the public interest but would be viewed very differently if the motive was a revenge for being passed over for promotion.

Deontological Ethics

Deontological ethics addresses one’s duty and the rights of others. Immanuel Kant provides a major historical example, with his categorical imperative providing unconditional requirements for ethical behavior. More recently, John Rawls’s deontological approach defined moral acts as those that we would converge on if we were ignorant of, and hence not biased by, how they would affect our personal interests. This focus on personal duty and respect for the rights of others guides most efforts to delineate ethical standards and principles for evaluators.

Consequentialism

For consequentialists, behavior is judged by its consequences. John Stuart Mill and Jeremy Bentham provide historical examples of this approach with the development of utilitarianism, with its goal of maximizing aggregate happiness. Modern versions include the Kaldor–Hicks criterion which, again, maximizes aggregate utility but operationalizes utility in monetary terms.

Applied Ethics

Applied ethics takes on the task of articulating principles that help understand and guide ethical behavior in specific real-world situations, including evaluation. Business ethics, for example, addresses the tension between maximizing profit and promoting other outcomes, such as customer and community safety. The American Psychological Association has five general aspirational principles for guiding ethical behavior of practicing psychologists.

Ethical Challenges in Evaluation Tasks

Ethical concerns arise in each phase of an evaluation: (1) evaluation contracting, (2) evaluation planning, (3) data collection, (4) data analysis, (5) evaluation reporting, and (6) evaluation utilization. Highlighting the applied ethical challenges in these tasks can help evaluators counteract them or, if that is not possible, to decline to participate or at least be transparent about the ethics involved.

Evaluation Contracting

Even in the first meeting of evaluators with the people contracting the evaluation, ethical challenges are common. Those contracting the evaluation have vested interests in the outcomes and may try to ensure preordained outcomes by restricting the questions that can be addressed by the evaluation, or by restricting the set of stakeholders allowed to be involved in the evaluation. Challenges can revolve around the evaluators also, as when the evaluators have conflicts of interest or clear differences in values from other stakeholders.

Evaluation Design

Choices in the design of the evaluation invite similar ethical challenges. One critical issue is the choice of outcome indicators used to characterize program effectiveness (e.g., measures of government expenditures can yield very different conclusions about program effectiveness than measures of citizen hardships). Another choice is the type of research design selected (e.g., selection of intervention and comparison groups for a quasi-experimental design), which can easily preordain conclusions about impact and effectiveness (internal validity focuses on these design problems).

Data Collection

As addressed in measurement reliability and validity, data collection can be problematic for ethics in that often the data available are not what was anticipated or are corrupted either deliberatively (e.g., staff are biased in reporting) or through carelessness.

Data Analysis

As with other evaluation tasks, data analysis involves choices that affect the conclusions. It is not difficult, for example, to conduct numerous analyses and select only those that make the program look most promising or only those analyses that yield results consistent with the evaluators' values and biases. On the other hand, inadequate training may result in an evaluator using a data analysis procedure that is inappropriate for the situation (e.g., where a technique is used despite its assumptions being violated; this is the focus of statistical conclusion validity).

Evaluation Reporting

Research indicates that the majority of evaluators have felt pressured, or at least strongly encouraged, to modify how the results of the evaluation are reported. This is particularly problematic for internal evaluators but also occurs when external evaluators have reasons to maintain good relationships with program staff or evaluation funders. Alternatively, reporting may result in program administrators requesting a violation of confidentiality because they want to know which program participants reported negative experiences.

Utilization of Evaluation Findings

How the evaluation results are used can create additional ethical challenges, as when the evaluation sponsor deliberately misrepresents the evaluation findings, either through suppressing negative findings, cherry-picking positive findings, or altering the findings. Evaluators can be complicit in misuse by not providing sufficient guidance for proper use or not objecting to the evaluation findings being used for inappropriate purposes.

Formal Statementson Ethics in Evaluation

Given the many ethical challenges in evaluation, associations around the world have attempted to counter them, with one approach (in Canada and Japan) employing credentialing to increase the competencies of practicing evaluators and another approach (in Canada, United States, and Europe) being to articulate expectations of evaluators. Two of the more established approaches for codifying elements of good evaluations are the *Program Evaluation Standards* developed by the Joint Committee on Standards for Educational Evaluation and the American Evaluation Association's *Guiding Principles for Evaluators*.

Program Evaluation Standards

The *Program Evaluation Standards*, first published in 1981 and updated in 1994 and 2011, provide criteria for what evaluators and others involved in evaluation should do in the interest of promoting effective evaluations. The standards consist of five overarching categories (utility, feasibility, propriety, accuracy, and evaluation accountability) with enumerated standards that all include the word "should," specifying the expectations, or duties, of "good" evaluators and others involved in the tasks. This focus on what should be done is most consistent with the deontological approach to ethics described earlier.

Utility

The utility standards direct evaluations to yield results that meet the information needs of identified stakeholders. For this, evaluations should be led by credible evaluators and should attend to the information needs of the main evaluation audiences. Engaging multiple audiences of people affected by the program in designing the evaluation and interpreting the results helps ensure stakeholders truly need and will use the findings.

Feasibility

The feasibility standards focus on evaluation effectiveness and efficiency. Evaluations should involve project management strategies that make effective and efficient use of resources, should be responsive to local program contexts, and should be sensitive to often-competing multiple cultural and political interests.

Propriety

The propriety standards, with an emphasis on doing the right things in evaluation (including what is fair, legal, and just), state that evaluations should be responsive to stakeholders, negotiating agreements with explicit obligations that are appropriate for the needs and cultural contexts of stakeholders. Evaluations also should protect the human and legal rights and dignity of stakeholders and be fair in addressing stakeholder needs. Finally, there should be transparency in communicating evaluation findings, limitations, and potential conflicts of interest.

Accuracy

The accuracy standards address good methodology, including concerns with reliable and valid measurement, data management, technically adequate designs for valid conclusions, and clear reasoning from findings to conclusions. However, accuracy is also framed in terms of appropriateness in serving intended purposes in targeted cultures and contexts.

Evaluation Accountability

The evaluation accountability standards concern the adequacy of evaluation processes and products and ways to improve them. This requires documenting all activities and also meta-evaluation, the systematic review, or evaluation, of an evaluation, conducted by both internal and external evaluators.

Guiding Principles for Evaluators

The American Evaluation Association established five guiding principles (similar to the five general principles of the American Psychological

Association) for evaluators, rather than standards that provide criteria with which to judge behavior. These five principles are introduced with quotes from the American Evaluation Association statement *Guiding Principles for Evaluators* and then described.

Systematic Inquiry

“Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.” This principle addresses (a) the quality of the design and implementation of an evaluation and (b) ethical challenges from pressures that would weaken the evaluation design and implementation.

Competence

“Evaluators provide competent performance to stakeholders.” The concern here is the extent to which the evaluator is competent in the sense of having the needed (a) education or training, (b) evaluation experience, (c) evaluation expertise, and (d) cultural competence.

Integrity/Honesty

“Evaluators ensure the honesty and integrity of the entire evaluation process.” This principle embodies aspects of virtue ethics—evaluators need the quality of character to be honest and to act with the integrity necessary to build trust with stakeholders.

Respect for People

“Evaluators respect the security, dignity, and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.” Respect for people is addressed by institutional review boards (ensure minimal risks, potential for positive outcomes, confidentiality, and informed consent). Respect also requires recognizing the vulnerability of those being evaluated and ensuring they have meaningful voice, as in the dictum of people with disabilities regarding policy evaluation “not about me without me.”

Responsibilities for General and Public Welfare

“Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.” This principle is challenging both because it requires some notion of what constitutes the “public interest” and because it is based on consequentialist ethics—right and wrong evaluator behavior is not just a matter of doing one’s duty and respecting others, it also needs to have a positive impact on the public interest. This contributes to a tension between evaluation neutrality and advocacy of a particular view of the public interest, entailing the obligation to engage with diverse views of the public interest.

Emerging Perspectives for Evaluation Ethics

Michael Morris and others have written about the need for more research on ethics in evaluation. There has also been consideration of how evaluators might be more self-aware and strategic about the ways that virtue, deontological, and consequentialist ethics can be incorporated in ethical guidelines. Evaluators also need to be more open to emerging perspectives relevant for evaluation ethics.

For example, Thomas Schwandt promotes a form of ethics based on critical systems thinking that highlights the “boundedness” of evaluative inquiry in which the selective inclusion and exclusion of possible facts and values are both inevitable, given the complexity of the systems being evaluated, and yet also something akin to people understanding an “elephant” differently when experiencing only one of its differing parts. Dialogue among those with different ethical and evaluative framings is essential and leaves open the possibility that such dialogue could lead to “better” understandings of what professionalism and ethics mean in evaluation.

Pragmatic ethics (with its evolutionary metaphor) offers similar hope in viewing our current ethical approaches as what Mill called “experiments of living.” Accordingly, our views of ethical behavior are supposed to evolve, and in line with the principle of requisite variety, this evolution benefits multiple, competing frameworks on ethics. This process orientation entails not only hope of improving frameworks for ethics but also a view of ethics as tools that, employed properly, help us make sense of what “doing the right thing” means for evaluators.

George Julnes

See also Advocacy in Evaluation; Conflict of Interest; Interviewer Bias; Social

Justice; Validity; Values

Further Readings

American Evaluation Association. (n.d.). Guiding principles for evaluators. Retrieved from http://www.eval.org/p/cm/ld/fid=51

Fitzpatrick, J. L., & Morris, M. (Eds.). (1999). Current and emerging ethical challenges in evaluation. In New directions for evaluation. San Francisco, CA: Jossey-Bass.

House, E. R. (2015). Evaluating: Values, biases, and practical wisdom. Charlotte, NC: Information Age Publishing.

Joint Committee on Standards for Educational Evaluation. (2011). The program evaluation standards (3rd ed.). Thousand Oaks, CA: Sage.

Julnes, G. (2015). Managing evaluation theories, practices, and communities. American Journal of Evaluation, 36(4), 584–600.

Morris, M. (2011). The good, the bad, and the evaluator: 25 years of *AJE* ethics. American Journal of Evaluation, 32(1), 134–151.

Schwandt, T. A. (2015). Evaluation foundations revisited. Stanford, CA: Stanford University Press.

Schwandt, T. A. (2015). Reconstructing professional ethics and responsibility: Implications of critical systems thinking. Evaluation, 21(4), 462–466.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (Eds.). (1999). Guiding principles for evaluators. In New directions for evaluation. San Francisco, CA: Jossey-Bass.

Dan Ispas Dan Ispas Ispas, Dan

Dragos Iliescu Dragos Iliescu Iliescu, Dragos

Ethical Issues in Testing Ethical issues in testing

616

618

Ethical Issues in Testing

Ethical principles are stated in codes of ethics. Codes of ethics attempt to address professional behaviors that are consistent with (or, by contrast, that violate) moral principles that are broadly accepted in a society and that may also be enforced by public policy. However, professional ethics goes beyond such codes of ethics: Although formal standards are critical benchmarks, they are no substitute for deliberate and conscious ethical judgment. As a result, ethical judgment is about right and wrong beyond the law and more than blind adherence to a standard. This entry discusses ethical issues in testing and describes ethics' codes and guidelines that address testing.

Testing can have important consequences for individuals, groups, and organizations. Tests provide data used to describe and explain past and current characteristics and behaviors, as well as predict future characteristics and behaviors. Tests are important instruments used to inform professional service delivery at the planning, monitoring, or follow-up stage. Tests are also used as gatekeepers in high-stakes contexts such as in the case of personnel selection decisions (helping decide who gets hired for a job), admission to a school or university, release from or admittance to a state-supported program, and others.

Ethical Principles of Psychologists and Code of Conduct

Testing is one of many technologies developed and designed to serve society. Having such important results for society, communities, groups, and individuals, testing is performed under a set of generally accepted rules. Specifically, as with testing is performed under a set of generally accepted rules. Specifically, as with any other professional activity in psychology, testing is covered by the general ethical principles under which psychologists operate.

The American Psychological Association’s *Ethical Principles of Psychologists and Code of Conduct* includes explicit and comprehensive ethical guidelines for testing. The code outlines five general principles: beneficence and nonmaleficence, fidelity and responsibility, integrity, justice, and respect for people’s rights and dignity. In addition to the general principles, the code outlines a set of ethical standards. The assessment part of the code addresses the following 11 issues:

Bases for assessments

The code emphasizes that the opinions and conclusions made by psychologists should be based on sufficient and adequate information and techniques supported by scientific and professional standards.

Use of assessments

The code emphasizes that psychologists use testing in an appropriate manner, based on evidence regarding the psychometric characteristics (validity and reliability) and usefulness of instruments employed, in reference to the specific population tested.

Informed consent

The code emphasizes that psychologists must obtain informed consent, explains the exceptions to this (when testing is mandated by law or other government regulations, when it is a routine activity, or when it is used to evaluate decisional capacity), and details both the information that must be provided and the process of obtaining consent.

Release of test data

The code explains what is understood by the term *test data* and outlines to whom these data can be released and under what circumstances. Test data include raw and scaled scores, responses to stimuli, psychologists’ notes, and recordings.

Test construction

Test construction

The code emphasizes that test authors should use both current substantive scientific knowledge and appropriate psychometric procedures in their test development work.

Interpreting assessment results

The code emphasizes the need to consider the purpose of the test as well as various other circumstances (both test and person related) that may affect the interpretation of the test as well as the need to indicate any significant limitations of any interpretation.

Assessment by unqualified persons

The code emphasizes that psychologists should not promote the use of psychological tests by persons who are not qualified to use them (with the exception of tests used for training under appropriate supervision).

Obsolete tests and outdated test results

The code emphasizes that psychologists should avoid basing their interpretations on test data or testing procedures that are outdated or obsolete.

Test scoring and interpretation services

The code states that psychologists retain responsibility for the scoring of the test, regardless of whether they input data and score the test themselves or use scoring services.

Explaining assessment results

The code states that psychologists have to make reasonable efforts to provide explanations of test results to test takers (unless the testing relationship precludes such feedback).

Maintaining test security

The code emphasizes that psychologists undertake efforts to maintain the integrity and security of test materials.

integrity and security of test materials.

International Codesof Ethics and Guidelines

The trends toward globalization and internationalization have important consequences for testing. Although many countries have laws or codes governing the activity of psychologists, relatively few of these discuss ethics in testing, according to Mark Leach and Thomas Oakland. The lack of specific standards in many codes does not imply that no ethical rules apply to testing in those countries. The general principles of a code, such as the obligation to promote dignity, caring, fairness, or beneficence, are applicable to testing situations. However, the specific way in which such general principles can be applied to a testing situation may not be very clear for test users and may be subject to interpretation.

It is necessary to exercise contextual judgment in testing because more and more often test users are confronted either with issues that are not specifically described in ethics codes or with contexts where no ethical provisions regarding testing exist. For example, many psychologists work in multiple countries, including those without codes of ethics or countries where the ethics codes do not refer to testing. In these situations, test users will turn to the implicit moral standards that are the foundation for laws and ethics codes.

International codes of ethics and guidelines can act as references for best practice, informing professionals, encouraging ethical reflection and decision making, and providing a basis for the development of enforceable standards at the national level. As psychology becomes internationalized, local and international organizations with an interest in tests and testing have begun to assume leadership for the development of an internationally applicable set of norms. One especially important document is the *Universal Declaration of Ethical Principles for Psychologists*, a moral framework and set of principles developed with the support of the two largest international psychological associations, the International Union of Psychological Science and the International Association of Applied Psychology, and also supported by the International Association for Cross-Cultural Psychology. It provides an ethical framework that is based on generally accepted and shared human values and that can be applied to testing situations.

Several other regional documents (i.e., documents developed and promulgated by two or more sovereign countries) are important as general codes. The *Meta-Code of Ethics* of the European Federation of Psychologists’ Associations was first approved in 1995 and revised in 2005. This code incorporates principles found commonly in the codes of member countries, focusing on four principles: respect for a person’s rights and dignity, professional competence, responsibility, and integrity. The Meta-Code encourages psychologists to reflect on these principles when engaged in ethical decision making in their professional work. Although testing is not specifically mentioned, the principles of the Meta-Code of the European Federation of Psychologists’ Associations apply to testing situations.

The five Scandinavian countries (i.e., Denmark, Finland, Iceland, Norway, and Sweden) developed another important regional ethics code, *Ethical Principles for Nordic Psychologists*. Rather than setting standards, the Nordic code highlights principles intended to promote reflection on ethical dilemmas and is generally applicable in professional work. Tests and testing are not discussed, but the statements of the code can be applied to testing situations.

The International Test Commission was established in 1976 and is an international professional association of national test commissions and national psychological associations (full members), other professional associations and testing agencies (affiliate members), and individuals active in the domain of testing. The International Test Commission has developed a set of guidelines and statements that address important issues, among them test use; quality control in scoring, test analysis, and reporting test scores; test adaptation, computer-based and Internet-delivered testing; test security; the use of test revisions, obsolete tests, and test disposal; the use of tests with immigrants and second-language learners; and the use of tests for research purposes. These guidelines discuss issues typically not addressed, or not addressed to this extent, in national, regional, or other international codes of ethics. The International Test Commission guidelines inform, educate, and provide guidance based on scientific scholarship and based on accepted professional practice in the domain of testing.

Dan Ispas and Dragos Iliescu

See also Ethical Issues in Educational Research; Ethical Issues in Evaluation; Second Language Learners, Assessment of; Testing, History of; Tests

Further Readings

Further Readings

Gauthier, J. (2008). Universal declaration of ethical principles for psychologists. In J. E. Hall & E. M. Altmaier (Eds.), Global promise: Quality assurance and accountability in professional psychology (pp. 98–105). New York, NY: Oxford University Press.

Leach, M. M., & Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35 countries. International Journal of Testing, 7, 71–88.

Lindsay, G., Koene, C., Øvreeide, H., & Lang, F. (2008). Ethics for European psychologists. Gottingen, Germany: Hogrefe … Huber.

Oakland, T. (2005). Selected ethical issues relevant to test adaptations. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment (pp. 65–92). Mahwah, NJ: Erlbaum.

Oakland, T., & Iliescu, D. (IN PRESS). Ethical standards, guidelines, and related issues pertinent to international testing and assessment. In F. Leong, D. Bartram, F. Cheung, K. Geisinger, & D. Iliescu (Eds.). The international handbook of testing. Oxford, UK: Oxford University Press.

Oakland, T., Poortinga, Y. H., Schlegel, J., & Hambleton, R. K. (2001). International test commission: Its history, current status, and future directions. International Journal of Testing, 1, 3–32.

David Bloome David Bloome Bloome, David

Judith L. Green Judith L. Green Green, Judith L.

Ethnography

Ethnography

618

623

Ethnography

The use of ethnography in educational research, measurement, and evaluation has become commonplace; yet, there is confusion over what ethnography means and what contribution it makes to educational research, measurement, and evaluation. The meaning of ethnography has evolved since its origin in the early 1900s. Anyone doing ethnography steps into a history of conversations about what the word means and what contribution it makes. This entry summarizes a few of the key conversations about what ethnography means as a way of defining ethnography and its contribution to educational research, measurement, and evaluation.

Ethnography asWriting About the “Other”

The word *ethnography* was originally located in the field of anthropology referring to the written product of an anthropological study of a people, community, or group (not to a methodology). Many scholars name Bronislaw Malinowski as the progenitor of ethnography, although some scholars name W. E. B. Du Bois and his field work in Philadelphia as the historical foundation for their ethnographic research.

During World War I, Malinowski studied the “culture” of the people on the Trobriand Islands. He lived among them, taking field notes of how they organized, gave meaning to, and did family, government, food gathering and meals, rituals, life cycle events, religion, work and economy, and so on. His efforts were long term and based on being there and living with the people of the Trobriand Islands. His perspective was holistic and sought to understand how the parts constitute the whole and the whole constituted the parts; he sought to understand what and how things meant from the perspective of the people studied and he focused on their culture.

One can see in Malinowski's research some of the foundational constructs associated with ethnography: long-term study in which the researcher becomes a participant in the site, the study of a bounded site (i.e., the conceptualization of an identifiable distinct community), holism (i.e., understanding the parts within the context of the whole), attention to the culture of the people, and a privileging of an emic understanding. However, one can also see the foundations of debates and disputes associated with ethnography. His field notes included ethnocentric commentary, raising questions not only about his work but about whether ethnographers can truly eschew their own cultural positionalities. Questions have also been raised about assumptions underlying holism and functionalist perspectives; that is, is it indeed the case that the various cultural systems and institutions of a culture nicely fit together and create a monolithic and coherent whole? Although Malinowski studied a group of people who were relatively isolated and whose culture might be studied as a bounded whole, it seems *prima facie* that overwhelmingly people live in and across many communities and groups and thus the concept of a bounded, integral whole seems a *non sequitur*. Also of importance is the relationship of the researcher and the people in the study. For Europeans, the people of the Trobriand Islands were an exotic other, and the ethnographic study of them—no matter how well intended—marginalized them as strange without sufficiently causing Europeans to see themselves as the other nor to question their own assumed hierarchy and the power relations that underlie exoticism. Such exoticism leaves unspoken questions about the responsibilities of the ethnographer to the participants and the difficulties of seeing the participants as more than objects of a study.

Although fewer scholars identify Du Bois as the scholar whose work was foundational for ethnography, it is useful to contrast his ethnography with Malinowski's to make visible the conversations within which current ethnographers conduct their research. Du Bois's ethnography took place in what was then the seventh ward of Philadelphia, an area that was predominately African American. Du Bois's research approach was methodological triangulation employing statistical data, survey data, interviewing, and participant observation (he lived in the seventh ward during the study). The use of methodological triangulation foreshadowed a discussion later in the field of education and elsewhere about whether ethnography should be labeled qualitative research or whether it constitutes an approach that is distinct from the binary of quantitative and qualitative. Although Du Bois brought to bear an interdisciplinary framework (bringing together the fields of sociology, economics, political science, history, and other social sciences), Du Bois's approach was essentially inductive, leading to the generation of new constructs about the relationship of poverty, structural inequality, racism, and what he called "social uplift." His book *The Philadelphia Negro* was an effort to engage in empirical, nonbiased research that would illuminate social problems and lead to solutions. The emphasis on empiricism was consistent with efforts of that period to demonstrate that the social sciences were scientific (analogous to the natural sciences), but the research was also political and aimed at social change. Du Bois sought to provide an empirical, scientific basis for generating directions for African Americans to address poverty, inequality, and prejudice as well as directions for Whites for addressing the persistence of race-based prejudice and discrimination.

Differences in the studies by Malinowski and Du Bois presage ongoing conversations and debates about ethnography. Although Malinowski focused on a foreign and exotic other, Du Bois focused on his own society and people who looked like him. The contrast leads to questions such as, "What are the advantages, disadvantages, and ethical issues in studying the other or alternatively one's own community?" If one is not a member of the cultural group being studied, does the distance help make visible that which is taken for granted by members? But, if one is not a member, is it possible to truly understand not only the cultural knowledge that members hold but also the affective dimensions of such knowledge and the complexities of consciousness? Although Malinowski's study involved a bounded community, the community Du Bois studied was defined in large part by its structural, historical, economic, and power relationships to the dominant society: How should one conceptualize the boundaries of the group of people in a study? What are the implications of how boundedness is conceptualized for what counts as knowledge and how it is generated? Although Malinowski engaged in ethnography for the purpose of building the knowledge base of a nascent field (and also because he was sequestered in the research site), Du Bois's study was problem oriented and intended to address racism and its consequences for African American people. If a study is problem oriented, does cultural knowledge get inappropriately framed by that problem? Or, is it, as Dell Hymes states, that, "If you want to know a certain thing or a certain class of things directly, you must personally participate in the practical struggle to change reality, to change that thing or class of things" (1974, p. 209).

Ethnography as the Study of Culture

One purpose of ethnographic studies is describing the culture of a group, community, institution, and so on. However, recently within anthropology and in the social sciences more generally, questions have been raised about the concept of culture (and thus relatedly, about what ethnography is the study of). The concept of culture has traditionally been used as a bounded attribution, as in the "culture of ____________" where the blank can be filled in with the name of a particular group of people (e.g., Jews, Latinos, women), a place (e.g., America, New York, Levittown, the Western Cricket Club), or an institution (e.g., Westville High School, the London Road Madrasah, Ms. Lee's preschool classroom). Such formulations of culture are problematic because they presuppose separation and integrity of the cultural category, whereas scholarship and experience suggest not only contact and influence among cultures but hybridity, integration, and dynamic flow within and across such socially constructed categories. In response, some social scientists have abandoned the use of culture, while others have redefined culture as a verb.

A traditional definition of culture defines it as the complex of knowledge, belief, art, morals, law, custom, and any other capabilities and habits that people of a particular group share. But such a definition suggests that culture is a set of discrete phenomena, given and static rather than produced and producing. More recent definitions of culture focus not on material phenomena, behavior, rites and rituals, institutions, and so on but rather on the organization, forms, and models of activity that people in a group collectively hold—shared mental models and standards for perceiving, interpreting, thinking, acting, feeling, believing, valuing, and using language. Culture is located in the minds of individuals. However, some scholars view this definition of culture as problematic because it does not address the social nature of everyday life, as people together make meaning and act on the world over time. Thus, another recent definition of culture focuses on the semiotic nature of public action: How people through their interactions with each other assign importance to what they are doing. Still another definition holds that culture is a verb placing emphasis on asking what culture does and on culture as a signifying process, the active construction of meaning. The meaning here goes beyond the transmission and decoding of a communicative message and the meaning of a sign *per se* to meaning that is socially constructed and reconstructed, both situated and in flow, denotational and indexical, intertextually and intercontextually contextualized, in time and over time, embodied and entextualized, and essentially dialectic. By defining culture as a verb, questions can be asked about who is doing what, with whom, how, when, where, and with what significance and impact for people's lives.

Ethnography as a Logic of Inquiry

A distinction can be made between an ethnography and an ethnographic study. An ethnography employs theoretical frameworks grounded in cultural, social, and linguistic anthropology to the study of a community. An ethnographic study, while also employing theoretical frames derived from anthropology, may take as its object of study a smaller unit than a community such as a classroom. Both an ethnography and ethnographic studies are distinct from ethnographic methods (tools) such as long-term participant observation and interviewing. Ethnographic tools do not necessarily make a study ethnographic. It depends on how the tools are used. Ethnographic tools may be embedded in a qualitative, phenomenological study; that study would not be an ethnography or an ethnographic study *per se*. In brief, ethnography and ethnographic research is not a method or a set of research tools but a logic of inquiry grounded in anthropological theories with a distinct orientation to epistemology and ontology.

The logic of inquiry of ethnography and ethnographic studies can be characterized as abductive reasoning that is iterative and recursive. Cultural patterns, cultural models, and the recurrence of shared standards and expectations for how people should act, believe, feel, talk, and value, and cultural models are identified through field work, interviewing, collection of artifacts, and so on, and then examined for the relationship of parts and wholes and emic interpretations within and across social situations. Both within the corpus of data of the study and as applied to other settings, activities, groups, and social situations, one seeks the recurrence of these patterns.

The search for patterns needs to be understood broadly. They may be a pattern of explicit actions such as classroom conversations structured by conversational moves of teacher initiation followed by student response followed by teacher evaluation; the recurrence of a particular cultural ideology such as holding the individual as a unit of analysis across social institutions such as law, education, and religion (as opposed to, e.g., the family as a unit of analysis); a narrative or narrative structure that is recurrent across situations or institutions such as narratives in which good always triumphs over evil; the structuring of social and cultural life by particular binaries such as moral/immoral; by recurrent definitions of time such time as linear progress as opposed to time as cyclical; among other ways in which patterns might be formulated. Questions can be asked about how these patterns provide meaning in people's lives and how they structure their daily lives.

The logic of inquiry in ethnographic studies contributes to the knowledge base in various fields by providing cases and insights. The cases provided may be of particular cultural patterns within and across various settings, social institutions, groups, activities, and so on, or they may be cases of how ethnographic studies have been conducted and knowledge generated. Heuristically, there are four types of cases. A typical case reports a pattern that is recurrent in analogous situations; a representative case reports the recurrence of a related set of diverse patterns, all of which represent a common process; a critical incident case reports an event that stands at the nexus of rising and falling action (analogous to the climax of a story); and a telling case reports an event (or series of events) whose nature is such that taken for granted and "invisible" cultural processes are made visible (usually because of a need to repair a situation by making those processes visible to those present).

One of the key questions asked of ethnographic case studies is how people might gain knowledge from them either for the purpose of policy making and implementation or for the purpose of improving educational practice. One answer to this question is that ethnographic case studies provide a grounded warrant (a warrant derived from the empirical ethnographic study) for identifying the processes and phenomena that need to be studied with regard to a particular situation or educational problem (sometimes called grounded theoretical constructs) as well as for generating appropriate research questions (sometimes called grounded hypotheses). Another answer requires reconceptualizing the nature of knowledge derived from educational research, measurement, and evaluation. Rather than rules and principles to follow based on statistical generalizations of the relationships of various factors, knowledge from ethnographic studies is abductive (analogic reasoning). Like chess masters who continue learning to improve their games by reading the games of others so that they can anticipate patterns of moves even if no two games will ever be exactly the same, researchers and practitioners across fields learn from ethnographic studies by examining the patterns reported and using them (and recontextualizing them) to understand abductively new situations.

Ethnography of and in Education

A distinction can be made between ethnography *of* education and ethnography *in* education. Ethnography *of* education involves anthropologists and other scholars using education as a research site to which they bring theoretical frames, tools of inquiry, and a history from their fields and disciplines. Their research may be oriented to contributing to their own fields and disciplines or it may be applied; the theories, questions, and goals of this research are framed by the home disciplines and academic fields of the researcher and not necessarily by educators' needs, issues, or concerns. Ethnography *in* education, meanwhile, is grounded in knowledge derived from the field of education, the historical background of ethnography in anthropology and sociology, the historical background of ethnographic studies *in* education, and guided by education questions, purposes, needs, and concerns primarily derived from ethnographers in the field of education.

The distinction between ethnography of education and ethnography in education is key to understanding the contribution of ethnographic studies to conceptualizing education and to education policies and practices. In the ethnography of education, schools are only one site of education, and when studying schools, the teachers and students are others. The distance between ethnographer and participant is productive in making the familiar strange and thus producing new insights. Ethnography in education refers to ethnographic studies by scholars and practitioners who are located in education and who are conducting ethnographic research that is problem oriented. Although still grounded in cultural, social, and linguistic anthropology, other disciplinary perspectives may also be incorporated. One contribution of ethnography in education is to offer new ways of conceptualizing core educational processes such as learning, instructional, achievement, failure, literacy, curriculum, and assessment. Also included in ethnography in education are those studies reporting the efforts of educators to engage their own students in ethnographic study as part of the academic curriculum and instructional processes.

Ethnography and the Definition of Education

Ethnography and the Definition of Educational Research, Measurement, and Evaluation

The contribution of ethnography to educational research, measurement, and evaluation, while including findings and knowledge about what happens in diverse classrooms, schools, and other education settings through careful, detailed field work, also and perhaps primarily has provided nuance, complexity, and problematization to taken-for-granted concepts in educational practice, policy, and research. It has questioned how who the researcher is (and what researcher relationship is with the research site and participants) affects what knowledge is generated and the nature of that knowledge. It has emphasized knowledge that is emic-oriented, placing such knowledge in the context of the educational site itself as well as across sites. It embodies a logic of inquiry that seeks to understand the culture of a people and a place while simultaneously questioning the concept of culture. It defines knowledge as an abductive practice requiring the engagement of both the producer of the knowledge (the researcher) and the consumer of the knowledge (educators and others addressing the nature, conduct, and policies of educational institutions).

David Bloome and Judith L. Green

See also Cross-Cultural Research; Naturalistic Inquiry; Qualitative Research Methods

Further Readings

Du Bois, W. E. B. (1996). The Philadelphia Negro: A social study. Philadelphia: University of Pennsylvania Press. (Original work published 1899) Geertz, C. (1973). The interpretation of cultures: Selected essays. New York, NY: Basic Books.

Goodenough, W. (1981). Culture, language, and society. Menlo Park, CA: Cummings.

Green, J., Dixon, C., & Zaharlick, A. (2003). Ethnography as a logic of inquiry. In J. Flood, D. Lapp, & J. Squire (Eds.), The handbook for research in the teaching of the English language arts. Mahwah, NJ: LEA.

Malinowski, B. (1922). Argonauts of the Western Pacific. London, UK: Routledge.

Spindler, G. (Ed.). (2000). Fifty years of anthropology and education 1950–2000: A Spindler anthology. Mahwah, NJ: Erlbaum.

Street, B. V. (2013). Anthropology in and of education. Teaching Anthropology, 3(1), 57–60.

Dominica McBride Dominica McBride McBride, Dominica

Evaluation

Evaluation

623

624

Evaluation

Evaluation is a process, discipline, and, in some cases, an intervention in and of itself. It entails the systematic application of social science research to plan for and learn about the impact of policy, performance, programs, or initiatives in order to create, further, or sustain social change. The policies, performances, and initiatives being evaluated are called *evaluands*. Evaluation is performed in sociopolitical environments and political influences, and their implications must be considered throughout the process.

From struggles to provide quality education and public health, to environmental dilemmas, societies across the world face issues that often require planning, policy, and subsequent action to address. Unfortunately, strategies often do not obtain the desired effect because projects are not implemented as planned, policies are disconnected from the communities they are supposed to benefit, or programs are not well planned.

According to evaluation expert Michael Scriven, evaluation examines the merit and worth of the evaluand. However, the examination is often not the end but the means to making change through contributing to a decision or using the results for advocacy purposes, as in the transformative paradigm, a framework for evaluation that places importance on groups that have been marginalized. For example, the purpose of an evaluation of a school district’s new program for reading by third grade would be to assess how effective the program is for all students, especially disadvantaged students in the district. Using this example, the process of evaluation would include:

1. identifying and engaging stakeholders, or people who have different stakes in the process, such as students, teachers, parents, school administrators, and communities.
2. constructing relevant and answerable questions, such as “To what extent did students enhance their reading skills?” “What worked and did not work? For whom? Why?” “What are the most pressing needs for low-income K–3rd grade students in the district?” and “What community assets can be used to address those needs? How?”
3. choosing data collection methods, such as tracking reading grades, interviewing students, surveying teachers, and holding focus groups with parents.
4. collecting and analyzing the data. The data and subsequent analysis used to answer questions such as these are, ideally, used to make changes necessary to effectively address the problem of focus, such as reading by third grade.
5. synthesizing and disseminating the results.
6. taking appropriate action to help those results make a difference in the evaluand and ultimately for the intended community.

Given the complexity, human dynamics, and sociopolitical nature of evaluation, there are various skills needed to successfully complete an evaluation. Skills in areas such as analysis, social skills, project management, critical self-reflection, negotiation, and advocacy allow evaluators to execute the technical components and navigate the social world of evaluation to make social change.

Dominica McBride

See also American Evaluation Association; Culturally Responsive Evaluation; Data-Driven Decision Making; Goals and Objectives; Outcomes; Program Evaluation; Transformative Paradigm

Further Readings

Davidson, E. J. (2005). Evaluation methodology basics: The nuts and bolts of sound evaluation. Thousand Oaks, CA: Sage.

Mertens, D. M., & Wilson, A. T. (2012). Program evaluation theory and practice: A comprehensive guide. New York, NY: Guilford Press.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). Evaluation: A systematic approach (7th ed.). Thousand Oaks, CA: Sage.

Dominica McBride Dominica McBride McBride, Dominica

Evaluation, History of

Evaluation, history of

624

626

Evaluation, History of

Many disciplines have contributed to the field of evaluation, but education and the social sciences have long played major roles in this area. There have been both philosophical and cultural shifts in the field of evaluation over time. Egon Guba and Yvonna Lincoln have described evaluation's ideological progression as proceeding through four generations: measurement, description (objectives focused), judgment (decision focused), and constructivism. Furthermore, the history of evaluation reflects the broader societal story of economic ebbs and flows, the burgeoning of measurement and standardization, and the emergence of multicultural inclusion. This entry looks at influences on evaluation and discusses shifts in the field's ideology and practice over time, with a focus on the 20th and early 21st centuries.

The first recording of evaluation was in the late 18th century, with the onset of testing to assess performance. Around this time, evaluation was also used to assess programs for public health and occupational training. In the 19th century, the use of evaluation became more formal both within the field of education and in the military with the standardization of production. In this era, an emphasis on testing emerged as well as the use of scientific experiments using control and comparison groups for evaluating education.

In the early 20th century, a focus on district-level efficacy in education arose. By the 1930s, evaluation was being used in a variety of areas with social science research being applied to learn the effectiveness of various social programs. Evaluation in this era ostensibly emphasized measurement.

In the 1930s and 1940s, Ralph Tyler's work on the 8-year study created a foundation for evaluation as a discipline. Some think of Tyler as “the father of educational evaluation.” Prior to this period, there had been a focus on quantitative methods and testing in particular. Tyler emphasized objectives and outcomes, framed objectives in terms of behavior, and urged that multiple methods be used in learning the extent to which objectives were manifested as outcomes. He also asserted the need to include teachers in the development of curricula and programming as well as tailor the research on effectiveness to the school setting. He recognized the necessary variance in objectives between schools and that one method or set of methods cannot be used in all situations.

Stafford Hood, a prominent African American evaluator, found that notable African American educational evaluators in the Tylerian period, such as Reid E. Jackson and Leander Boykin, were left out of the overarching discussion on evaluation, despite scholarly contributions to methods and approaches. Boykin, for example, was beyond his time, in asserting that multiple stakeholders, from students and parents to teachers, should play important roles in the evaluation process. This exclusion reflected the overarching social predicament in the United States around race and equity.

1940s and Beyond:Government Policies and Boost to Evaluation

World War II marked another pivotal point in evaluation’s history. During the war, the U.S. Army applied social science methods to study the morale of soldiers and citizens and evaluate personnel policies. Following the war, the United States had a strong sense of optimism and abundance. With this wealth and perceived power, there was greater investment in more and enhanced education and social programming. It was thought that poverty could be alleviated through programs and everyone would rise economically. With greater funding toward education and social programs, evaluation benefited and grew in uses, applications, and structure. By the 1950s, the use of evaluation had spread to programs from psychopharmacology to community initiatives. In 1958, as the U.S. government invested millions into curricula to hone students’ knowledge and skills in areas related to defense through the National Defense Education Act, it included evaluation to measure effectiveness. Evaluation also was being used in many countries outside of the United States from Latin America to Africa.

The 1960s War on Poverty sparked another change and boost in evaluation. The Elementary and Secondary Education Act of 1965, intended to ensure the education needs of disadvantaged students were met, included a requirement for educators to measure their effectiveness in this area. This shift in national focus placed more emphasis and value on social justice. Stakeholders wanted to know how the War on Poverty and related initiatives were affecting their intended beneficiaries. Evaluation was used to make that judgment and affected subsequent decisions.

Many of the programs begun under the War on Poverty were found to be ineffective or having an insufficient benefit in relation to the cost. Therefore, there was reluctance to support social programs through the 1970s. With this doubt, government funding eventually decreased and so did the investment in evaluation. This lull in the general use and application of evaluation continued through the 1980s, but investment in evaluation rebounded in the 1990s. Throughout these three decades and beyond, the field continued to build structure.

In the 1970s, academic journals, professional associations, and university courses on evaluation emerged. *Evaluation Review* was the first evaluation journal published in 1976. The Evaluation Research Society and Evaluation Network combined in 1986 to create the American Evaluation Association. Its mission is:

> to improve evaluation practices and methods, increase evaluation use, promote evaluation as a profession, and support the contribution of evaluation to the generation of theory and knowledge about effective human action. (American Evaluation Association, 2016, n.p.)

The association provides scholarly work on evaluation through two journals, *American Journal of Evaluation* and *New Directions for Evaluation*, an annual conference and institute, a daily blog, and webinars. It also hosts a listserv called EvalTalk. In addition to these journals, there are several others focused on evaluation such as *Evaluation: The International Journal of Theory, Research and Practice*, and *Educational Evaluation and Policy Analysis*.

2000s and Beyond: Greater Focus on Culture and

Social Justice

Prior to 2000, there was not a prominent focus on culture and marginalized groups within the field of evaluation, despite the external focus on social and educational programs for disenfranchised communities. The 2000s saw the development of more constructivist evaluation approaches that explicitly promote social justice (e.g., deliberative democratic evaluation and transformative participatory evaluation).

In 2003, Hood and Melvin Hall started the Relevance of Culture in Evaluation Institute, a project dedicated to defining culturally responsive evaluation (CRE). Also around this time, a small group of evaluators convened to write a statement on cultural competence in evaluation. In 2011, the Public Statement on Cultural Competence in Evaluation was adopted by the American Evaluation Association and promoted thereafter by a working group dedicated to the dissemination and use of the statement. This period also included initiatives to increase the racial/ethnic diversity of evaluators, such as the Graduate Education Diversity Internship program, which provides training and service learning opportunities in evaluation to graduate students of color.

Also during this time, the *Program Evaluation Standards* were developed by the Joint Committee on Standards for Educational Evaluation. These standards provide direction and criteria for quality evaluations, including around utility, feasibility, propriety, accuracy, and for meta-evaluations. The professionalization and explicit identification of what constitutes a “good” evaluation have led to discussion of the desired competencies to meet these criteria.

Ongoing Shifts in Field

Although professionals in the field continue to strengthen evaluation, it is also being influenced by shifts in many disciplines, including business, finance, and economics. An example of this has come with the increasing emphasis in business on corporate social responsibility and the development of market-based social impact structures (e.g., social impact bonds). Out of this milieu, there has been an emerging focus on social return on investment. These business-related ideas and terms are becoming part of the field of evaluation, affecting how some frame criteria for program success. As other fields affect evaluation, evaluation professionals are reaching out to other professional groups for collective learning and collaboration.

Dominica McBride

See also American Evaluation Association; Culturally Responsive Evaluation; Data; Data-Driven Decision Making; Educational Research, History of; Evaluation; Great Society Programs; Testing, History of

Further Readings

American Evaluation Association. (2016, January). AEA mission, vision, values and governing policies. Retrieved from http://www.eval.org/p/cm/ld/fid=13

Guba, E. G. & Lincoln, Y. S. (1989). Fourth Generation Evaluation. Newbury Park, CA: Sage.

Hogan, R. L. (2007). The historical development of program evaluation: Exploring the past and present. Online Journal of Workforce Education and Development, 4. Retrieved July 14, 2016, from http://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1056…context=ojwed

Hood, S. (2001). Nobody knows my name: In praise of African American evaluators. New Directions for Evaluation, 92, 31–43.

Madaus, G. F. (2004). Ralph W. Tyler’s contribution to program evaluation. In M. C. Alkin (Ed.), Evaluation roots: Tracing theorists’ views and influences (pp. 69–79). Thousand Oaks, CA: Sage.

Mertens, D. M., & Wilson, A. T. (2012). Program evaluation theory and practice: A Comprehensive guide. New York, NY: Guilford Press.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). Evaluation: A systematic approach (7th ed.). Thousand Oaks, CA: Sage.

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2000). Evaluation models: Viewpoints on educational and human services evaluation (2nd ed.). Boston, MA: Kluwer Academic Publishers.

David Fetterman David Fetterman Fetterman, David

Jason Ravitz Jason Ravitz Ravitz, Jason

Evaluation Capacity Building Evaluation capacity building

626

628

Evaluation Capacity Building

Evaluation capacity building (ECB) is an approach for helping people learn how to conduct an evaluation and think evaluatively in the process. ECB is designed to help people acquire evaluation knowledge, skills, and attitudes and apply them appropriately in practice. According to Donald Compton, Michael Baizerman, and Stacey Hueftle Stockdill, ECB involves efforts to develop and sustain practices within organizations and make the use of evaluation processes and practices routine. The goal of ECB is to increase evaluation capacity in order to increase the probability staff members will assess and document the implementation and impact of their programs.

ECB is being used by community-based organizations, religious institutions, government agencies, foundations, and private industry. It represents a group or institutional understanding of the value of evaluation to improve performance. There are many contextual factors that influence the effectiveness of ECB professional learning and development. For example, access, expense, motivation, incentive, and degree of difficulty influence the type of ECB training selected and/or needed. A culture of inquiry and supportive leadership enhances the quality of ECB.

The approaches used in ECB training vary from providing templates and information sources to complete immersion and control of an evaluation. Experienced evaluators may only need a tip in a blog posting to enhance their capacity. However, an individual new to the field may enroll in a workshop or course for a more extensive introduction to evaluation. The remainder of this entry discusses resources for ECB training and different approaches to involving stakeholders in conducting an evaluation.

stakeholders in conducting an evaluation.

Training Resources

Professional Associations

Professional associations offer ECB training, including the American Educational Research Association, American Public Health Association, Australasian Evaluation Society, Canadian Evaluation Society, Southeast Evaluation Association, and the American Psychological Association. For example, the American Evaluation Association has numerous ECB training resources, including its AEA365 blog; its 20-minute AEA Coffee Break webinars; its longer, more in-depth eStudy webinars; and workshops at its conferences.

Coursework

Formal coursework in ECB is offered in both face-to-face and online programs, including those at Claremont Graduate University, Western Michigan University, University of Minnesota, University of Connecticut, Syracuse University, University of Illinois Urbana–Champaign, University of California, Los Angeles, University of Cape Town, and University of Wisconsin–Stout.

Additional OrganizationsProviding Evaluation Training

Government agencies and private organizations, including the U.S. Government Accountability Office, the National Science Foundation, and the RAND Corporation, provide both online and printed evaluation training materials, including the U.S. Government Accountability Office and the National Science Foundation. One of the most widely used evaluation capacity building training documents is *Getting to Outcomes 2004*, a publication authored by Matthew Chinman, Pamela Imm, and Abraham Wandersman and published by RAND. It provides a 10-step approach to conducting a self-assessment.

Extensive materials are also available from the Centers for Disease Control and Prevention. Its Program Performance and Evaluation Office provides an extensive list of useful ECB resources. Program-specific evaluation training extensive list of useful ECB resources. Program-specific evaluation training materials include “Introduction to Program Evaluation for Comprehensive Tobacco Control Programs,” “Framework for Program Evaluation,” and “Key Outcome Indicators for Evaluating Comprehensive Tobacco Control Programs.” The Republic of South Africa distributes “Basic Concepts in Monitoring and Evaluation.” Foundations provide evaluation training materials, such as the W. K. Kellogg Foundation, Annie E. Casey Foundation, the Robert Wood Johnson Foundation, the United Way, Amherst H. Wilder Foundation, and the Northwest Connecticut Community Foundation.

There are extensive online resources for ECB; some of these are listed in the suggested readings and websites at the end of this entry.

Stakeholder InvolvementApproaches to Building Capacity

One of the most effective and sustainable forms of ECB involves direct stakeholder involvement in conducting an evaluation. The primary stakeholder involvement approaches to evaluation, with a focus on capacity building, include collaborative, participatory, and empowerment evaluation. A brief comparison of these approaches, focusing on the role of the evaluator, can help evaluators select the most appropriate approach for the task at hand, given their local context. Elements to consider when selecting an approach include the purpose of evaluation, time, resources, existing community, staff members and leadership, program participants, evaluator capacity, organizational and/or community culture, and commitment to capacity building.

Collaborative evaluators are in charge of the evaluation, but they create an ongoing engagement between evaluators and stakeholders. This can contribute to a stronger evaluation design, enhanced data collection and analysis, and results stakeholders understand and use. *Participatory* evaluators jointly share control of the evaluation. Participatory evaluations range from having program staff members and participants participate in an evaluator’s vision of the evaluation to having an evaluation that is jointly designed and implemented by the evaluator and program staff members. Participants are encouraged to become involved in defining the evaluation, developing instruments, collecting and analyzing data, and reporting and disseminating results.

Empowerment evaluators view program staff members, program participants, and community members as in control of the evaluation. However, the empowerment evaluators do not abdicate their responsibility and leave the community to conduct the evaluation solely by itself. They serve as critical friends or coaches to help keep the process on track, rigorous, responsive, and relevant. Empowerment evaluations are not conducted in a vacuum. They are conducted within the conventional constraints and requirements of any organization. However, participants determine how best to meet those external requirements and goals.

The type and level of ECB is determined by the type of stakeholder approach selected. For example, collaborative evaluation capacity training will enhance skills associated with evaluation design, data collection, and analysis. However, collaborative evaluation capacity training does not prepare stakeholders to lead an evaluation in the future, in part because the evaluator remains in charge and there is little provision or opportunity for staff or participants to take responsibility for the evaluation. Similarly, participatory evaluation may provide evaluative capacity building in the areas of evaluation design, data collection, analysis, and reporting. It may also provide development in the area of shared decision making. However, evaluative capacity building in a participatory approach may only partially prepare people to conduct an evaluation in the future because decision making is not completely handed over to staff and/or community members.

Empowerment ECB training enhances evaluation design, data collection and analysis, and reporting. It is also expected to prepare staff and participants to implement their own evaluations in the future in part because staff and participants are placed in charge of the evaluation. Empowerment ECB training also contributes to self-determination for programs and stakeholders.

David Fetterman and Jason Ravitz

See also American Educational Research Association; American Evaluation Association; American Psychological Association; Empowerment Evaluation; Evaluation, History of; Evaluation Versus Research

Further Readings

Centers for Disease Control and Prevention Program Performance and

Evaluation Office. (2016, November 17). A framework for program evaluation. Retrieved from cdc.gov/eval/framework

Chinman, M., Acosta, J., Hunter, S. B., & Ebener, P. (2015). Getting to outcomes: Evidence of empowerment evaluation and evaluation capacity building at work. In D. M. Fetterman, S. Kaftarian, & A. Wandersman (Eds.), Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability. Thousand Oaks, CA: Sage.

Fetterman, D. M., Kaftarian, S., & Wandersman, A. (Eds.). (2015). Empowerment evaluation: Knowledge and tools for self-assessment, evaluation capacity building, and accountability. Thousand Oaks, CA: Sage.

Garcia-Iriarte, E., Suarez-Balcazar, Y., Taylor-Ritzler, T., & Luna, M. (2011). A catalyst-for-change approach to evaluation capacity building. American Journal of Evaluation, 32(2), 168–182.

Harvard Family Research Project. (n.d.). Evaluation. Retrieved from hfrp.org/evaluation

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. American Journal of Evaluation, 29(4), 443–459.

Ravitz, J. (2016). Evaluation planning worksheet. In D. M. Fetterman & J. Ravitz (Facilitators), American Evaluation Association, “Coffee Break” webinar. Using the power of rubrics and technology for empowerment evaluation at Google and Beyond.

Stockdill, S., Baizerman, M., & Compton, D. (2002). Toward a definition of the ECB process: A conversation with the ECB literature. New Directions for Evaluation, 93, 1–25, 233–243.

Websites

American Evaluation Association Collaborative, Participatory, and Empowerment (CP…E) Evaluation Topical Interest Group (TIG): comm.eval.org/cpetig/home

American Evaluation Association STEM Education and Training Topical Interest Group: comm.eval.org/stemeducationandtraining/home

Better Evaluation: betterevaluation.org

Community Tool Box: ctb.ku.edu/en/table-of-contents

EvalPartners: evalpartners.org

Foundation Center’s Tools and Resources for Assessing Social Impact: trasi.foundationcenter.org/

Free Resources for Program Evaluation and Social Research Methods: gsociology.icaap.org/methods/

Online Evaluation Resource Library: oerl.sri.com

University of Wisconsin Cooperative Extension, Program Development and Evaluation (PD…E): fyi.uwex.edu/programdevelopment/

Gail Vallance Barrington Gail Vallance Barrington Barrington, Gail Vallance

Evaluation Consultants Evaluation consultants

628

629

Evaluation Consultants

Evaluation consultants provide their expertise in evaluation and applied research on a temporary basis to a wide range of organizations. They assess program effectiveness and efficiency, answer policy questions, provide advice, and support organizational change. They design and implement tools to collect relevant information; review, analyze, and synthesize that information; and make judgments, report findings, and provide recommendations to improve organizational performance. As contractors, not employees, they have no authority to implement the changes they recommend.

Of course, they must demonstrate the knowledge, skills, and competencies required of any evaluator, as James Altschuld suggests, including knowledge about ethics; performance measurement; systems theory; history, theories, models, and types of evaluation; research design; sampling and measurement; capacity building; communications; and both interpersonal and project management.

To gain these skills, they should have a master’s degree in their discipline of choice; in some research-oriented fields, such as health care, credibility can be enhanced through doctoral studies. Specific evaluation training is also available through specialized postsecondary programs or training opportunities offered by professional evaluation organizations such as the American Evaluation Association. Ultimately, extensive field experience is the best trainer.

As independent practitioners, evaluation consultants need strong business skills to select an appropriate ownership structure, understand price-setting methods, track and manage time, calculate fees and expenses, and manage accounts payable and receivable. Managing their cash flow issues is critical to their survival.

survival.

They need to obtain appropriate insurance for risk management. They must understand and manage contracting processes so that expectations and timelines are clear and potential pitfalls are avoided. To work on larger, more challenging projects, they can hire subcontractors or associates on a short-term basis, but team work requires additional management and supervision skills.

While addressing these many demands, consultants must continue to search for their next evaluation project. Proven marketing techniques include subscribing to government bidding services, registering on vendor lists, creating a presence on social media, offering webinars and workshops, and networking informally. Once a potential project is identified, the consultant prepares a proposal and must win the contract before the cycle begins again.

This independent role requires a number of personal strengths—self-confidence, negotiation skills, an ability to multitask, reflexivity, resilience, and enough financial stability to sustain themselves between projects. Consultants must model ethical business and research practice because their livelihood depends upon the reputations they build.

Gail Vallance Barrington

See also American Evaluation Association; Educational Researchers, Training of; Ethical Issues in Evaluation

Further Readings

Altschuld, J. W. (2005). Certification, credentialing, licensure, competencies, and the like: Issues confronting the field of evaluation. The Canadian Journal of Program Evaluation, 20(2), 157–168.

Barrington, G. V. (2012). Consulting start-up and management: A guide for evaluators and applied researchers. Thousand Oaks, CA: Sage.

Helen L. Chen Helen L. Chen Chen, Helen L.

Evaluation Versus Research Evaluation versus research

629

632

Evaluation Versus Research

Although the title of this entry implies a dichotomy between evaluation and research, in practice, evaluation is a category of applied research that employs similar methodologies but for a different purpose, focus, and audience. *Evaluation* is described as a systematic process by which the value, effectiveness, or significance is determined and can be applied to a range of processes, programs, policies, products, and personnel. *Research* is referred to in the U.S. regulation governing research on human subjects as the "systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge" (Protection of Human Subjects, 45 C.F.R. §46.102(d)).

This entry addresses the similarities and differences between evaluation and research according to their purpose and intention, methods and analysis approaches, and outcomes and audience. A case study is used to provide a more concrete illustration of how evaluation and research are applied in practice. The entry concludes with additional issues for consideration and recommended readings.

Purpose and Intention

The fundamental difference between evaluation and research lies in the purpose and motivation behind the work. The goal behind research is to contribute to a body of knowledge and theory in a field. An intellectual question and the opportunity to contribute to a broader understanding of a subject or body of study often motivates researchers. In contrast, the focus of evaluation, as a form of applied research, is to judge merit or quality as determined by the interests of various stakeholders, such as funding agencies. As a result, evaluators often view the work that is being conducted with a different lens and a broader perspective than that of researchers. These evaluators may work internally within the organization or be commissioned externally under a contract.

It should be noted that the emphasis on “producing generalizable knowledge” is a critical criterion for determining whether a proposed study requires institutional review board or human subjects review approval. Typically, plans to publish results in an abstract, poster, conference presentation, or journal article demonstrate an intention to contribute to “generalizable knowledge.”

Although evaluation often involves interactions with living individuals, human subjects review may not be necessary if the evaluation activities are part of the normal educational activities within a course and individually identifiable information is not being collected or shared beyond the specific context. For example, if the purpose of the data collection is primarily to improve upon the teaching activities and techniques for a particular course or program, then this is not considered human subjects research. However, the distinguishing characteristic of research is based on the *intention* to advance knowledge and not whether findings are actually published or not. Irrespective of the institutional review board approval, it is generally a good practice in evaluation to err on the side of following the guidelines of informed consent and ensuring that all participants understand the instructional purpose of the tasks or activities.

Methods and Analysis Approaches

As articulated in their respective definitions, while the questions that guide and drive evaluation and research are different, both employ similar methods, theories, and analysis approaches to achieve results and ensure reliable and valid outcomes. In designing rigorous studies to address a particular question, quantitative, qualitative, or mixed methods from a range of disciplines and professions may be employed. Although researchers typically have formal training in these areas, evaluators tend to be more diverse in their knowledge and skills. One approach to measuring the quality of the research design is through the peer review process. In contrast, the credibility and usefulness of evaluation findings do not rely solely on the results themselves but require a broader understanding of context and the ability to integrate and synthesize across multiple perspectives, data sets, and circumstances.

Outcomes and Audience

The outcomes of most research studies are represented in conference presentations, journal articles, and other forms of publication. This research output may become part of the scholarly literature in a particular field or discipline where it is accessed by a broad academic community consisting of researchers, practitioners, and educators. In contrast, evaluation products typically consist of some kind of internal report or presentation designed specifically to address questions or needs identified by a particular stakeholder. The purpose of these deliverables is to inform current and future decision making for a particular policy, program, or project. The audience is intentionally narrow and any expectations for the sharing of findings are left to the discretion of the stakeholder(s).

Case Study: TheEngineering Majors Survey

In 2015, the research arm of the National Science Foundation-funded National Center for Engineering Pathways to Innovation (Epicenter) launched a major longitudinal study of engineering students' interests and career goals surrounding innovation and entrepreneurship. The Engineering Majors Survey (EMS) is an online instrument consisting of 35 questions and designed to measure a comprehensive range of undergraduate learning experiences that may influence students' beliefs about their ability to innovate and includes measures of students' entrepreneurial activities, past, present, and future. A nationally representative sample of U.S. engineering schools was identified and invitations to participate were extended to the deans. A cohort of 27 institutions with a collective student population of over 30,000 engineering juniors and seniors agreed to take part.

In preparation for deployment, the EMS research team collaborated with liaisons to develop campus-specific recruitment plans and data collection timetables, as well as promotion materials, relevant resources, and e-mail text for survey recruitment and administration. Approximately 6 months after the final data collection, each participating campus received a report with preliminary findings and analyses on the responses of their students. The external evaluators contracted by Epicenter to focus on the efficiency and effectiveness of the portfolio of programs and teams within the organization closely observed the design and deployment of the EMS.

EMS as Research

As part of a broader research program focused on innovation and entrepreneurship, the following three research questions guided the design of the EMS study:

1. What are undergraduate engineering students' innovation and entrepreneurial interests, abilities, and achievements?
2. How do these interests, abilities, and achievements change over time?
3. Which educational and workplace environments/experiences influence the development of their innovation and entrepreneurial interests, abilities, and achievements?

The research team was led by a faculty member with expertise in engineering and engineering education and included undergraduate students, graduate students, postdoctoral scholars, and several senior research staff members. Quantitative methods were employed to design a survey for data collection. The research findings were published in conference posters, journal articles, and other presentations as contributions to a growing body of scholarship on engineering students' innovation pathways. Because of the longitudinal research design and the need to identify and track individual subjects over time, institutional review board approval was secured at the lead institution and at several of the participating institutions that required additional human subjects review.

Evaluation of the EMS

In collaboration with the EMS research team, Epicenter's external evaluators codeveloped and administered a survey for the deans and campus liaisons at the 27 institutions that participated in the EMS. The purpose of this study was 2-fold: first, to understand the value of the EMS to the participating schools; and second, to solicit feedback on the recruitment, administration, and reporting processes of the EMS in order to learn from the users' experiences. The survey consisted of a combination of items using Likert scales and open-ended questions covering topics such as the adequacy of the preadministration communication with the EMS team, the quality and usefulness of the campus report as well as interest in future involvement with the EMS research.

Descriptive statistics summarizing the responses were shared with the EMS

Descriptive statistics summarizing the responses were shared with the EMS research team. The insights from these findings prompted the research team to streamline the survey administration process for future deployments, assess their current strategies around community-building among the participating schools, and consider how the EMS experience as a whole—from recruitment to reporting—can encourage new thinking about the value of research to inform decision making and actions around student programming and promotion of these activities internally and externally.

The EMS case study illustrates how the efforts of evaluation and research can be complementary and synergistic. With shared expertise in research methods and analysis approaches, the insights gained from the recommendations made by the external evaluators directly supported and led to improvements to enhance the quality and the impact of the research.

Additional Considerations

A helpful visualization created by John Lavelle uses the analogy of an hourglass to describe how the differentiating characteristics of evaluation and research—purpose and intention, guiding questions, and audience focus—converge around the methodologies and analysis approaches used to address the questions at hand. Divergence occurs again at the bottom of the funnel around the types of recommendations that are made, products and deliverables, and whether reporting is focused on internal stakeholders or a broader research community.

Noted evaluator Michael Scriven recognized that social science research methods are essential in order for evaluation to be conducted. He attributed the differences between evaluation and social science research to the development of program evaluation as a profession in the 1960s and its eventual emergence as a discipline with its own unique methodologies, theoretical frameworks, and paradigms. As the discipline of evaluation continues to mature, its distinguishing features and components as compared to research are likely to become more obvious.

Helen L. Chen

See also American Evaluation Association; Evaluation; Evaluation, History of; Evaluation Consultants; Federally Sponsored Research and Programs; 45 CFR Part 46; Human Subjects Research, Definition of; Institutional Review Boards

Further Readings

Karberg, A., Davis, T., & Cloth, A. (2005, January). Program evaluation … human subjects research. The University of Texas at Austin Updated UT Policies and Position Papers. Retrieved from https://research.utexas.edu/ors/human-subjects/special-topics/updated-ut-policies-and-position-papers/

LaVelle, J. (2010, February 26). John LaVelle on describing evaluation (Weblog post). American Evaluation Association AEA365 blog. Retrieved from http://aea365.org/blog/john-lavelle-on-describing-evaluation/

Quinn, P. M., Aisbey, E., Dean-Coffey, J., Kelley, R., Miranda, R., Parker, S., & Fariss Newman G. What is evaluation? American Evaluation Association Online Resources. Retrieved from http://www.eval.org/d/do/492

Scriven, M. (2004). Michael Scriven on the differences between evaluation and social science research. The Evaluation Exchange, IX(4). Retrieved from http://www.hfrp.org/evaluation/the-evaluation-exchange/issue-archive/reflecting-on-the-past-and-future-of-evaluation/michael-scriven-on-the-differences-between-evaluation-and-social-science-research (Original work published 2003).

Legal Citations

Protection of Human Subjects, 45 CFR 46 (2009).

Anna J. Egalite Anna J. Egalite Egalite, Anna J.

Every Student Succeeds Act

Every student succeeds act

632

634

Every Student Succeeds Act

The Every Student Succeeds Act (ESSA) is a federal education law that was signed by President Barack Obama on December 10, 2015. It replaces the No Child Left Behind Act (NCLB) as the most recent reauthorization of the landmark Elementary and Secondary Education Act of 1965, which was the nation's first federal education law and a key component of President Lyndon B. Johnson's War on Poverty. It has been planned to take ESSA to full effect in the nation's public elementary and secondary schools in the 2017–2018 school year. This entry describes the passage of ESSA and the key components of this federal education law and explains how ESSA differs from its predecessor, NCLB, which was signed into law in 2002 by President George W. Bush. The entry concludes with an overview of the transition to ESSA's full implementation.

The Passage of ESSA

ESSA is a federal law focused on education policy reform that replaces NCLB, which was scheduled for reauthorization as early as 2007. Years of political stalemate in Congress delayed efforts to reauthorize NCLB. As a result, NCLB endured for several years beyond its anticipated expiration date, and as the law's 2014 deadline for 100% of students to be proficient in math and reading approached, an increasing number of schools were classified as "failing" to make adequate yearly progress toward this goal. In response, the Obama administration allowed states to petition for relief from some of the law's requirements. By 2014, 43 states, the District of Columbia, Puerto Rico, and a group of California school districts had received waivers granting flexibility in how they met certain requirements of the law.

Then-Education Secretary Arne Duncan drew on Section 9401 of the ESEA to grant the conditional flexibility waivers, which offered states increased flexibility under the law in exchange for the implementation of a particular set of education policies selected by the administration. These policies included the incorporation of student test scores into teacher evaluation systems and the adoption of “college and career-ready” standards.

In 2015, both houses of the 113th Congress produced proposals to reauthorize the ESEA. In the House of Representatives, bipartisan collaboration between Rep. Bobby Scott (D-Va.) and Rep. John Kline (R-Minn.) led to the passage of the Student Success Act on July 8, 2015. The Every Child Achieves Act of 2015, sponsored by Sen. Patty Murray (D-Wash.) and Sen. Lamar Alexander (R-Tenn.), passed the Senate less than 2 weeks later on July 16, 2015. Despite significant political polarization in Congress at the time, bipartisan compromise during the conference committee ensured that a single proposal emerged from the two houses of Congress for the president to sign into law.

Key Components of ESSA

Key elements of the law are as follows:

Testing

As before, states must test students annually in math and reading in Grades 3 through 8 and once in high school. Results must be publicly reported at the school level and broken out by key subgroups defined by student race/ethnicity, disability, English-language status, and poverty status. At the high school level, states are permitted to administer nationally recognized college entrance exams such as the Scholastic Assessment Test or American College Testing instead of traditional standards-based assessments.

Low-performing schools

At least once every 3 years, states must identify the lowest performing schools, defined as the bottom 5% of all schools or those with graduation rates below 67%. Districts are required to intervene in these low-performing schools using evidenced-based practices chosen in partnership with school staff. States must monitor these district intervention efforts and if improvements haven’t been monitor these district intervention efforts and if improvements haven’t been observed after 4 years, states are free to implement their own intervention plan.

ESSA also requires states to identify schools in which student subgroups are underperforming. Schools are responsible for designing an evidence-based plan to improve student subgroup performance, which is subject to district and state supervision. Finally, ESSA consolidates what was previously a stand-alone grant, the School Improvement Grant Program, into general Title I funding. To accommodate this change, ESSA increases the percentage of Title I funds that states can set aside for school improvement efforts from 4% to 7%.

State accountability plans

States are required to submit an accountability plan to the U.S. Department of Education for approval. These plans describe each state’s accountability goals as well as outlining the details of the accountability system they have designed. Accountability systems must include at least four types of indicators: academic achievement, another academic indicator such as students’ academic progress over time, progress toward English proficiency for English-language learners, and one other valid, reliable indicator of school quality or student success. For the fourth indicator, states have broad discretion in choosing how to measure school quality or student success; this indicator could be, for instance, a measure of school climate, faculty retention, or students’ noncognitive skills.

Goal setting

While NCLB mandated that all states demonstrate 100% proficiency on state assessments by 2014, ESSA defers to the states to choose their own long-and short-term goals with regard to student performance on standardized assessments, English-language proficiency, and graduation rates. In setting these goals, however, states must demonstrate how gaps will be closed in terms of student academic achievement and graduation rates.

Academic standards

ESSA requires states to adopt challenging academic standards but doesn’t specify what those might be. Further, the U.S. secretary of education is explicitly prohibited from incentivizing or coercing states to choose a particular set of standards such as the Common Core State Standards.

Teacher quality

The Teacher and School Leader Innovation Program replaces the Teacher Incentive Fund to offer grants to states to experiment with performance pay schemes or other programs to improve teacher and principal quality. ESSA also includes several provisions that aim to strengthen alternative teacher and school leader preparation programs.

Grant-funded programs

ESSA consolidates more than 20 previously stand-alone grants for arts education, computer science, student counseling, student health and safety, and other forms of “student support and academic enrichment” into a single US$1.6 billion block grant.

Pilot programs

ESSA contains several notable pilot programs, such as a weighted student funding program, which will allow 50 districts to comingle funds from federal, state, and local sources to create a school finance system that relies on student-based budgeting.

How ESSA Differs From NCLB

Although ESSA maintains a focus on standards, testing, and accountability, it reverses NCLB’s trend toward centralization by devolving authority over key education policy decisions back to the states. ESSA explicitly limits the power of the U.S. secretary of education, prohibiting the secretary from promoting a particular set of academic standards, for instance. This includes the Common Core State Standards, which the Obama administration publicly supported prior to ESSA’s passage. ESSA also grants states the freedom to determine their own short-and long-term goals for school and district performance rather than setting a target for all schools to achieve.

States are still required to submit an accountability plan to the U.S. Department of Education that describes how they intend to monitor and improve student achievement, English-language proficiency, and graduation rates, but states now have the freedom to design interventions for the lowest performing schools rather than choosing from a narrow menu of federally approved turnaround rather than choosing from a narrow menu of federally approved turnaround options. Also notable is that although previously permitted under NCLB waivers, states can no longer generate “super subgroups” for accountability purposes and teacher evaluation systems are no longer required to incorporate student test scores. Finally, ESSA eliminates NCLB’s “highly qualified teacher” provision, which required states to measure and publicly report progress toward ensuring that all teachers (a) hold a bachelor’s degree, (b) possess full state certification or licensure, and (c) prove that they know each subject they teach.

Transitioning to ESSA

The U.S. Department of Education published guidelines to govern states during the interim period before ESSA was to take effect in 2017–2018. Effective immediately, states were no longer required to provide and notify the public of supplemental educational services and public school choice for students in low-performing public schools. Nonetheless, schools in waiver states that have been identified as in need of targeted support and intervention (“priority” and “focus” schools) had to continue to receive such supports.

In terms of accountability requirements in the interim period, states were no longer required to establish and report progress against annual measurable objectives or to hire “highly qualified teachers.” Finally, during 2016–2017, each state had to distribute any funds received under a state formula grant program to schools in the same manner and using the same formulas that were employed in 2015–2016.

Anna J. Egalite

See also Adequate Yearly Progress; Common Core State Standards; Federally Sponsored Research and Programs; Great Society Programs; No Child Left Behind Act; Policy Evaluation; Standardized Tests

Further Readings

ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015–2016).

Hess, F. M. (2016, February 16). The Every Student Succeeds Act and what lies ahead. Washington DC: American Enterprise Institute. Retrieved from:

https://www.aei.org/publication/the-every-student-succeeds-act-what-lies-ahead/

Manna, P. (2010). Collision course: Federal education policy meets state and local realities. Washington, DC: CQ Press.

U.S. Department of Education. (2016). Transitioning to the Every Student Succeeds Act (ESSA): Frequently asked questions. Washington DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/policy/elsec/leg/essa/essafaqstransition62916.pdf

Karen Badger Karen Badger Badger, Karen

Evidence-Based Interventions Evidence-based interventions

634

636

Evidence-Based Interventions

The No Child Left Behind Act of 2001 articulates the value of research and scientific evidence in education. Evidence-based education is a model in which educational practices are expected to be evaluated and results disseminated, so that empirical data informs decision making and educational intervention selection. Interventions for which there is reliable evidence confirming their effectiveness to bring about behavior change/desired results are referred to as evidence-based interventions (EBIs). Such interventions have been evaluated and, based on strong resulting evidence, determined to be capable of achieving specific outcomes. This entry first provides a brief history of the development of EBIs, then describes the planning process, the evaluation of evidence, locating EBIs, and the role of treatment fidelity in implementation.

Historically, the concept of EBI can be traced to *evidence-based medicine*, a term first introduced in 1992 and later referred to more broadly as *evidence-based practice* (EBP). The application of EBP has grown beyond medicine to include allied health specialties, clinical and counseling psychology, education, and school psychology.

Initially, the EBP paradigm heavily emphasized research findings, with the randomized controlled trial (RCT) representing the gold standard of rigorous evaluative research. This emphasis on research was designed to move from making opinion-driven clinical decisions to those grounded in data. However, some criticized this as too narrowly focused on research to the exclusion of other factors. In 1996, David Sackett and colleagues expanded the evidence-based medicine model to include the use of clinical expertise and consideration of the characteristics, needs, and preferences of the population or individual receiving the intervention. Later EBP models also accounted for the cultural and environmental context where the intervention would be implemented.

environmental context where the intervention would be implemented.

EBIs are validated for a limited application. To be effective, EBIs need to be used to address the specific problem/desired outcomes for which they were designed and with similar populations. Changes to the intervention could impact results. It is also important to remember that EBIs are validated on large groups—they may not be effective with every single case.

EBI planning is a process generally consisting of (a) formulating a researchable question that identifies the type of client/problem, the intervention(s) of interest, and the desired outcome or behavior change; (b) searching for the best quality evidence in response to the question; (c) critically examining the evidence for its applicability, quality, and strength; (d) choosing and implementing the best intervention by integrating the best available evidence with clinical expertise and client/population-specific information; (e) evaluating its effectiveness; and (f) sharing the results and revising the practice as indicated.

The body of education intervention research is growing, but designing and conducting rigorous evaluation research takes time and resources. Thus, not all interventions in use have been tested. Outcome studies may differ in the rigor of their research designs and the strength and reliability of their results. One cannot assume that an intervention that has been published or described as research based meets the standard of an EBI.

Because EBIs in education, as in other areas, draw upon the best available evidence—whether at the program, policy, or individual level—assessing the quality of evidential support is necessary when selecting interventions. There are numerous schemas depicting levels or grades of evidence that consider factors such as the design of the study, its reliability, and the steps taken to minimize bias. Systematic reviews and meta-analyses of RCTs, as well as individual studies using an RCT design, are considered to constitute the highest level of evidence.

RCTs include larger samples, randomization of the sample, and a control. RCTs produce more convincing results than non-RCTs, but the results generated using either of these designs are considered stronger than studies using observational designs, particularly ones with no control group. Data sources such as professional/expert opinion, observational studies with no control, and single case studies are considered to produce evidence of lower quality. Evidence appraisal includes assessing the scientific rigor of the study, the sufficiency of detail to determine effectiveness and guide replication, and peer review of the detail to determine effectiveness and guide replication, and peer review of the findings.

Using established and trusted sources to locate already vetted EBIs can be helpful in identifying effective and research-informed practices. As examples, the Cochrane Collaboration and the Campbell Collaboration provide access to systematic reviews that collectively examine studies that tested particular interventions. Specific to education, the What Works Clearinghouse established by the Institute of Education Sciences, U.S. Department of Education, is a credible, federally endorsed and holistic source of information about reviewed and graded EBIs. The National Center for Education Evaluation and Regional Assistance carries out its own studies, disseminates best practices/research findings, and educates readers about topics such as how to evaluate the strength of available evidence. The University of Missouri houses a comprehensive resource—the Evidence Based Intervention Network—that provides information about the EBI process, resources, and reviewed EBIs.

The efficacy of EBIs is also tied to treatment fidelity—the extent to which the delivery of the intervention aligns with how it was envisioned and implemented during outcome evaluations. To protect treatment fidelity, staff must be trained to carry out the intervention and must follow intervention manuals of procedures as they were designed. There also needs to be ongoing oversight and monitoring of intervention delivery.

Karen Badger

See also American Educational Research Association; Data-Driven Decision Making; Joint Committee on Standards for Educational Evaluation; Merit; Meta-Analysis; National Council on Measurement in Education; No Child Left Behind Act; Outcomes; Program Evaluation; Single-Case Research; Treatment Integrity

Further Readings

Freeman, J., & Sugal, G. (2013). Identifying evidence-based special education interventions from single-subject research. Teaching Exceptional Children, 45(5), 6–12.

Institute of Education Sciences, U.S. Department of Education. What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/default.aspx

Satterfield, J. M., Spring, B., Brownson, R. C., Mullen, E. J., Newhouse, R. P., Walker, B. B., & Whitlock, E. P. (2009). Toward a transdisciplinary model of evidence-based practice. The Milbank Quarterly, 87(2), 368–390.

Stanovich, P. J., & Stanovich, K. E. (2003). Using research and reason in education: How teachers can use scientifically based research to make curricular … instructional decisions. Washington, DC: National Institute of Child Health and Human Development; Department of Education; Department of Health and Human Services. Retrieved from http://lincs.ed.gov/publications/pdf/Stanovich_Color.pdf

University of Missouri. (2011). Evidence-based intervention network. Retrieved from http://ebi.missouri.edu/

U.S. Department of Education Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. Retrieved from http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

Yoon Jeon Kim Yoon Jeon Kim Kim, Yoon Jeon

Evidence-Centered Design Evidence-centered design

Evidence-Centered Design

Evidence-centered design (ECD) originated from the need for a *principled* process of assessment design to meet ever-increasing demands for employing various forms of authentic and interactive tasks and for utilizing advances in cognitive science and computer technology to support those types of tasks. An evidence-centered assessment design framework was created by Robert Mislevy, Linda Steinberg, and Russell Almond at the Educational Testing Service during the 1990s. This entry further defines ECD and discusses how it is used.

With ECD in place, assessment designers can shift to an evidence-based approach to assessment, as opposed to a task-centered approach, where they engineer features of tasks to support the chain of inferences explicit from construct to task. The central principle of ECD is that educational assessment is inherently an evidentiary argument and that ECD guides the design and implementation of assessment as a principled process by formalizing the structure of the assessment argument.

The ECD design process can be described using the *layer* metaphor—where levels of interconnected work are characterized by cycles of iteration and refinement both within and across layers to ensure coherence among assessment models. A comprehensive application of ECD will include five layers: domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. The conceptual assessment framework layer yields operational models that in turn can be used as design objects.

The competency model consists of student-related variables (e.g., knowledge, skills, and other attributes) on which the designer wants to make claims. For example, suppose that the designer wanted to make claims about a student’s ability to design excellent electronic presentation slides. The competency model variables (or nodes) would include technical as well as visual design skills. The evidence model would show how, and to what degree, specific observations and artifacts could be used as evidence to inform inferences about the levels or states of competency model variables. For instance, if the designer observed that a learner demonstrated a high level of technical skill but a low level of visual design skill, it could be estimated that the learner's overall ability to design excellent slides would be approximately "medium"—if both the technical and aesthetic skills were weighted equally. The task model in the ECD framework specifies the activities or conditions under which data are collected. In the slide presentation example, the task model would define the actions and products (and their associated indicators) that the student would generate comprising evidence for the various competencies.

Due to its comprehensiveness and flexibility, ECD has been adopted widely by researchers and practitioners who wish to develop alternative forms of assessment. First, with alternative assessments, what is being assessed is often complex and not immediately apparent. ECD's strength resides in the development of performance-based assessments where assessment designers can begin by figuring out just what they want to assess (i.e., the claims they want to make about learners), thereby clarifying the intended goals and outcomes of learning. Second, the ECD framework can support a sociocognitive view of learning supports where people learn in action. Learning in such environments involves continuous interactions between the learner and the environment (in which tasks are embedded), as learning is inherently situated in context. The interpretation of knowledge and skills as the products of learning cannot be isolated from the context and neither should assessment. When using the ECD framework, assessment is clearly tied to learners' actions within learning environments and operates without interrupting what learners are doing or thinking. One application of this approach in gaming environments is commonly known as game-based assessment.

ECD is not a lockstep tool kit (or recipe book) but rather an iterative conceptual framework. The strength of ECD resides in its flexibly guiding an abstracted way of thinking about assessment and providing an integrated and comprehensive language among various participants of assessment design. The assumption that it is a tool kit can lead to misalignment of expectations and the rejection of ECD as too complicated or theoretical. In addition, just like any other design and development framework, ECD is not a solution to all assessment-related issues. Instead, it requires both formative evaluation and assessment-related issues. Instead, it requires both formative evaluation and revision throughout the process of design and development and summative evaluation of psychometric qualities of the developed assessment.

Yoon Jeon Kim

See also Formative Evaluation; Item Development; Summative Evaluation; Tests; Validity

Further Readings

DiCerbo, K., Shute, V. J., & Kim, Y. J. (in press). The future of assessment in technology rich environments: Psychometric considerations. In J. M. Spector, B. Lockee, & M. Childress (Eds.), Learning, design, and technology: An international compendium of theory, research, practice, and policy. New York, NY: Springer.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), Handbook of test development (pp. 61–90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3–67.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. The Journal of Technology, Learning and Assessment, 8(4).

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. Computer Games and Instruction, 55(2), 503–524.

Phillip D. Payne Phillip D. Payne Payne, Phillip D.

Excel

Excel

637

642

Excel

Microsoft Excel is a widely used spreadsheet application that stores, organizes, and analyzes data across a number of disciplines including education and is available for Windows, Mac OS X, Android, and iOS. Excel relates to education research, measurement, and evaluation because it functions as an efficient and effective data analysis application. This entry describes the basic functions and data analysis tools included in Excel and provides an overview of its statistical capabilities. This entry concludes with a list of the most commonly used formulas, examples of possible outputs, and purchase information.

Basic Functions

The primary purpose of Excel is to organize and analyze large amounts of information. Data are stored in a worksheet (grid) organized in columns and rows that can be manipulated, sorted, and analyzed to meet specific needs of the user. One file can contain multiple worksheets that can be interconnected for efficiency of use and delivered through various forms such as line graphs, charts, and histograms. The most common file extensions are .xlsx, .xlsm, .xlsb, and .xls.

Built into the application are an assortment of procedures to address statistical, scientific, engineering, and financial needs including, but not limited to, pivot tables (which Microsoft refers to as PivotTables), the “what-if” analysis suite, and descriptive statistics. Excel also includes an option for the user to function as a programmer with Visual Basic for Applications as well as the ability to display the spreadsheet as a decision support system and function more as an application would on a computer or smartphone. More frequent and experienced users will be able to use these functions to interactively link with Microsoft Word to generate regular reports and Microsoft PowerPoint to develop slide shows and to send these files out to a subscription list at predetermined intervals.

PivotTables

PivotTables are a function that allows the user to simplify the organization and summarization of a large amount of data. They also provide unique views of the data set as constructed by the user. Among the many possibilities are sorting, counting, totaling, and averaging a large amount of data in one table. All results are constructed in a new table to aid in creating a report for a specific situation as designed by the user. They derive their name from the rotating, or “pivoting,” of data fields graphically to create a new structure providing a unique view of the raw data set.

PivotTables are not created automatically, so the users must design each one specifically to meet their analysis needs. The fields must first be defined, then the raw data in the main table must be assigned to one of the given fields. These fields are typically provided for selection as row or column headers and defined as a report filter, column label, row label, or summation value. Once the fields are defined and the pivot table designed, the user can customize reports based on the needs of the requesting agency or entity. Pivot tables can be a very powerful tool; however, they are intricate in their design, so ensuring there are no errors in the raw data and construction is critical. Error codes are provided if a faulty design exists and the “Help” function will allow for swift analysis of the design flaw.

What-If Suite

What-if analysis is the action of providing multiple values within a range of cells to determine the impact these varying values have on the outcome of the formulas within the workbook. Three types of what-if analysis tools are provided in Excel: scenarios, data tables, and goal seek. The two former types are forward-looking and try to predict outcomes; whereas, the latter is backward-looking in trying to determine one of the initial values that produces a specific result.

A scenario is a set of values defined by the user to create multiple iterations of a

A scenario is a set of values defined by the user to create multiple iterations of a given data set. Scenario Manager is an extension of the scenario tool and stores all of these options and applies them to the data set to determine how the results will be impacted as the scenarios change. Scenarios can accommodate up to 32 different values and the user can create unlimited scenarios. The user also has the option to shift between scenarios during analysis. Data can be collected and merged from multiple workbooks into these scenarios. Once all scenarios are entered, the user can create a summary report of all of the scenarios incorporating all of the information contained in Scenario Manager.

Data tables are another option provided in the “what-if” suite and consist of a range of cells that provide information on how altering one or more of the cells impact the outcomes of the formulas within the spreadsheet. Unlike the scenario option, data tables can only accommodate one table on one worksheet and calculate a maximum of two variables. Any analysis containing more than two variables should use a scenario.

The final type of “what-if” analysis is Goal Seek, which is designed to determine initial values that produce a specific result. Goal Seek is a backward-looking analysis tool and essentially solves for *x*. This allows the user to determine a specific value based on an already known answer. This function is useful when trying to isolate values of given variables.

Visual Basic for Applications

Visual Basic for Applications is another function of Excel that is included in the Office Suite. The implementation of this language allows the user to build macros and control the application in a variety of ways. Among the options for implementing this code are manipulating the interface of the application, altering the toolbars, and creating custom forms or dialog boxes. This option can be useful for data entry of outside users or a way to collect data from a large amount of people. Although this is a useful function, mastery of the language and coding of VBA is essential before attempting the implementation of this function.

Statistical Use

Excel comes standard with preset functions for engineering, finance, and descriptive statistics. It also contains an add-in for statistical analysis including analysis of variance, *F* tests, *t* tests, correlations (Pearson *r* only), Fourier analysis, and multiple regression.

Three types of analysis of variance are provided in Excel: single factor, two factor with replication, and two factor without replication. All three are designed to assume normal distribution and the output is generated in a new worksheet in table format. Furthermore, post hoc analyses are not provided within the add-in; therefore, another program or additional software should be used if detailed analysis is sought for a specific study. *F* tests are also available for two-sample studies to compare the variances of the two populations to determine whether they are equal.

t Tests are available as an analysis tool and come in three forms: paired sample, two-sample assuming equal variances, and two-sample assuming unequal variances. The user can also select “one-tail” or “two-tail” *t* statistic when using the manual function. If using the Analysis ToolPak, the output provided consists of the descriptive statistics, *t* value, and significance in a new worksheet, whereas, when entered manually, only the significance value is reported in the selected cell.

The correlation function employs Pearson *r* and requires two sets of scale data. Other correlations such as Spearman rank-order correlation coefficient or Kendall’s tau can be calculated; however, the user must build the formula into each cell. For these specific correlations as well as others that will help establish reliability of testing measurements, a more powerful statistical package or additional software to run in concert with Excel is recommended.

Regression is calculated in Excel using the “least squares” method of fitting of a line through the set of provided data. This method allows the user to determine how a single dependent variable is affected by one or more independent variables. The function “LINEST” is used to calculate regression in any given worksheet.

Limitations

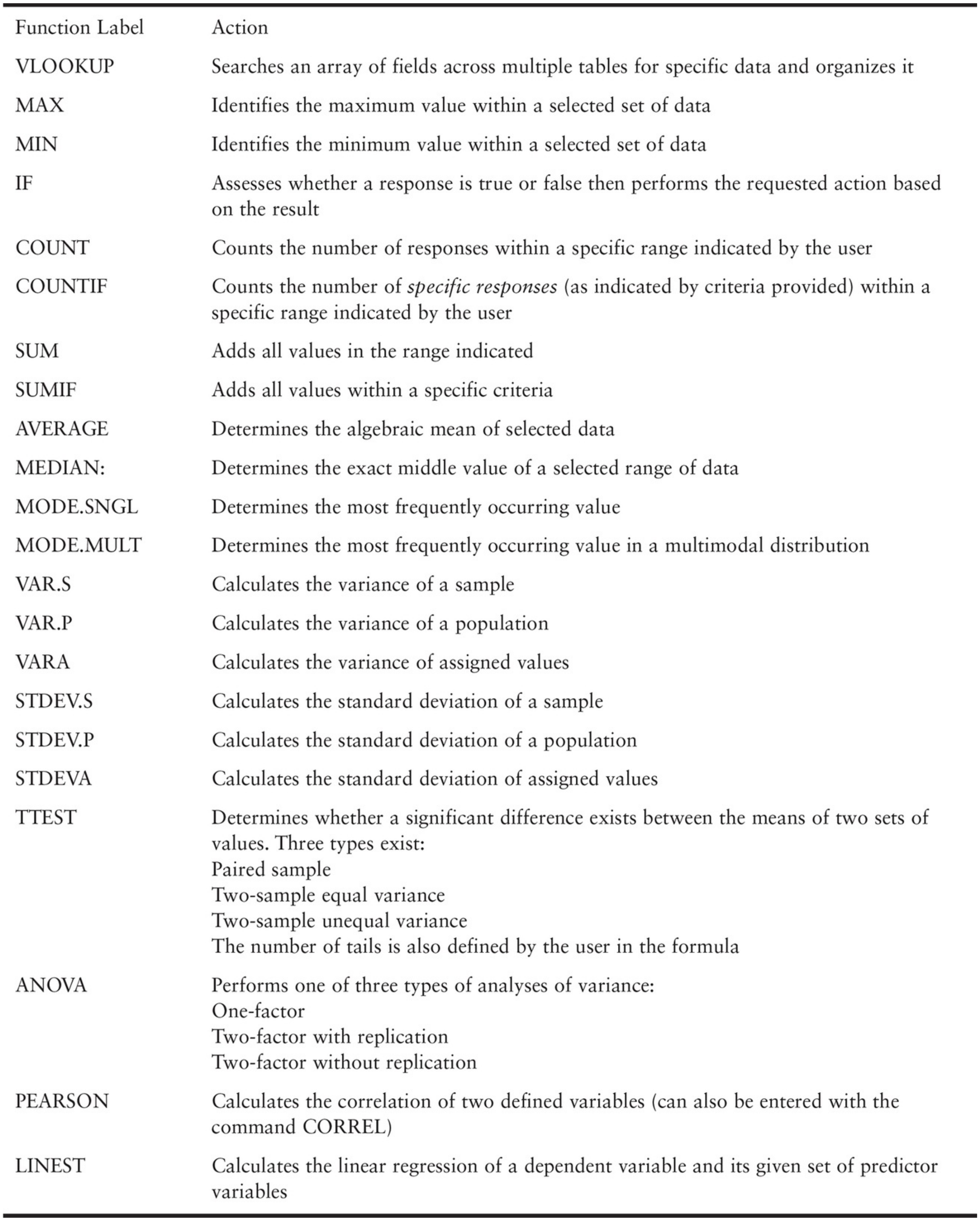
Although many statistical procedures can be applied using Excel, limitations of both the 2016 and older versions of the software create some issues when using the application as a sole source for statistical analysis. As of 2016, both Mac and

Windows platforms have analysis ToolPak add-ins. With the exception of not providing post hoc analysis in analyses of variance, many of the issues regarding the treatment of data have been addressed. However, users of any Mac version prior to 2010 must install a third-party software to gain full use of the statistical package. Furthermore, any version (Windows or Mac) 2007 or prior are subject to flaws in calculating many of the statistical procedures because of the way the software treated empty cells or rounded specific results. Users should be aware that statistical functions can only be used in one worksheet at a time regardless of whether multiple worksheets have been previously linked. The function must be recalculated for each worksheet in the given file.

Commonly Used Formulas

Excel provides an array of formulas for across multiple disciplines. Using the Analysis ToolPak, the user will access the various formulas through the “Data” menu in the top ribbon of the interface. The user will then be guided through a set of dialogue boxes to specify the range of data, the desired α level, and the location of the output table (on the same worksheet or on a new worksheet).

However, the user also has the option to build the formulas directly into the current worksheet. If the user selects this option, the user will need to first type an equals sign (=) into the desired cell followed by a formula listed in Table 1 and an open parenthesis. Excel will then prompt the user with the data required. These are then selected from the spreadsheet with multiple data separated by commas. Once the formula is complete, hitting “enter” will calculate the formula in the current cell. Furthermore, the output is just the final calculation with no corresponding data or explanation. For instance, when calculating a *t* test in this format, the *p* value is provided in the selected cell; however, none of the descriptive statistics or *t* value are provided for the user. Table 1 provides a list of commonly employed formulas and their intended use within the application.



Function Label	Action
VLOOKUP	Searches an array of fields across multiple tables for specific data and organizes it
MAX	Identifies the maximum value within a selected set of data
MIN	Identifies the minimum value within a selected set of data
IF	Assesses whether a response is true or false then performs the requested action based on the result
COUNT	Counts the number of responses within a specific range indicated by the user
COUNTIF	Counts the number of *specific responses* (as indicated by criteria provided) within a specific range indicated by the user
SUM	Adds all values in the range indicated
SUMIF	Adds all values within a specific criteria
AVERAGE	Determines the algebraic mean of selected data
MEDIAN:	Determines the exact middle value of a selected range of data
MODE.SNGL	Determines the most frequently occurring value
MODE.MULT	Determines the most frequently occurring value in a multimodal distribution
VAR.S	Calculates the variance of a sample
VAR.P	Calculates the variance of a population
VARA	Calculates the variance of assigned values
STDEV.S	Calculates the standard deviation of a sample
STDEV.P	Calculates the standard deviation of a population
STDEVA	Calculates the standard deviation of assigned values
TTEST	Determines whether a significant difference exists between the means of two sets of values. Three types exist: Paired sample Two-sample equal variance Two-sample unequal variance The number of tails is also defined by the user in the formula
ANOVA	Performs one of three types of analyses of variance: One-factor Two-factor with replication Two-factor without replication
PEARSON	Calculates the correlation of two defined variables (can also be entered with the command CORREL)
LINEST	Calculates the linear regression of a dependent variable and its given set of predictor variables

Note: ANOVA = analysis of variance.

Output

Output can be generated in many ways and is at the discretion of the user

Output can be generated in many ways and is at the discretion of the user. Charts, tables (not to be confused with data tables), histograms, reports, and presentations can all be created within Excel.

While they can be created specifically in Excel, the user can also generate reports in Word and presentations in PowerPoint linking directly to data sets and analyses in an Excel spreadsheet. Table 2, Figure 1, and Figure 2 are, respectively, examples of a table, chart, and histogram created in Excel.

Figure 1 Chart comparing pretest and posttest scores of 20 participants

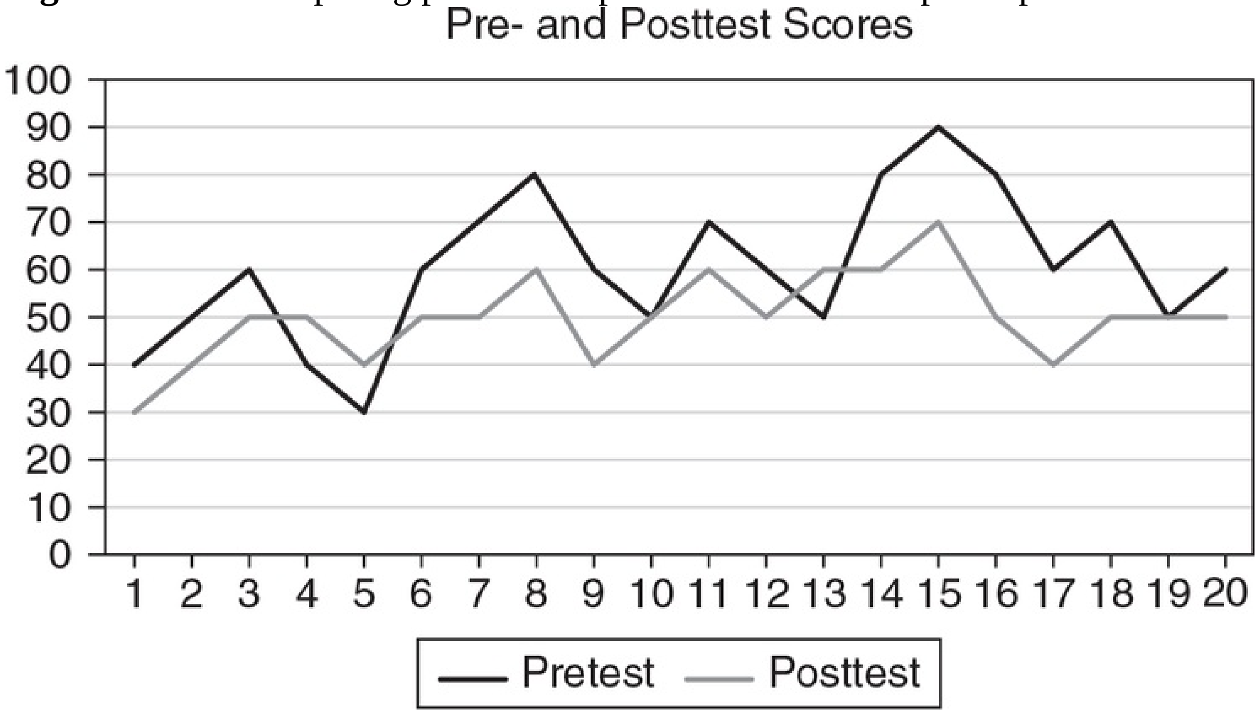
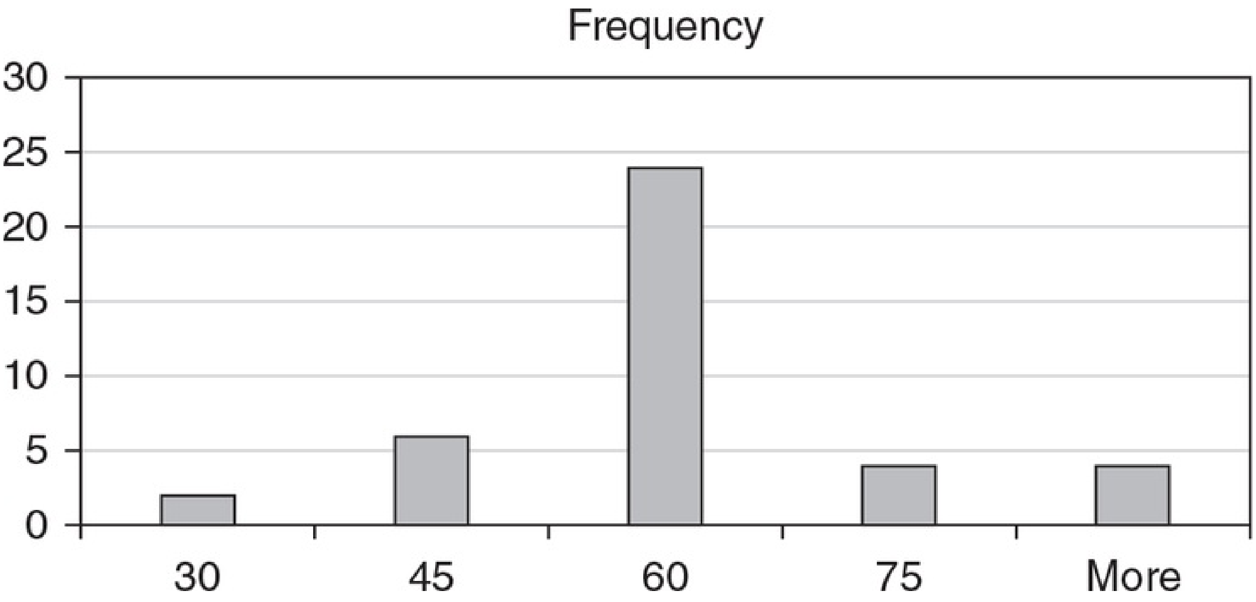
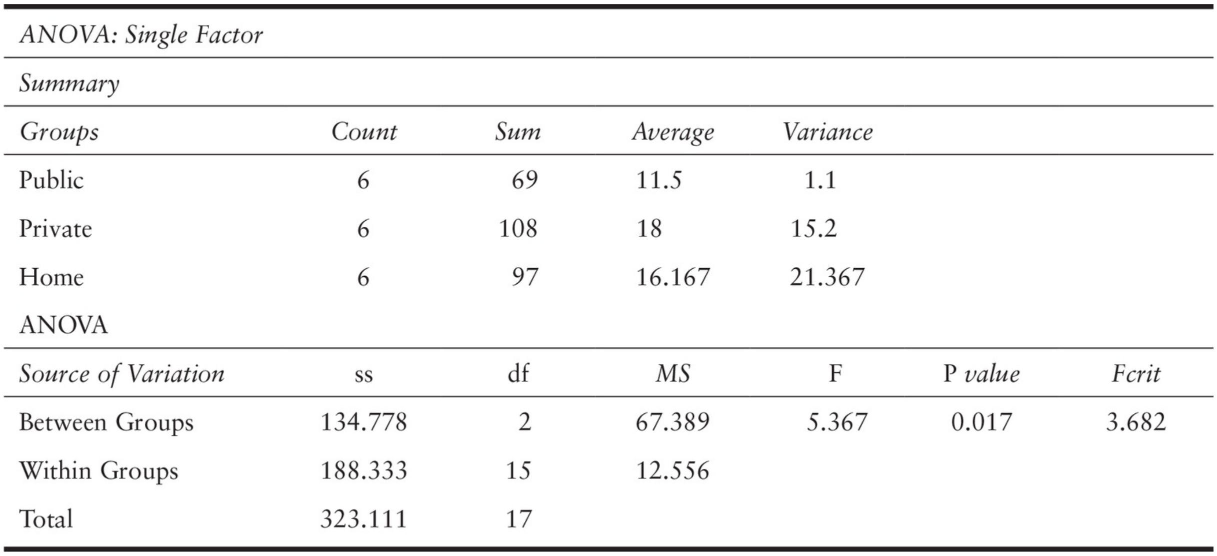


Figure 2 An example of the histogram output in Excel





ANOVA: Single Factor

Summary

Groups	*Count*	*Sum*	*Average*	*Variance*
Public	6	69	11.5	1.1
Private	6	108	18	15.2
Home	6	97	16.167	21.367

ANOVA

Source of Variation	ss	df	*MS*	F	P *value*	*Fcrit*
Between Groups	134.778	2	67.389	5.367	0.017	3.682
Within Groups	188.333	15	12.556			
Total	323.111	17				

Note: ANOVA = analysis of variance.

Purchase and Installation

Excel is part of Microsoft Office 365 and can be downloaded at the Microsoft website or found at a physical or online retailer selling computers or office supplies. Prices will vary based on the selection of the user, and the software can be purchased as a one-time cost or a recurring cost with automatic updates. Installation is available through download or disk and will depend on the mode of purchase.

Phillip D. Payne

See also Analysis of Variance; Correlation; Descriptive Statistics; Multiple Linear Regression; SAS; SPSS; *t* Tests

Further Readings

Abbott, M. L. (2011). Understanding educational statistics using Microsoft Excel and SPSS. Somerset, NJ: Wiley.

Alexander, M., & Kusleika, R. (2016). Excel 2016 formulas. Indianapolis, IN: Wiley.

Salkind, N. J. (2017). Statistics for people who (think they) hate statistics: Using Microsoft Excel 2016. Thousand Oaks, CA: Sage.

Walkenbach, J. (2015). Microsoft Excel 2016 bible: The comprehensive tutorial resource. Indianapolis, IN: Wiley.

Yalta, A. T. (2008). The accuracy of statistical distributions in Microsoft® Excel 2007. Computational Statistics … Data Analysis, 52(10), 4579–4586. Retrieved from http://dx.doi.org/10.1016/j.csda.2008.03.005

Websites

ExcelFunctions.net www.excelfunctions.net

Microsoft Office Help … Training www.support.office.com

Real Statistics Using Excel www.real-statistics.com

Laura O’Dwyer Laura O’Dwyer O’Dwyer, Laura

Experimental Designs

Experimental designs

642

646

Experimental Designs

Experimental designs are used to examine the effect of a treatment or intervention on some outcome. In the simplest two-group case, a treatment is implemented with one group of participants (the treatment group) and not with another (the control group). Experiments can be conducted with individual participants or with clusters of individuals. That is, the unit of assignment may be at the individual level or at the cluster level. This entry refers to individual participants as the unit of assignment with the understanding that the same designs may be used with clusters of individuals. The entry further describes experimental designs, looks at the role of randomization in experimental designs, and discusses some commonly used experimental designs.

Experimental designs require that the researcher assign participants to the treatment or the control group using random assignment, a process known as randomization. Subsequent to applying the intervention to the treatment group and observing participants in both conditions, the researcher hypothesizes about the intervention’s effect on each group. Treatment exposure is the independent variable that is hypothesized to lead to changes in the outcome or dependent variable.

When correctly implemented, experimental designs provide unbiased estimates of the effect of a treatment on observed outcomes. The primary purpose of experimental designs is to establish “cause and effect” or more technically, to make causal inferences. The researcher aims to conclude that the treatment *caused* the differences that were observed between the groups on the attribute that is being studied.

The Role of Randomization in Experimental Designs

Establishing cause and effect requires that several conditions be met. For X to cause Y, X must occur before Y; changes in X must be associated with changes in Y; and all other plausible explanations for the observed association between X and Y must be controlled. The condition that all other plausible explanations are controlled is one of the defining characteristics of experimental research. When researchers conduct an experiment, they apply these three conditions by (1) manipulating the hypothesized cause and observing the outcome afterward, (2) testing whether variation in the hypothesized cause is associated with variation in the outcome, and (3) using randomization to reduce the plausibility of other explanations for the results observed. In technical terms, the final condition stipulates that plausible *threats to internal validity* are controlled. These include subject characteristics threats, testing threats, instrumentation threats, history threats, attrition threats, and regression to the mean.

Arguably, the *subject characteristics* threat is the most important threat minimized by the process of randomization. A subject characteristics threat occurs when individuals in the groups being compared are not equivalent to each other prior to the implementation of the treatment. In this case, equivalence implies that on average, the two groups are approximately the same on all measured *and* unmeasured characteristics. Without group equivalence, the researcher cannot be confident that any observed posttreatment differences were caused by the treatment. It is important to note that randomization does not eliminate all threats to internal validity; instead by reducing subject characteristics threats, randomization aims to ensure that threats are distributed evenly across conditions and are not conflated with participants' condition membership. In experimental designs, randomization is the primary mechanism for minimizing plausible internal validity threats, distinguishing them from quasi-experimental designs, which without randomization cannot fully minimize all plausible threats.

Randomization requires that each participant has a nonzero probability of being assigned to condition, implying that all participants *could* be assigned to either condition. Groups created using a random process are probabilistically equivalent having been equated on the expected values of all attributes prior to the implementation of the intervention, regardless of whether those attributes are measured. When the randomization process is fair, the center of the distribution of all possible sample statistics (e.g., means, standard deviations) will be the same in the treatment and control groups. It is important to note, however, that expectation is about the mean of all possible sample statistics and that the results of a single randomization process may create groups that are, by chance, different from each other on some attributes. In this case, researchers may still conclude that a single experiment provides an unbiased estimate of the treatment effect because the difference between the observed treatment effect and the population treatment effect only occurs by chance.

Commonly Used Experimental Designs

Experimental designs fall into several broad categories. In a *between-subject design*, participants are randomly assigned to serve in only one of the treatment conditions. In a *crossover design*, participants serve in one condition first and then cross over to participate in the other condition. In a *longitudinal design*, researchers collect data at multiple time points before and after the implementation of the treatment. In a *factorial design*, the effects of two or more treatments and their interactions are evaluated simultaneously. In the case where all possible combinations of treatments are evaluated, the design is referred to as full factorial experimental design, and when only some combinations are evaluated, the design is referred to as a fractional factorial experimental design.

Between-Subject Experimental Designs

In the simplest form of a between-subject experimental design, participants are randomly assigned to either the treatment or control condition, pretest data are collected, the treatment is implemented with one group and not with the other, and at the end of the intervention phase, posttest data are collected. A two-group experimental design with randomization to groups (denoted by R), a single treatment (X), and pre-and posttest measures (O) is configured as follows:

```
 Treatment group R  O   X  O

                                      -------------------------

  Control group R  O    O
```

Alternatively, the researcher may randomly assign participants to treatment and control conditions *after* the pretest data are collected. Under this approach, the pretest data may be used to create homogenous strata from within which participants are randomly assigned to either the treatment or control condition. Compared to the simple random assignment process that relies on chance to make the group's equivalent, this alternative approach may result in treatment and control groups that are more similar to each other, particularly if the sample size is small.

Treatment group O R X O

Control group O R O

Between-subject designs can easily be extended to include more than two groups so that the outcomes can be compared across the conditions to determine which treatment (X_A or X_B) or amount of treatment produces the greatest effect:

Treatment group A O R X_A O

Treatment group B O R X_B O

Control group O R O

In between-subject designs, the pretest allows researchers to empirically examine the equivalence of the treatment and control groups on the measured variables; statistically control for preexisting differences on the pretest; and monitor attrition rates in the treatment and control groups. Although there is the possibility of a testing threat, this can be mitigated by using psychometrically equivalent pre-and posttest measures and/or maximizing the time between the data collection points. Another variation on the between-subject designs would be to eliminate the pretest measures; while this may ameliorate the effects of a be to eliminate the pretest measures; while this may ameliorate the effects of a testing threat, it would preclude being able to evaluate the equivalence of the groups and monitor the effects of attrition.

The Solomon four-group design is a variation of the between-subject design that allows researchers to examine testing threats empirically. Using a complex four-group design, this configuration allows researchers to compare groups that do and do not complete a pretest and do and do not receive the treatment. By comparing the posttest scores for Groups A and C, and Groups B and D, researchers can evaluate whether testing threats are likely to have led to the results observed. The configuration for the Solomon four-group design is as follows:

Group A R O X O

Group B R O O

Group C R X O

Group D R O

Crossover Designs

Crossover experimental designs require that participants be randomly assigned to receive one of at least two treatments first and, subsequent to the posttest, receive the second treatment. A typical configuration with two treatment conditions (X_A and X_B) would be as follows:

R O X_A O X_B O

R O X_B O X_A O

The primary advantage of crossover designs is that individuals serve in every condition, making it possible to look at the effects of individual treatments and if there is an interest, in the cumulative effects of participating in both conditions. That being said, crossover designs are generally considered useful when carryover effects are not expected from the first treatment being implemented, and when attrition from the study and testing threats are not be expected to be an issue.

Longitudinal Experimental Designs

When implementing a longitudinal design, researchers collect data from randomly formed groups at multiple time points prior to and after the implementation of a treatment. The following configuration indicates pre-and posttest data collection at four time points before and after the implementation, but as many time points as are feasible may be added:

Treatment group R O O O O X O O O O

Control group R O O O O O O O O

Because multiple measures of an attribute provide a more stable and consistent (i.e., reliable) estimate compared to a single measure, the researcher can have greater confidence in the researcher’s measurement of the attribute in the groups. In addition, this type of configuration allows the researcher to formulate statistical models of change over time as a consequence of the intervention. This approach would be particularly useful if, say, maturation was expected to be a concern. In this case, the researcher could build a measure of maturation into the design and during the data analysis phase could explicitly model maturation while also examining the treatment effect. Finally, the multiple posttest measures allow researchers to examine the immediacy of the treatment effect and whether it endures over time. This virtue makes longitudinal designs ideal for examining interventions that aim to create sustainable change in attributes or behaviors.

Despite these strengths, however, this configuration has several weaknesses, some of which may preclude its implementation, and others that can weaken the validity of any causal claims. Specifically, longitudinal designs are vulnerable to testing threats, particularly if the same measurement instruments are used at each time point and are often weak with regard to attrition. Depending on the duration of the study and the extent of the commitment required on the part of the participants, it can be difficult and costly to maintain a sufficiently large sample that also remains representative of the population. Overall, longitudinal designs are costly and time intensive to implement, and depending on the research area, it can be difficult to recruit subjects to studies that are conducted over long periods of time.

Factorial Designs

When the effects of two or more treatments and their interactions need to be estimated together, researchers may choose a factorial experimental design. In this type of design, two or more treatments are considered factors, each with at least two levels. Although these types of designs can be extended to include many factors with many levels, the following configuration represents a two-factor (A and B) design, each with two levels (1 and 2):

R O X_{A1B1} O

R O X_{A1B2} O

R O X_{A2B1} O

R O X_{A2B2} O

Although this design is often difficult to implement in the field, it offers researchers several advantages when the aim is to examine treatments together.

For instance, these designs allow researchers to investigate the joint effect of the treatments and whether the effect of one treatment is constant across all levels of the other treatment(s). The latter is referred to as an interaction effect. Moreover, all else being equal, factorial designs require fewer participants than conducting two or more between-subject studies to estimate treatment effects independently.

Conclusion

The clear advantage of experimental research designs rests on their capacity for supporting causal inferences. With the combined virtues of random assignment and counterfactual evidence provided a control group, experimental research designs are the gold standard for isolating causal mechanisms. However, experimental designs can be difficult to implement in the real-world environments such as classrooms and schools. For example, it is often difficult to assign (randomly, or otherwise) teachers in the same school to different conditions, to assign students to classrooms, and to assign students in the same classroom to different conditions. As a consequence, experimental designs in education often use schools (i.e., clusters) as the assignment unit whereby schools, along with every teacher and student in that school, are assigned to either the treatment or control condition. This approach typically requires many schools and so can be quite expensive to implement.

Laura O’Dwyer

See also Longitudinal Data Analysis; Random Assignment; Random Selection; Threats to Research Validity; Validity; Validity Generalization

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 228–293). Chicago, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design … analysis issues for field settings. Boston, MA: Houghton Mifflin.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past 15 years. Annual Review of Psychology, 45, 545–580.

Fraenkel, J. R., & Wallen, N. E. (2011). How to design and evaluate research in education (6th ed.). Boston, MA: McGraw-Hill.

Gall, M. D., Borg, W. R., & Gall, J. P. (2003). Educational research: An introduction (7th ed.). White Plains, NY: Longman.

Keppel, G., & Wickens, T. D. (2004). Design and analysis: A researcher's handbook (4th ed.). Prentice Hall.

Kirk, R. E. (2012). Experimental design: Procedures for the behavioral sciences (4th ed.). Thousand Oaks, CA: Sage.

McMillan, J. H., & Schumacher, S. (2006). Research in education: Evidence-based inquiry (6th ed.). Boston, MA: Pearson.

Mertler, C. A., & Charles, C. M. (2010). Introduction to educational research (7th ed.). Boston, MA: Pearson.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Shavelson, R. J., & Towne, L. (2002). Scientific research in education. Washington, DC: National Academies Press.

Wiersma, W., & Jurs, S. (2009). Research design in quantitative research. In Research methods in education: An introduction. Boston, MA: Pearson.

Norman J. Lass Norman J. Lass Lass, Norman J.

Experimental Phonetics Experimental phonetics

646

648

Experimental Phonetics

Experimental phonetics is the branch of general phonetics that applies the experimental method to the study of sounds and other human speech units. This scientific field includes basic areas of phonetics: articulatory phonetics, acoustic phonetics, and auditory phonetics. Moreover, the experimental method is used to investigate numerous topics, including segmental phonetics (the study of the individual sounds, or phonemes, of a language) and suprasegmental phonetics (the study of nonsegmental features of a language, including stress, intonation, and timing, overlaid on segmental features). This entry first provides an overview of phonetics, then discusses topics studied in experimental phonetics research.

Phonetics

Phonetics is the study of speech sounds, including the isolated speech sounds of vowels, diphthongs (combinations of two vowels), and consonants as well as their physiological production and acoustic features. Articulatory phonetics involves the various configurations (shapes) of the human vocal tract determined by the vocal folds of the larynx (“voice box”), pharynx (throat), oral cavity (mouth), nasal cavity (nose), and lips used to produce speech sounds. Acoustic phonetics involves the acoustic properties of speech sounds, whereas linguistic phonetics involves the mechanism for combining speech sounds to produce syllables, words, phrases, and sentences.

Phonemes are the individual sounds of a language. They include consonants, vowels, and diphthongs (combinations of two vowels). Consonants can be classified as *voiced* (example: *b* and *d*) or *voiceless* (example: /*p*/ and /*t*/). All vowels are voiced (i.e., they involve vibration of the vocal folds of the larynx), the anatomical structure in the neck region below the trachea (windpipe) responsible for producing speech and other nonspeech sounds. Examples of English vowels include *i* (as in b***ee***), *e (as in* c***a***se*)*, *ε* (as in n***e***t), *ae* (as in b***a***t), *u* (as in s***oo***n), and //(as in s***aw***). Diphthongs are combinations of two vowels. Examples are *I* (as in n***oi***se), *eI* (as in p***ai***d), and *a* (as in n***ow***).

The traditional approach to describing speech sounds is based on the movements of the anatomical structures that produce them. These structures include the articulators (tongue, lips, teeth, alveolar [gum] ridge, hard palate, and soft palate), as well as the respiratory system (airstream from the lungs), and, for voiced sounds, the vocal folds of the larynx. The airstream from the lungs passes between the vocal folds, which are two small muscular folds located in the larynx at the top of the trachea. If the vocal folds are apart, as they are normally for vegetative (nonspeech) breathing, the air from the lungs will have a relatively free passage into the pharynx and the oral cavity. But if the vocal folds are modified to create a narrow passage between them, the airstream will cause them to be sucked together, with no flow of air, and the pressure below them will build up until they are blown apart. The flow of air between them will then cause them to be sucked together again, and the vibratory cycle will continue. Sounds produced when the vocal folds are vibrating are *voiced sounds* (all vowels, diphthongs, and voiced consonants); when the vocal folds are apart, they are *voiceless sounds* (all voiceless consonants).

Topics inExperimental Phonetics Research

Experimental phonetics research has advanced our understanding of the speech production and speech perception processes through the study of topics such as:

models of speech production;

the role of laryngeal jitter (a measure of frequency instability) and shimmer (a measure of amplitude instability) in speech production;

laboratory techniques for the investigation of speech articulation;

aerodynamics of speech production;

brain mechanisms responsible for speech motor control;

the role of formant frequencies (resonances of the human vocal tract) and fundamental frequency (lowest frequency of a complex periodic sound) in speaker identification;

cognitive processes in speech perception;

evidence-based practice in the treatment of communication disorders;

the function and dysfunction of the temporomandibular joint;

spectrographic analysis of speech;

production and perception of speech rate;

listener perception of time-altered speech;

suprasegmental (prosodic) features of speech;

feedback mechanisms in speech production;

contemporary instrumentation for the study of speech acoustics;

contemporary instrumentation for the study of speech physiology (respiration, phonation, and articulation);

motor control systems in speech production;

speech synthesis techniques;

auditory illusions such as the verbal transformation effect (when listeners begin to report changes in the verbal form of a word after it is repeated multiple times) and their implications for speech perception;

dichotic listening (while listening under headphones, listeners are presented with two different auditory stimuli presented simultaneously, one to the right ear and one to the left ear, and are asked about the content of each message);

infant/developmental speech perception;

computer speech recognition; and

theories of speech perception.

In addition, experimental phonetics research is employed to test theories and/or hypotheses in order to support or disprove them, thereby providing important information on the speech production and speech perception processes.

Experimental phonetics research is conducted by investigators from numerous disciplines, including education, special education, deaf education, speech science, hearing science, physiology, anatomy, otolaryngology, linguistics, neurology, neuroanatomy, neurophysiology, prosthodontics, psychology, neuroscience, speech-language pathology, audiology, and sociology. Among the areas of education studied in experimental phonetics research are the effect of bilingualism on the education of children and the perception of speech rhythm in the second language of bilingual children.

Norman J. Lass

See also Bilingual Education; Experimental Designs; Speech-Language

Pathology

Further Readings

Ferrand, C. T. (2014). Speech science: An integrated approach to theory and clinical practice (3rd ed.). New York, NY: Pearson.

Hardcastle, W. J., Laver, J., & Gibbon, F. E. (2013). The handbook of phonetic sciences (2nd ed.). West Sussex, England: Wiley.

Lass, N. J. (2013). Review of speech and hearing sciences. St. Louis, MO: Elsevier.

Lass, N. J., & Woodford, C. M. (2007). Hearing science fundamentals. St. Louis, MO: Elsevier.

Michael Quinn Patton Michael Quinn Patton Patton, Michael Quinn

Expert Sampling

Expert sampling

648

648

Expert Sampling

The logic and power of expert sampling lie in selecting people to study or interview who are especially knowledgeable about a topic and are willing to share their knowledge. Expert sampling involves identifying key informants who can inform an inquiry through their knowledge, experience, and expertise. Experts can provide valuable insights into the root of problems, what has been tried and failed, what has been tried and worked, and future trends to watch.

Using key informants began with ethnographers who needed indigenous expertise to help them understand cultures other than their own. Sociologists developed focus groups and interview methods with key informants to study issues in their own countries. A carefully selected group of naturally acute observers and well-informed people can serve as a panel of experts about a setting or situation, experts who can take the researcher inside a phenomenon of interest. Expert sampling to interview key informants can be part of research and evaluation on any specialized issue that requires in-depth knowledge of what goes on in a place and how things work. For example, expert sampling could be used with special education teachers to learn about the issues involved in teaching children with special needs.

Expert sampling can be used to gather data from experts through surveys. Interviews with experts are among the most common sampling strategies for qualitative inquiry. The challenge is identifying and gaining the cooperation of genuinely knowledgeable experts, whether through surveys or interviews. As with all sampling, what a researcher or evaluator ends up learning in a study depends on who is sampled. The credibility and utility of expert sampling results depend on the credibility and depth of knowledge of the experts surveyed.

depend on the credibility and depth of knowledge of the experts surveyed, interviewed, and/or observed.

Michael Quinn Patton

See also Convenience Sampling; Focus Groups; Qualitative Research Methods; Survey Methods; Surveys

Further Readings

Emmel, N. (2013). Sampling and choosing cases in qualitative research: A realist approach. London, UK: Sage.

Patton, M. Q. (2015). Purposeful sampling. In Qualitative research and evaluation methods (4th ed., pp. 264–315). Thousand Oaks, CA: Sage.

Zafar, M. B., Bhattacharya, P., Ganguly, N., Gummadi, K. P., & Ghosh, S. (2015). Sampling content from online social networks: Comparing random vs. expert sampling of the Twitter stream. ACM Transactions on the Web (TWEB), 9(3), 12.

Leandre R. Fabrigar Leandre R. Fabrigar Fabrigar, Leandre R.

Matthew P. H. Kan Matthew P. H. Kan Kan, Matthew P. H.

Exploratory Factor Analysis Exploratory factor analysis

648

653

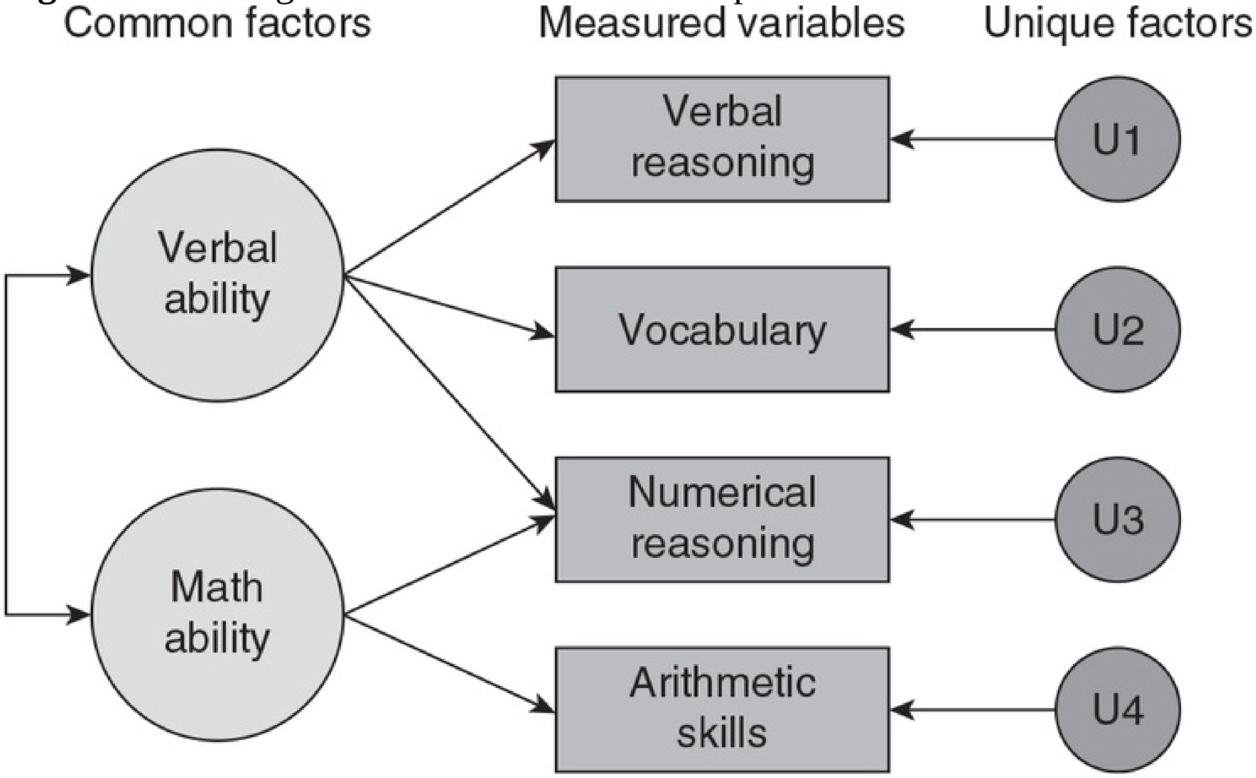
Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a set of statistical procedures used to determine the number and nature of constructs required to account for the pattern of correlations among a set of measures. EFA is used when there is little theoretical and/or empirical basis to generate specific predictions regarding the underlying structure of correlations among the measures. This entry further describes EFA and its purpose and then discusses how to conduct an EFA.

EFA is generally conducted for purposes of theory development (i.e., identifying fundamental constructs in a domain of interest) or measure development (i.e., determining specific measures that effectively represent constructs). For example, a researcher might be presented with several academic tests (e.g., verbal reasoning, vocabulary, numerical reasoning, and arithmetic skills) and unsure of the academic abilities (e.g., verbal and mathematical ability) underlying these tests. Although a researcher might speculate which tests reflect common underlying abilities, EFA provides a more formal method of assessing which measures reflect the same constructs.

To understand EFA, it is useful to have some insight into the *common factor model* (CFM), which is a mathematic framework upon which EFA is based. The CFM can be illustrated via a path diagram using the academic tests example in the previous paragraph (see Figure 1). Common factors, also called latent variables, are hypothetical constructs that cannot be directly measured and influence more than one measured variable. In the academic tests example, the common factors are verbal and mathematical ability. Verbal ability influences verbal reasoning, vocabulary, and numerical reasoning, whereas mathematical ability influences numerical reasoning and arithmetic skills.

Figure 1 Path diagram for academic tests example



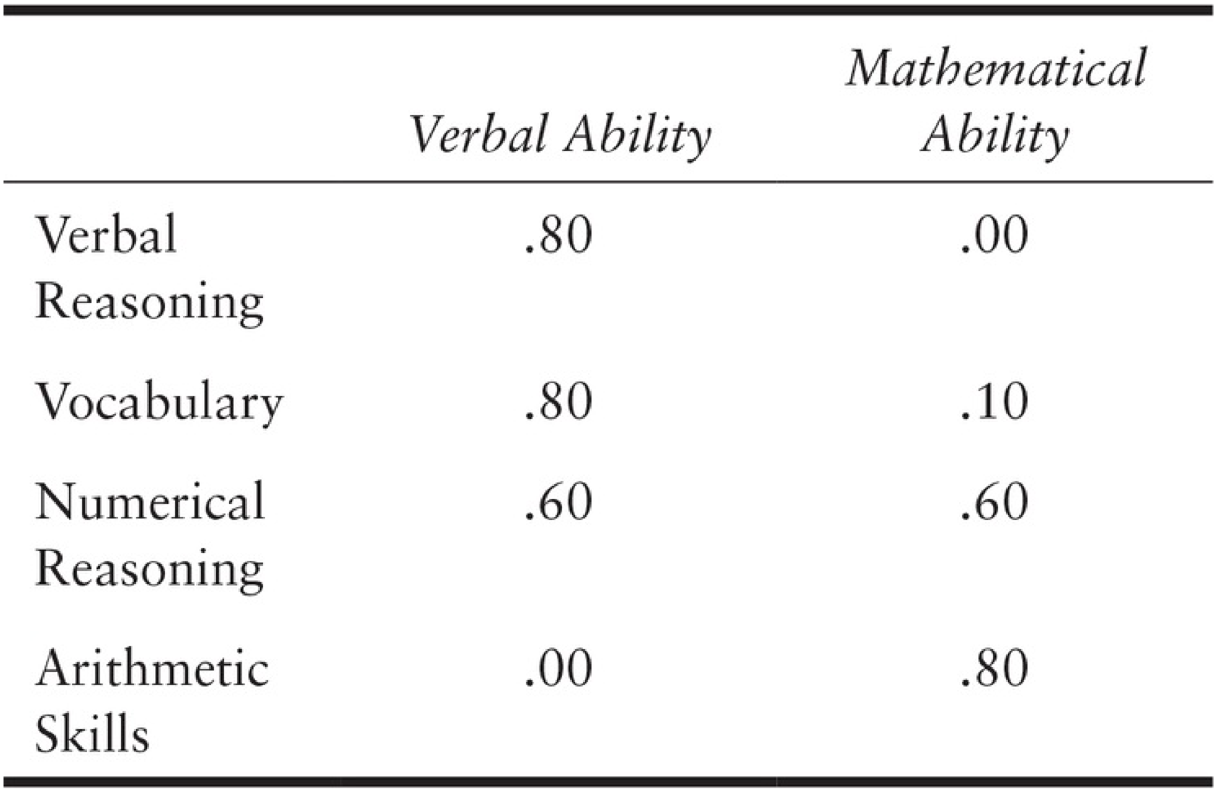
Measured variables, also called manifest or surface variables, are observed scores that can be directly computed and are drawn from the domain under investigation. In the academic tests example, the measured variables are verbal reasoning, vocabulary, numerical reasoning, and arithmetic skills tests. Unique factors are unobservable variables that account for the variance in measured variables that is unaccounted for by the common factors. Each unique factor only influences one measured variable and includes two components: the specific factor and measurement error. The specific factor consists of systematic sources of influence on a measured variable that are specific to that variable, while the measurement error is random influences on a measured variable.

Another way to represent the CFM is through its formal matrix algebra mathematical expression:

$$P = \lambda\Phi\lambda^T + D\psi,$$

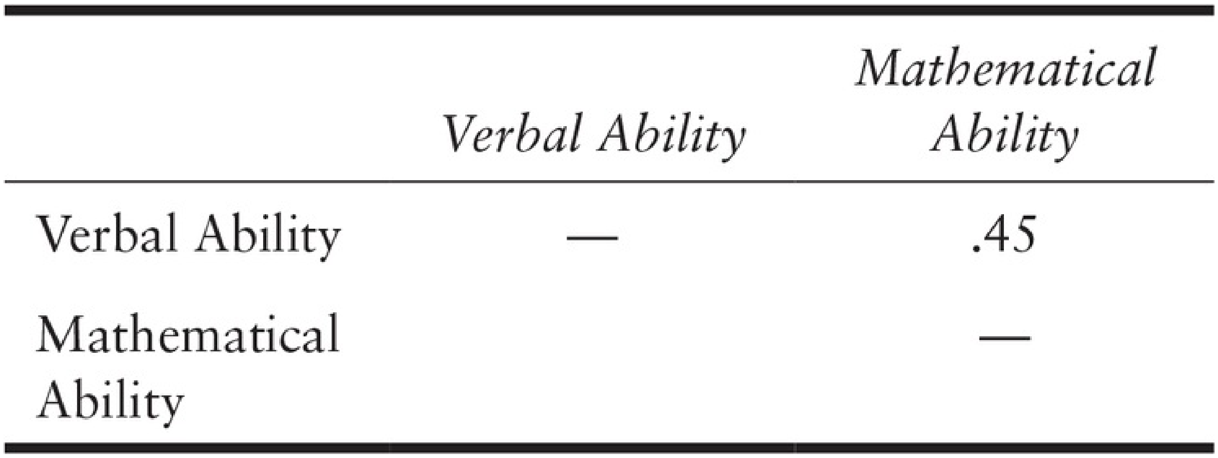
where P is the measured variable correlation matrix in the population and λ is the factor loadings matrix that contains numerical values representing the strength and direction of common factors’ influence on the measured variables. In this matrix, columns represent common factors and rows represent measured variables. The elements comprising the matrix reflect the influence of each common factor on each measured variable.

For example, as seen in Table 1, for every 1 unit of increase in verbal ability, there is a corresponding 0.8 unit increase in vocabulary. Φ is the matrix of correlations among the common factors. λ^T is the factor loadings matrix transposed (a reexpression of the columns of a matrix as rows). $D\psi$ is the unique factors covariance matrix. In this matrix, the diagonal elements are the unique variances associated with each measured variable, and the off-diagonal elements (covariances among unique factors) are assumed to be zero because unique factors are assumed to be independent of one another.

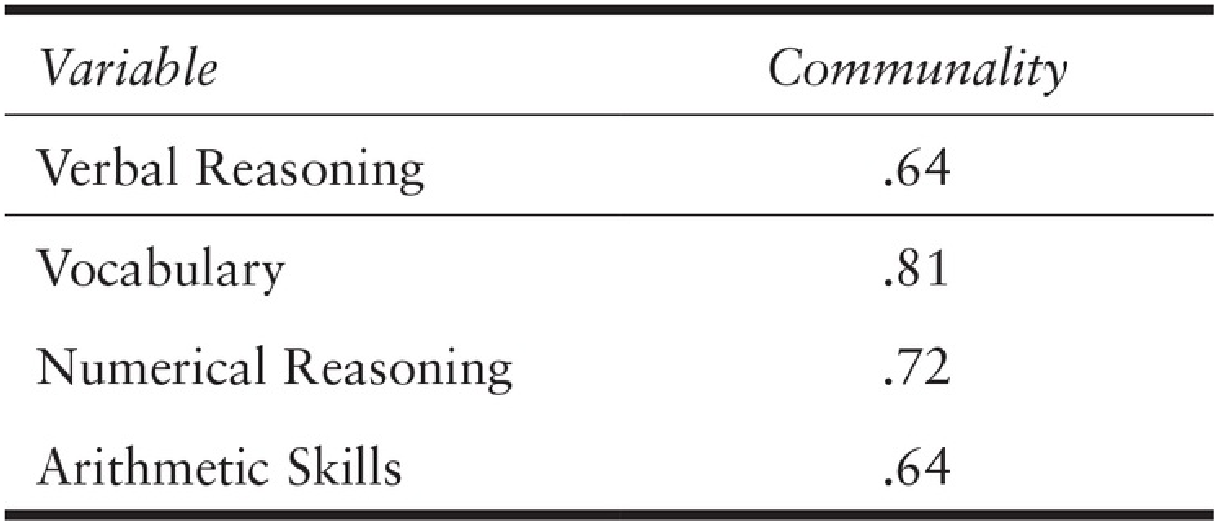


	Verbal Ability	*Mathematical Ability*
Verbal Reasoning	.80	.00
Vocabulary	.80	.10
Numerical Reasoning	.60	.60
Arithmetic Skills	.00	.80

Although the path diagram and matrix algebra expressions are different approaches to representing the CFM, they have the same conceptual implications. Both suggest that when two measured variables are strongly influenced by the same set of common factors, those two measured variables should be strongly correlated with one another. Conversely, if two measured variables are influenced by distinctly different sets of common factors, they should be uncorrelated with one another. Conducting an EFA allows researchers to estimate the values in the CFM equation. Typically, statistical software provides researchers with numerical values for factor loading matrix, common factor correlation matrix (see Table 2), and communalities (see Table 3). Communalities are inversely related to unique variances, as they are the proportion of variance in each measured variable accounted for by all common factors. Thus, higher communalities equal lower unique variances.



	Verbal Ability	*Mathematical Ability*
Verbal Ability	—	.45
Mathematical Ability		—



Variable	*Communality*
Verbal Reasoning	.64
Vocabulary	.81
Numerical Reasoning	.72
Arithmetic Skills	.64

How to Conduct an EFA

Choosing a Method of Model Fitting

The tables shown previously illustrate the final results of an EFA. The challenge of EFA, however, is that there are many steps in reaching this final set of results, and there are many procedures that can be used to accomplish each step. The EFA process begins at model fitting (also called factor extraction or parameter estimation). Model fitting is the process by which the numerical values are calculated for a given model that best account for the correlations among the measured variables. For most model fitting procedures, the CFM mathematical equation is rearranged into the following mathematically equivalent equation:

$$R - D_{\psi} = \lambda\lambda^{T},$$

where *R* refers to the correlation matrix of measured variables in a sample, *R* replaces *P* as the correlation matrix for the population is rarely available and thus a correlation matrix based on a sample drawn from the population is the best available approximation of *P*. Φ is dropped from the equation because in most fitting procedures, factors are initially assumed to be uncorrelated. $R - D_{\psi}$ represents the reduced correlation matrix, which is the unique factor matrix subtracted from the correlation matrix. In this matrix, the diagonals become the communalities and the off-diagonals remain the same because the D_{ψ} off-diagonals are 0.

There are two unknowns in this equation: D_{ψ} and $\lambda\lambda^{T}$. λ and λ^{T} are considered to be one unknown, as λ^{T} is simply λ transposed. D_{ψ} is estimated using communalities, which are initially unknown. For most fitting procedures, *square multiple correlations*, which are each measured variables' amount of variance that is accounted for by all other measured variables, are used as initial estimates of communalities. With square multiple correlations estimating the values of $R - D_{\psi}$ in the diagonal, λ remains the sole unknown and can therefore be solved.

The central task of model fitting methods is thus to arrive at a solution for λ (i.e., the factor loadings) that comes as close as possible to reproducing the values of the reduced correlation matrix. Different fitting procedures are distinguished by how they mathematically define the *closeness* between the elements of the predicted and observed reduced correlation matrices. The mathematical definition of closeness is referred to as the *fitting function*.

One common model fitting method, *noniterated principal axis factor analysis* (NIPAF), defines this closeness as the sum of the squared differences between the elements of the predicted and observed reduced correlation matrices. For this fitting function, a value of 0 would indicate perfect fit and larger values would indicate poorer fit. For NIPAF, the factor loadings are calculated using the estimated (or initial) communalities (i.e., square multiple correlations), and the final communalities are then computed from the factor loadings. Each final communality is calculated by summing the square of the factor loadings for each measured variable. In the academic assessments example, the final communality for the measured variable vocabulary is equal to:

$$(0.80)^2 + (0.10)^2 = 0.65.$$

A variant of the NIPAF is the *iterated principal axis factor analysis* (IPAF), which begins with the same calculation procedures as NIPAF. However, after calculating the final communalities, they are reinserted into the NIPAF’s algorithm as new communality estimates to recalculate λ. This is because the final communalities are likely a better estimate of the communalities than the initial estimates. Each cycle of recalculating communalities and factor loadings is referred to as an *iteration*. These cycles repeat until the procedure has *converged* on a solution (i.e., the initial communalities are nearly identical to the final communalities for a cycle). Simulation studies have shown that IPAF tends to produce slightly more accurate results than NIPAF. However, in some cases, IPAF can lead to improper solutions (i.e., impossible values).

Like IPAF, *maximum likelihood* (ML), another model fitting method, is an iterative procedure. However, ML differs from NIPAF and IPAF in its mathematical definition of closeness. ML’s fitting function is referred to as the *likelihood function*, which determines a set of parameter estimates for a model that are most likely to produce the observed data. ML is based on the assumptions that the data are based on a random sample drawn from some defined population and that the measured variables have a multivariate normal distribution. For its calculations, ML uses the *ML discrepancy function* (F_{ML}), which is inversely related to the likelihood function. In this function, zero indicates perfect fit and larger values indicate poorer fit.

The main advantage of ML is its ability to produce model fit information, referred to as *model fit indices*. Conversely, a limitation of ML is its lack of robustness to severe violations of multivariate normality. *Robust ML* (MLR), a variant of ML, can be used to deal with normality violations. MLR procedures are similar to ML except the fit indices and standard errors of parameters are adjusted. As a result, ML and MLR will always have the same parameter estimates but different fit indices and standard errors if there are normality violations.

One other model fitting approach that is commonly used is *principal components analysis* (PCA). Although PCA is regarded by many as another model fitting procedure, it is actually fundamentally different from the other model fitting procedures discussed so far. All prior model fitting procedures are different methods to solve the equation: $R - D_\psi = \lambda\lambda^T$. PCA, however, attempts to solve for another mathematical model: $R = \lambda\lambda^T$, as PCA assumes that all unique variances are 0. Thus, PCA solves for λ using R (i.e., the unreduced correlation matrix), as opposed to $R - D_\psi$ (i.e., the reduced correlation matrix). Critics of PCA argue that it is unrealistic to assume no unique variances and that PCA lacks formal indices of model fit.

Determining Number of Common Factors

After choosing a model fitting procedure, the researcher must determine the number of common factors to be specified in the model. Ideally, the number of common factors in a model should do well in accounting for the correlations among measured variables, and all common factors should be readily interpretable. Additionally, one less common factor should substantially undermine the performance of the model and one more common factor should not appreciably improve the model's performance. Methodologists have proposed a number of procedures for determining the appropriate number of factors to include in the model. Decisions regarding the number of factors to specify for a model should be based on multiple procedures.

Eigenvalues-greater-than-1 (Kaiser criterion)

Eigenvalues are numerical indices calculated from the reduced correlation matrix. These values correspond to the variance in measured variables accounted for by each common factor, with the largest value corresponding to the first factor, the second largest value the second factor, and so forth. Researchers should be aware that it is also possible to compute eigenvalues from the unreduced correlation matrix, but these values do not correspond to the variance accounted for by common factors but rather principal components. Thus, procedures based on reduced matrix eigenvalues are more conceptually sensible for factor analysis and unreduced matrix eigenvalues are more sensible for PCA.

One common procedure based on eigenvalues is the *eigenvalues-greater-than-1* procedure, which involves retaining as many common factors as eigenvalues greater than 1. Although this procedure is quite simple, simulation studies have shown that it performs poorly. Furthermore, this method is conceptually inappropriate for FA because the threshold value of 1 was developed for the eigenvalues of PCA—not FA.

Scree Test

The scree test involves plotting the eigenvalues generated from the reduced correlation matrix in a graph in descending order. The number of common factors can be determined by counting the number of eigenvalues that precedes the last major drop in the graph. Although the scree test has been criticized as a relatively subjective procedure, it performs reasonably well when strong major common factors are present in the data.

Parallel Analysis

Parallel analysis involves generating eigenvalues that would be expected from random data with the same sample size and number of measured variables as the observed data and subsequently, comparing these expected random eigenvalues with the real eigenvalues. The appropriate number of factors is the number of eigenvalues from the actual data set that are greater than their corresponding eigenvalues expected from random data. Parallel analysis procedures have performed well in simulated data sets where strong common factors are present.

Model Fit

If ML or MLR is used, the chi-square (χ^2) test of perfect fit can be used to determine the number of common factors. This index tests the hypothesis that the model holds perfectly in the population. The method begins by testing the goodness of fit for a one-factor model. This model is retained if it is nonsignificant. If not, then common factors are added to the model until a model is found to produce a nonsignificant test. Critics of this method argue that a test of perfect fit is an unrealistic standard and that its sensitivity to sample size makes it likely to overfactor in large samples and underfactor in small samples.

Because of these limitations, some methodologists have proposed using descriptive fit indices. These indices quantify the magnitude of the lack of fit between the model and data rather than simply categorizing the model into perfect or imperfect fit. Thus, these tests have a more realistic assumption of model fit. There are many descriptive fit indices, including the *root mean square error of approximation* and *nonnormed fit index* (also called the *Tucker–Lewis index*). Regardless of the descriptive fit index used, the number of factors can be determined by examining the model fit of a one-factor model and then examining if a model with an additional factor substantially improves model fit. The appropriate number of factors is a model that produces a good fit to the data and for which the addition of another factor produces no appreciable improvement in fit.

Stability and Interpretability of Solutions

Good models should produce replicable and interpretable parameter estimates. Thus, it is also useful to compare models with differing numbers of factors with respect to the extent to which the parameters of the models are stable across data sets or subsets of a given data set. Researchers can also examine the interpretability of the solutions for models with differing numbers of factors. When interpreting the solution, all measured variables with substantial loadings on a common factor should share a readily interpretable common theme. Difficulty generating a single common theme for a common factor may reflect underfactoring. Obtaining common factors with only a single measured variable with a substantial loading or no measured variables with substantial loadings are common symptoms of overfactoring.

Rotating Factor Analysis Solutions

After model fitting, a solution must be transformed, or rotated, to enhance interpretability, as there will be an infinite number of solutions that fit the data equally well when a model has two or more factors. Rotation refers to the process of selecting the most readily interpretable solution among these equally fitting solutions. Most rotation procedures attempt to select the solution with the best *simple structure*. The criteria of simple structure imply that each common factor should have high loadings for a subset of measured variables and low loadings for the remaining variables. Additionally, these subsets defining different factors should not substantially overlap. Furthermore, each measured variable should be influenced by only a subset of the common factors.

Different mathematical functions called *simplicity functions* have been developed to define simple structure. The goal of *analytic rotation* is to find a solution out of an infinite number of solutions that best satisfies the simplicity function of a rotation method. Because this rotation process does not alter model fit, all measures and tests of fit as well as communalities are not affected by the rotation process.

Orthogonal analytic rotation assumes there are no correlations among common factors. The most common orthogonal rotation is *varimax*, which is a procedure that selects the solution that maximizes the variance of the factor loadings for each common factor. Varimax has typically performed well when its assumption of orthogonal factors holds. However, many methodologists have questioned the assumption of orthogonal factors. Thus, it has been suggested that *oblique analytic rotation*, which allows common factors to correlate, is a more appropriate approach to rotation. The three most common oblique analytic rotation methods are *promax*, *orthoblique* (Harris–Kaiser rotation), and *direct quartimin rotation*.

Despite the more realistic assumption of oblique rotation, orthogonal rotation, especially varimax, has been the most widely used rotation in the literature. The popularity of orthogonal rotation could in part be based on several misconceptions about the two types of rotation. First, some researchers incorrectly believe that oblique rotation requires common factors to be correlated. Rather, oblique rotation merely permits factors to correlate. If better simple structure exists for a solution with orthogonal factors, oblique rotations will produce solutions with uncorrelated common factors. Second, some believe that orthogonal rotation leads to better simple structure than oblique rotation. However, the opposite is true. In an oblique rotation, the spurious effects between a common factor and a measured variable are removed via the control of the influence of other common factors on the measured variable. Finally, some researchers believe that if they wish common factors to be uncorrelated, this can be accomplished using an orthogonal rotation. However, using an orthogonal rotation does not change the underlying structure of the data. If the latent variables underlying a set of measured variables are correlated, rotating to solution that assumes uncorrelated factors merely masks but does not change this reality.

Leandre R. Fabrigar and Matthew P. H. Kan

See also Confirmatory Factor Analysis; Path Analysis; Structural Equation Modeling

Further Readings

Fabrigar, L. R., & Wegener, D. T. (2012). Exploratory factor analysis. New York, NY: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272.

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Harman, H. H. (1976). Modern factor analysis (3rd ed.). Chicago, IL: University of Chicago Press.

Gail Vallance Barrington Gail Vallance Barrington Barrington, Gail Vallance

External Evaluation External evaluation

External Evaluation

An external evaluation is conducted by an evaluator who is not employed by the organization that has commissioned the evaluation. An external perspective brings credibility to the process because it is perceived as more objective and accountable than an evaluation conducted by internal staff. It may also foster innovative thinking.

The external evaluator or team is governed by the terms of a contract that specifies the evaluation tasks and the duration of the project. The contract focuses the evaluator on the parameters of the evaluation process itself and limits involvement in broader organizational issues. It can be easily (though not necessarily painlessly) severed by both parties and so the relationship is both temporary and accountable. Its arm’s length nature allows the external evaluator to interact objectively with staff and stakeholders without fear of reprisal and report findings (both negative and positive) without affecting their own career aspirations. The external evaluator appears to have less to gain or lose from the evaluation findings and is less vulnerable to conflict of interest.

On the other hand, internal evaluators have a personal stake in the success of the organization. They have a clear advantage in terms of their understanding of organizational history, culture, context, and the players involved but can be hampered by the implications of reporting negative findings to their employer. In larger organizations, being a union member or located in a particular department can limit the likelihood of staff interacting openly with them.

The compensation arrangements made for an external evaluation tend to heighten accountability and enhance the drive for completion. While the internal evaluator receives a salary regardless of the stage of the evaluation project, the external evaluator must adhere to completion timelines to get paid. Transparency external evaluator must adhere to completion timelines to get paid. Transparency is also increased by contractual requirements for interim documentation as status or technical reports tied to milestone payments.

While an internal evaluator understands the politics of the program under review, an external evaluator has worked with many other programs and communities and brings a broader perspective. This cross fertilization can result in fresher, more innovative approaches to the evaluation problem. Through a series of engagements in different settings, external evaluators have learned many ways to interact with stakeholders, collect data, and present findings. Thus, they are better able to respond to the unexpected as the evaluation unfolds.

Selecting an internal or external evaluator may be situational and sometimes budget driven, because internal evaluators are available at no additional cost. In complex organizations such as the federal government, a joint internal–external team may be the best solution combining a credible and objective external perspective with the internal knowledge required to shepherd the study through to completion.

Gail Vallance Barrington

See also Ethical Issues in Evaluation; Internal Evaluation; Objectivity

Further Readings

Harvey, L., & Newton, J. (2004). Transforming quality evaluation. Quality in Higher Education, 10(2), 149–165.

Kemmis, S. (1986). Seven principles for programme evaluation in curriculum development and innovation. New Directions in Educational Evaluation, 2, 117–140.

Gregory Mitchell Gregory Mitchell Mitchell, Gregory

External Validity

External validity

654

658

External Validity

External validity refers to the degree to which the relations among variables observed in one sample of observations in one population will hold for other samples of observations within the same population or in other populations. External validity is often treated as synonymous with the generalizability of results. Whenever empirical research makes use of a sample to examine how two or more variables are related within a larger population or whenever one seeks to extend results drawn from one population to a new population, questions of external validity arise.

External validity is often contrasted with internal validity or the question of whether valid inferences can be reached regarding the existence and nature of a relationship between two variables. Efforts to increase internal validity often reduce a study’s external validity. The key method for increasing external validity is to employ representative sampling with respect to all aspects of empirical design. External validity can be assessed inductively and deductively: Inductive assessments involve reviewing the relevant empirical literature to determine the conditions under which a research finding did or did not generalize; deductive assessments involve applying existing theoretical and empirical knowledge to deduce conditions on generalizability. External validity should be of particular concern when empirical research aims to serve as a guide to public policy. This entry contrasts external validity with internal validity, then discusses ways to increase external validity and describes the assessment of external validity.

External ValidityContrasted With Internal Validity

Many textbooks and articles on research design published before 1957 discuss how the sampling techniques used to gather observations will affect the generalizability of research findings. After 1957, it is much more common to find generalizability discussed in terms of the external validity of a research design. This change in terminology, and the increase in attention given to questions of generalizability, followed publication in 1957 of an article by the psychologist and methodologist Donald Campbell in which Campbell reframed the generalizability question as one of "external validity" to be contrasted with questions of "internal validity." In this 1957 article and a series of subsequent influential publications, Campbell and coauthors examined threats to external validity and ways of increasing and assessing external validity. Although many scholars have made important contributions on the topic of external validity, Campbell and his colleagues' work on this topic continues to serve as the foundation for most other discussions.

The primary insight of Campbell, and his reason for placing internal and external validity in contrast, was that steps taken in the research design process to increase internal validity often decrease external validity, and steps taken to increase external validity often decrease internal validity. Thus, if we limit who may participate in a study to reduce the chance that individual-level differences will confound the result (resulting in greater internal validity), we cannot be sure the results will generalize to other groups of persons (resulting in less external validity). A compromise would be to study a more representative sample drawn from the population about which one wants to draw inferences while measuring the individual difference-level variables that the researcher has reason to believe may affect the nature or degree of relationship between the target variables (e.g., the researcher may record the sex of students who serve as participants in a study on the relation of teaching style to course grades in order to examine whether male and female students exhibit the same pattern of results).

Researchers wisely focus on internal validity when designing empirical research because drawing valid inferences about what relates to what and why (when the research design permits causal inferences) is the basic reason for conducting empirical research. However, a focus on internal validity to the exclusion of external validity may produce internally valid results that fail to generalize or perhaps even fail to address the aspects of the real-world phenomena that motivated the study in the first place.

Ways to Increase External Validity

Within the social sciences, concerns about external validity often focus on the characteristics of the persons who served as participants in a study, asking whether these participants were good representatives of the population under study. This concern is valid, but other aspects of research design also impact external validity. Questions of external validity arise whenever empirical research makes use of a sample with respect to the units of analysis (e.g., a sample of students drawn from the full population of students), settings (e.g., elementary schools within a particular school district), treatments or explanatory variables (e.g., particular types of teaching or testing methods), outcomes of interest and measures of those outcomes (e.g., course grades or student engagement as measured through a questionnaire vs. through behavioral observation), and time periods. The same issues arise when one is able to observe all members of a population, but one seeks to extrapolate from this population to another population. In addition, the manner in which a study is conducted may, by necessity, create alterations in the environment or behavior being observed, and this reactivity to the research design may reduce the external validity of the study's results.

The trade-off between internal and external validity is most severe in research programs that use convenience sampling (i.e., use of samples based on who or what can be conveniently studied) and laboratory experiments to study how humans navigate complex social environments. In these areas of research, highly controlled experiments produce internally valid findings with suspect external validity. However, steps can be taken to increase external validity even within research domains that rely heavily on convenience samples and experimental designs.

The primary means of increasing external validity is the use of representative sampling with respect to the units of analysis and other research design elements that may impact external validity, and the primary means of achieving a representative sample will be through random sampling from the population of interest. Although truly representative sampling of units of analysis may be difficult to achieve due to lack of access to the full population, representative sampling along other dimensions may be more manageable. Education researchers, for instance, may be able to sample from the full population of schools, teaching materials, and assessment methods employed within a school system and may be able to sample across the full academic year or even over multiple years. Too often, researchers use convenience samples based on the ease of access or implementation, even when more representative sampling could be used.

If representative sampling is not possible due to resource or feasibility constraints (e.g., with children as participants, parental or guardian assent must be obtained, which increases the cost and difficulty of obtaining a random sample), a sampling plan should be used that prioritizes sample characteristics for each dimension of research design, from the unit of analysis to time periods. The priority among sample characteristics will depend on the purpose behind the research and the theory or hypothesis being tested. For instance, if an educational intervention is primarily aimed at increasing achievement among low-income students but it is hoped that the intervention will have general positive effects, deliberate sampling from low-income students should be a priority. Too often, researchers accept the sample to which they have easiest access whenever representative sampling is not possible, even though a more suitable sample could be employed than the convenience sample that was accepted.

Replication with new samples is another important method for increasing external validity. Replication provides useful information about conditions on the results of prior studies and, in many instances, will reveal the dimensions along which results generalize or fail to generalize. As with sampling, a priority plan for replication should be developed that seeks to address the most relevant external validity concerns given the purpose behind the research.

Where there is concern that the research design involves reactive elements (i.e., merely conducting the research will somehow alter the behavior under observation), consideration should be given to the use of placebo groups in addition to a control group (i.e., a group that is exposed to the same levels of interest and observation as the experimental group but that does not receive the treatment that the experimental group received). Where there is a concern about reactivity in a prior study, replications should employ designs that use nonreactive or low-impact observational measures to the extent possible.

Finally, researchers utilizing experimental or quasi-experimental designs should attempt to achieve both psychological realism and mundane realism. Psychological realism refers to engaging the research participant in the same way that the real-world phenomena of interest engage people (to the extent it is ethically possible to do so), and mundane realism refers to simulating the real-ethically possible to do so), and mundane realism refers to simulating the real world environment and tasks of interest as much as possible in the research setting.

Assessing External Validity

External validity may be assessed inductively and deductively. Inductive assessments involve comparing results across studies that employed different samples to examine the consistency of research findings across these samples and to identify predictable variations in the nature and strength of relations among the target variables of interest. The preferred means of inductively assessing external validity is through meta-analysis or the technique of systematically surveying a body of research and quantitatively synthesizing the studies that have examined a common research question to establish the robustness of the relationships among the variables of interest.

If an insufficient number of studies exists to conduct a meta-analysis, then new studies should be conducted to test whether an alteration in the sampling or research design affects the research outcome. When the initial studies on a topic have all been conducted in a laboratory or simulated settings, priority should be placed on replications in the field. Research has shown that many of the findings of experimental studies from psychology, for instance, fail to replicate in the field.

Deductive assessments involve applying existing theory and empirical research to determine the factors that are likely to produce differences in variable relations across samples (or populations). The confidence one should place in deductive external validity assessments depends on the nature and reliability of the background knowledge: If one factor is well established as a factor that does or does not affect outcomes on variables similar to those in the research of interest, then one may confidently predict external validity in samples (or populations) where that factor is present. However, because it is difficult to establish all potential conditions on external validity deductively, the primary function of deductive external validity assessments should be to serve as the basis for replication research aimed at inductively assessing external validity with respect to those factors most likely to produce a change in the previously observed results.

Assessing the external validity of a body of research is an important aspect of good science. Researchers understandably emphasize internal validity in good science. Researchers understandably emphasize internal validity in research design because drawing valid inferences about the relations among variables is a prerequisite to meaningful results. However, the purpose of most social science is to identify predictable patterns of behavior and cause-effect relations; thus, external validity is also crucial to developing a reliable theory that can explain and predict behavior outside the specific research setting. Where research is undertaken to inform public policy, particular attention should be given to external validity because internal validity is no guarantee of external validity. Incorrectly assuming that an internally valid result has external validity may be detrimental to the public and wasteful of public resources.

Gregory Mitchell

See also Convenience Sampling; Experimental Designs; Generalizability; Internal Validity; Meta-Analysis; Program Evaluation; Quasi-Experimental Designs; Representativeness; Threats to Research Validity; Validity

Further Readings

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. American Educational Research Journal, 5, 437–474.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297–312.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis for field settings. Boston, MA: Houghton Mifflin Company.

Marcellesi, A. (2015). External validity: Is there still a problem? Philosophy of Science, 82, 1308–1317.

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. Perspectives on Psychological Science, 7, 109–117.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Company.

F

David L. Raunig David L. Raunig Raunig, David L.

F Distribution

F distribution

659

662

F Distribution

The *F* distribution is behind perhaps only the *t* distribution as the most popular statistic in tests of statistical significance when using continuous response variables. The *F* statistic is directly related to the chi-square statistic and the explicit output of all analysis of variance (ANOVA) tests. It is the foundation of modern experimental design and is the most powerful test for comparing two variances when the assumptions of normality are met. The following sections covers the history and derivation and properties of the *F* distribution, a short review of the noncentral *F* distribution and an ANOVA example, the resulting *F* statistic, and relevance to the distribution of *F* values.

History of the *F* Distribution

When the objective of an analysis is to compare two variances, the ratio of the variances is preferred to the difference due to the statistical properties of the latter. In 1924, Sir Ronald Fisher drew from work done by Student (W. S. Gosset) and introduced the *z* distribution to describe the distribution of the ratio of two chi-square-distributed random variables, in this case sample variance estimates. George Snedecor modified the *z* slightly in 1934 to define *F*, named in honor of Fisher. Using the *z* statistic, two sample variance estimates were compared to determine whether their ratio was greater than a hypothesized population variance ratio when the population variances are known. For two normally distributed and independent populations,

$$X_1 \sim N\mu_1,\sigma_1^2$$

$$X_2 \sim N\mu_2,\sigma_2^2,$$

with μ_1 and μ_2 as the population means and and the population variances, z is defined for sample variances s_1 and s_2 as:

$$z = 12\ln s_1 s_2 = 12\ln\sigma_1^2\sigma_2^2 \cdot n_2 SS_1 n_1 SS_2,$$

where *ln* is the natural logarithm, SS_1 and SS_2 are the sums of squares for the two comparison groups and n_1 and n_2 are the degrees of freedom for the numerator and denominator. The sampling errors of s are proportional to σ, but the sampling errors for log(s) depend only on n, which will be important when describing the shape of the F distribution.

Snedecor restructured the z statistic to arrive at the F distribution where F is defined as:

$$F = s_1 s_2 = \sigma_1^2\sigma_2^2 \cdot n_2 SS_1 n_1 SS_2.$$

Under the null hypothesis that the two variances and are equal, F reduces to the more common form:

$$F = SS_1 n_1 SS_2 n_2$$

or

$$F = MSS_1 MSS_2,$$

where *MSS* is the mean sum of squares. Snedecor expanded on the interpretation of the ANOVA, reasoning that the within-groups variance correctly describes the error variance and that if the data are randomly sampled from the same homogeneous population, then the between-group variance would not be expected to be statistically different than the within-group variance. The F statistic is then defined for ANOVA as the ratio:

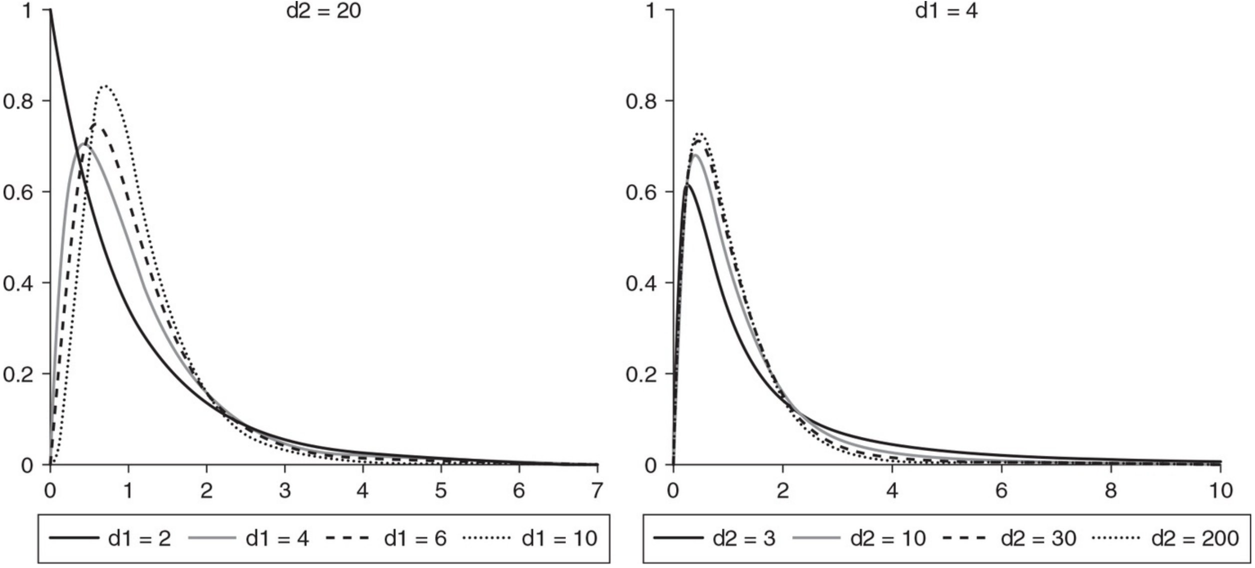
$$F = \text{between} - \text{groups} \cdot \text{within} - \text{groups},$$

and tested for equivalence to 1, or more practically, tested for $F > 1$.

Properties of the *F* Distribution

The distribution of the *F* statistic for the null hypothesis that all groups are sampled from the same distribution is dependent only on the degrees of freedom of both the numerator, n_1, and the denominator, n_2, often referred to as the numerator and denominator degrees of freedom, *ndf* and *ddf*, respectively. The *ddf* is also often the error degrees of freedom. The effects of *ndf* and *ddf* on the shape of the *F* distribution are shown in Figure 1. Larger *ndf* moves the mode (or peak) to the right and lowers the distribution in the far right tails. In fact, as *ndf* → ∞, the distribution approaches the chi-square distribution and *ndf* = 1 leads to the *t* distribution. Larger *ddf* results in a similar effect by lowering the far right tails. The effect of reducing the fatness of the right tails is to lower the probability that an *F* statistic larger than the one observed would occur by chance. This probability is the *p* value in ANOVA tables.

Figure 1 Different *F* distributions for (a) various numerator and (b) denominator degrees of freedom



Critical values, typically seen as $F(1 - \alpha, ndf, ddf)$, assumes a one-sided significance based on the expectation that between-group variance is not expected to be smaller than the error variance. Two-sided *F*-critical values can use the following relationship:

$$Fndf,\ ddf,\ 1-\alpha/2 = Fddf,\ ndf,\ \alpha/2-1.$$

Noncentral *F* Distribution

A word must be said about the *F* distribution when the null hypothesis is not true, or

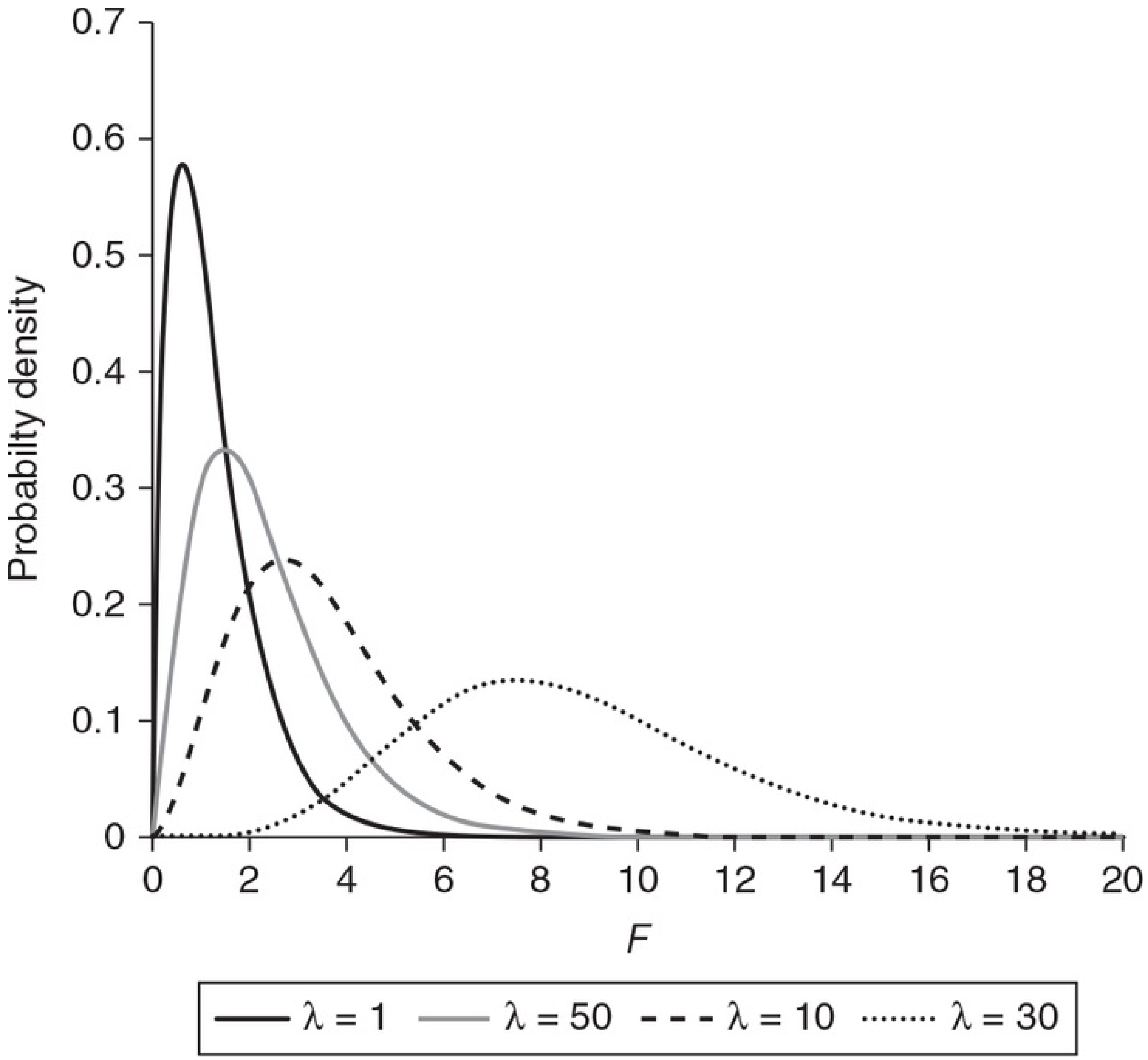
$$\sigma_1^2 \neq \sigma_2^2.$$

In this case, the numerator is noncentral chi-square distributed and the *F* distribution is also noncentral. The noncentral *F* distribution is very complex and critical values tabulated in statistical references. This distribution is used in sample size calculations. The central *F* distribution can be fully described by *ndf* and *ddf*, but the noncentral *F* distribution requires the noncentrality parameter, λ, to describe the location of the distribution. This parameter is calculated from the ANOVA outputs as:

$$\lambda = SS_{\text{between}} MSS_{\text{error}}.$$

Figure 2 shows the shift of the *F* distribution with *ndf* = 4, *ddf* = 20, and λ = various. The shift represents the difference in the between-group difference from the null hypothesis.

Figure 2 The effects of the noncentrality parameter on the shift to the right of the *F* distribution



Example

The following example simulates five independent groups, each with the following parameters:

Within-group means: 7, 9, 10, 5, and 3
Within-group standard deviation: 1
Number in each group: 5

All plots and analyses used JMP statistical analysis software, Version 10. The data are plotted in Figure 3 with the group means shown in Table 1 and ANOVA results in Table 2. The position of the *F* statistic on the *F* distribution is shown in Figure 4 on the next page.

Figure 3 ANOVA Example 5 groups. The Horizontal is the overall mean

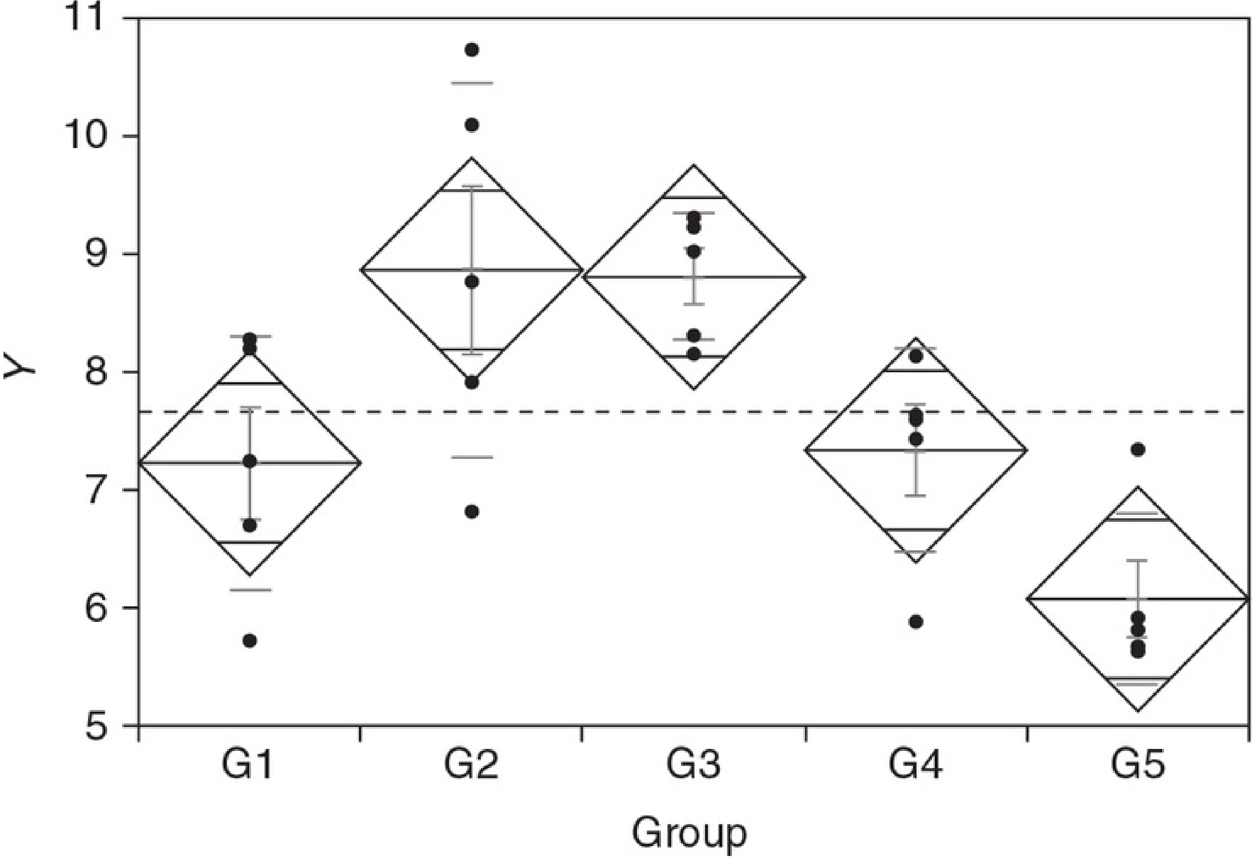
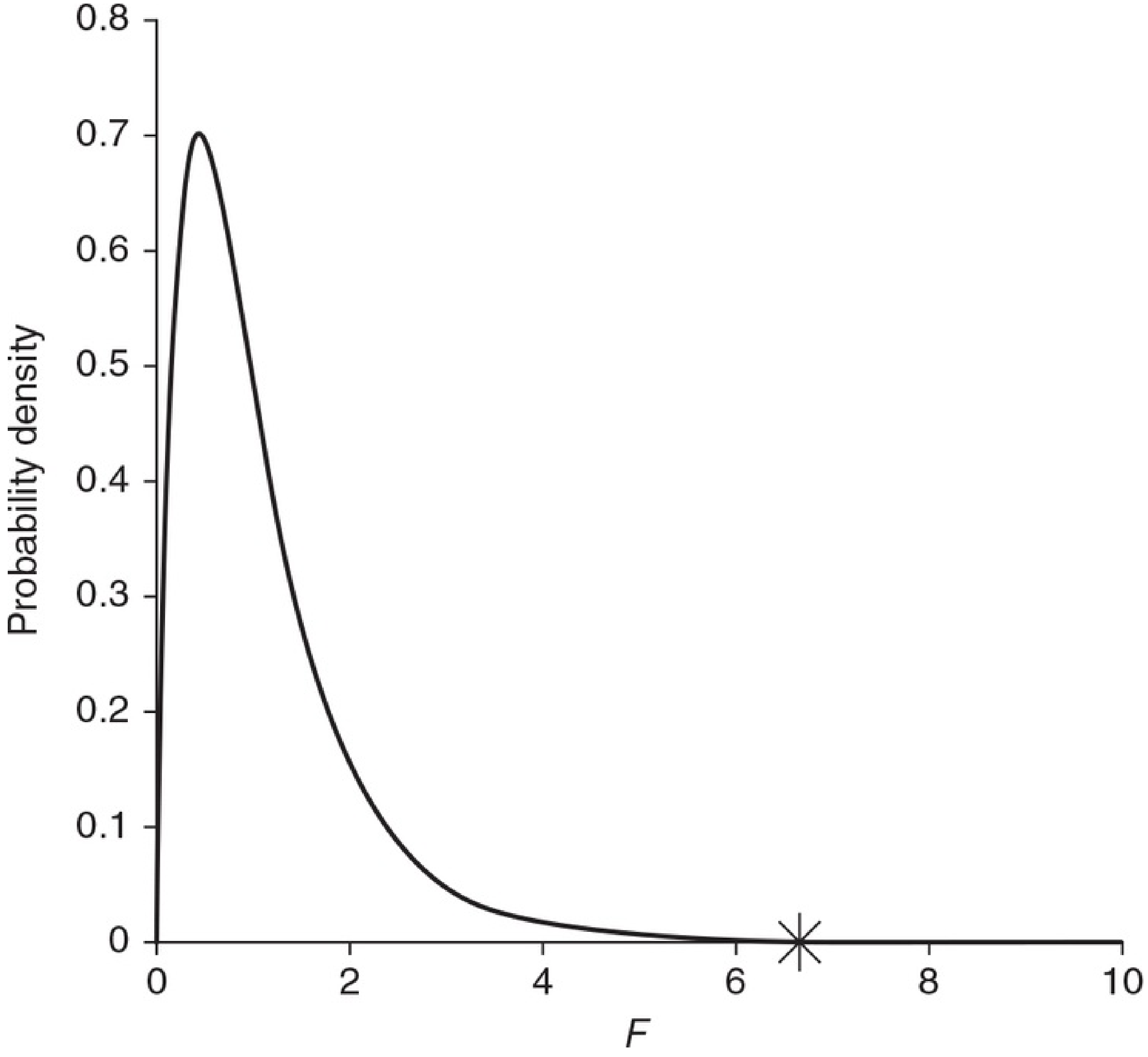
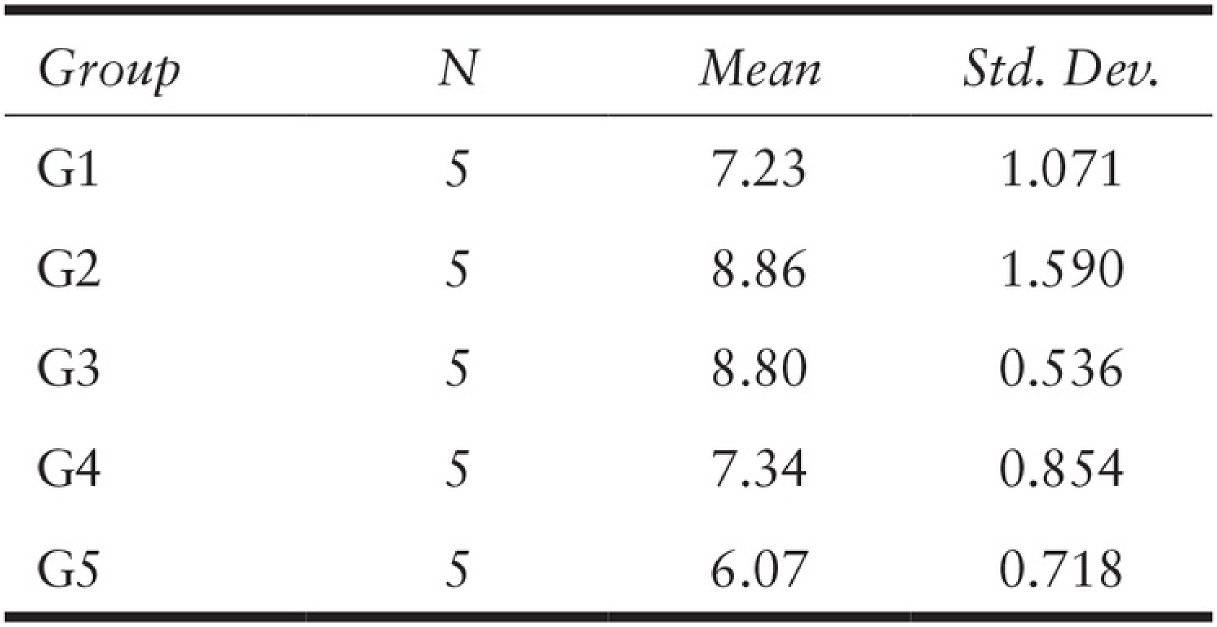
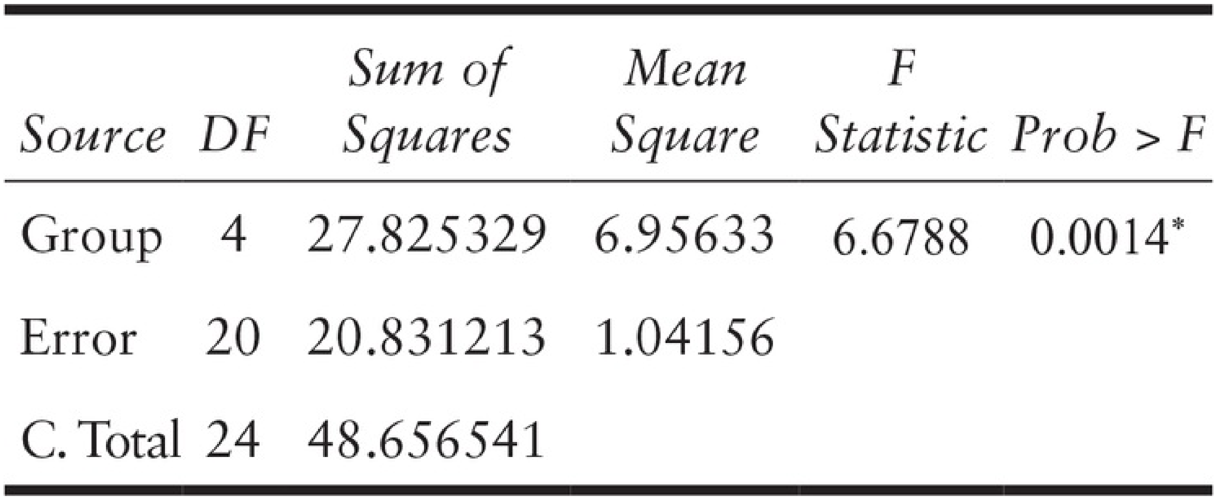


Figure 4 ANOVA Example *F* ratio on the *F* distribution. The asterisk is located at the *F* statistic


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Probability density
0
2
4
6
8
10
F



Group	N	Mean	Std. Dev.
G1	5	7.23	1.071
G2	5	8.86	1.590
G3	5	8.80	0.536
G4	5	7.34	0.854
G5	5	6.07	0.718



Source	DF	Sum of Squares	Mean Square	F Statistic	Prob > F
Group	4	27.825329	6.95633	6.6788	0.0014*
Error	20	20.831213	1.04156		
C. Total	24	48.656541			

David L. Raunig

See also Analysis of Variance; Chi-Square Test; *t* Tests

Further Readings

Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. Paper presented at the Proceedings of the International Congress of Mathematics.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). Continuous univariate distributions (Vol. 2, 2nd ed.). New York, NY: Wiley.

Snedecor, G. W. (1934). Calculation and interpretation of analysis of variance and covariance.

Natalja Menold Natalja Menold Menold, Natalja

Falsified Data in Large-Scale Surveys Falsified data in large-scale surveys

662

665

Falsified Data in Large-Scale Surveys

Falsifications of survey data may be classified by the source of the falsification: data providers, interviewers, or interviewed persons. This entry focuses on interviewers’ falsifications. Interviewers’ falsifications cause significant problems with the reported data and are difficult to identify reliably. The reported prevalence of falsifications by interviewers in cases where established quality standards and controls are used is low. Falsifications may seriously impact on data analysis results, regardless of how frequent they are in the data. Methods of detection include control procedures such as reinterview or ex-post data analysis methods. This entry first discusses large-scale survey data and forms of falsification before describing specific cases of interviewers’ falsifications, how data are contaminated through interviewers’ falsifications, and methods to detect falsifications.

Large-Scale Survey Data

Large-scale surveys are national, federal, or cross-cultural surveys aimed at providing high-quality data on the target population under investigation. The data can concern a broad range of topics, for example, consumption, skills, health, opinions, or behavior. Large-scale surveys are based on probabilistic sampling procedures; the representativeness of the data for a certain population can be evaluated via, among others, response and nonresponse statistics. Data production in large-scale surveys is a significant cost factor and data are used by secondary data analysis research or by official reports. Examples include the Programme for International Student Assessment, the Programme for the International Assessment of Adult Competencies, the European Union Statistics on Income and Living Conditions, and the European Social Survey.

Forms of Falsification

One potential source of falsification in large-scale probabilistic sample surveys is the duplication of cases by survey data providers to increase the number of observations and to fulfill specific requirements for response rates. Another potential source of falsifications relates to interviewers involved in the data collection process. Large-scale survey data can be collected via different modes of data collection, such as mail, Internet, telephone, or face-to-face interviews. Interviewers play a key role in telephone and face-to-face surveys. Face-to-face surveys are the most valuable data collection mode because they make it possible to obtain higher representation of the population under investigation and enable the collection of a significant amount of data during a single interview. Interviewers may be a relevant source of both higher data quality and biased survey statistics. Bias is possible due to mistakes and deviations from prescribed behavior.

The American Association of Public Opinion Research defines interviewers' falsifications as intentional deviations from standards and instructions that result in data contamination. The American Association of Public Opinion Research differentiates between four categories of falsifications by interviewers: 1. The first category is partly or fully falsified interview data and is referred to as *interview falsifications*. In the case of interview falsifications, data were not provided by a target person but by the interviewer. In the case of partly or fully falsified interviews, interviewers may collect some central data from the target person, such as gender, age, and characteristics of residence, and then respond to survey questions from the point of view of this person.

2. The second category is falsifications of process data (paradata), such as misreporting of the number and times of contact attempts needed to reach a target person or misreporting contact results. Examples include declaration of a sampling unit as not legible and reporting that a target person refused to be interviewed when in fact the person was not contacted.

3. Falsifications of certain responses, to shorten the interview, belong to the third category. Examples would be specific responses that filter questions to avoid subsequent, more detailed questions, or underreporting of the number of persons for whom additional information should be provided.

4. The fourth category of falsification is interviewing of persons outside the

sample and using their data to substitute for those who refused or were not contacted.

The last potential source of falsifications would be the respondents themselves. Faking is a research area in psychology that addresses lying, straightlining (answering the same way each time regardless of the question), impacts of social desirability, and other forms of misreporting by respondents.

Cases of Interviewers’ Falsifications

The prevalence of falsifications of data by interviewers seems to be relatively low when considering surveys for which established standards and extensive procedures to control interviewers are used. If controls are conducted, falsifications of interview data are often obtained, but their extent is relatively limited.

For the 1986 U.S. Census, controls for randomly selected interviewers were conducted. Irregularities were found for 3–5% of all interviewers, with 70% of these irregularities involving interview falsifications. The next case considers a German population survey (ALLBUS). The first controls were conducted in 1994. Information on gender and age from population registers was used to control the data collected by the interviewers. After the survey, respondents were contacted again in cases where data collected by interviewers deviated from the register information, revealing 45 cases (1.2%) that were classified as full interview falsifications. The ALLBUS has been controlling the data with this method ever since and publishes the results of its controls. The proportion of potentially, partly, or entirely fabricated interviews was 1.3% in 2010 and 0.8% in 2012. Falsifications are more likely to occur when a survey is on a sensitive topic and survey participation is more difficult to obtain, compared to surveys with nonsensitive topics. “Painful” experiences were reported when conducting a population survey on sexually transmitted diseases in Baltimore. Researchers observed that some interviewers had very high rates of cooperation from target persons, although, in general, it was difficult to obtain cooperation. The controls by recontact revealed that almost 50% of the interviews delivered by interviewers with high cooperation rates were falsifications. In the American National Survey on Drug Use and Health in 2002, controls were conducted by telephone recontacts and time stamps. These controls identified some interviewers who falsified nearly 70% of the interviews they claimed to have conducted.

Finally, falsifications would be a severe problem when collecting data under circumstances in which survey infrastructures and control procedures have been not established. Sebastian Bredl, Peter Winker, and Kerstin Koetschau conducted a small survey in rural areas in a less-developed country in 2007 and 2008. They became suspicious after receiving 50 interviews delivered by their five interviewers and conducted face-to-face reinterviews for all cases. With this method, all 50 interviews were found to be falsified.

Data Contamination Through Interviewers' Falsifications

Questions of bias associated with interviewers' falsifications have been raised by some of the cases described in the previous section and other cases of falsification. In the case of the National Survey on Drug Use and Health falsification, it was found that falsifiers reported higher drug consumption than was found in the real data. In the Baltimore survey, also described in the previous section, sexually active behavior was overestimated in the falsified data.

In several experimental studies that systematically compared falsified and real data, falsifications were produced by instructions using descriptions of real survey participants, so that parallel falsified and real data could be obtained and compared. The results demonstrate that interviewers were very often able to "predict" real responses, so that means and distributions with respect to the opinions, behavior, knowledge, or personal characteristics differed marginally between the real and the falsified data. However, for some specific contents, significant differences were obtained. For example, further political participation was underestimated and past political participation was overestimated by the falsifiers in a study conducted by Natalja Menold and Christoph Kemper. Moreover, a number of studies report that the data differ with respect to the dispersion masses because falsifiers provide less differentiated data.

Methods to Detect Falsifications

One method to detect falsifications is to collect and inspect paradata that are produced during interview processing, on the case level, for each contacted unit. Examples would be date and time stamps, in the case of computer-assisted interviews, or inspection of the results of contact attempts by an interviewer.

interviews, or inspection of the results of contact attempts by an interviewer. Transcription of each interview would be an effective method, but it is rarely used in face-to-face surveys.

A very commonly used method to control interviewers is recontact and reinterviewing of the respondents. During the reinterview, respondents are asked to confirm that the interview took place, to respond to some detailed questions about the interview, and also to repeat their responses to selected survey questions. Such reinterviews are usually conducted by sending postcards or by calling the respondents by phone. The method that is used depends on the availability of contact data and the survey budget.

Because it is hardly practicable to recontact all respondents, and due to the fact that random selection of respondents for the recontact would be less effective, methods are used that allow for a more focused selection of cases for the recontact. So-called at-risk interviewers have been identified by a combination of different methods, such as using paradata, register information on respondents, or other estimation methods.

A relatively new approach is the multivariate indicator–based method, developed by Winker and colleagues. The concept uses multivariate cluster analyses with optimized classification to separate falsified data from real data. With this method, in particular, content-independent differences in response behavior between real respondents and potential falsifiers are used, such as item nonresponse, filtering questions, acquiescence, and presentation order effects.

Each of the methods is associated with limitations; for example, using paradata such as time stamps requires computer assistance and is limited in its discovery of falsifications because these could also be provided with plausible time stamps. Recontacts are limited with respect to the availability of contact data, willingness of interviewed persons to respond a second time, memory effects, and the limited amounts of information that can be collected. Multivariate analyses methods, finally, can only locate suspicious cases, which cannot be declared as falsifications without additional information. Therefore, only a multitude and combination of methods can guarantee successful identification of falsifications. Methods also differ in their usefulness in detecting a certain kind of falsification; for example, recontact can deliver information on interviewing of nonsampled persons whereas other methods, such as transcription, cannot. Partly falsified interviews, in particular, can rarely be identified by any method. More research is needed not only on falsifications and detection methods but also on interviewer motivation and prevention strategies.

interviewer motivation and prevention strategies.

Natalja Menold

See also Attitude Scaling; Fraudulent and Misleading Data; Interviewer Bias; Nonresponse Bias; Survey Methods

Further Readings

American Association of Public Opinion Research. (2003). Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf.

Bredl, S., Storfinger, N., & Menold, N. (2013). A literature review of methods to detect interviewer falsification of survey data. In P. Winker, N. Menold, & R. Porst. (Eds.), Interviewers’ deviations in surveys: Impact, reasons, detection and prevention (pp. 3–24). Frankfurt, Germany: Peter Lang.

Bredl, S., Winker, P., & Koetschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. Survey Methodology, 38(1), 1–10.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). Survey methodology. Hoboken, NJ: Wiley.

Menold, N., & Kemper, C. J. (2014). How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys. International Journal of Public Opinion Research, 26(1), 41–65. Retrieved from http://dx.doi.org/10.1093/ijpor/edt017

Menold, N., Storfinger, N., & Winker, P. (2011). Development of a method for ex-post identification of falsification in survey data. Proceedings of New Techniques and Technologies for Statistics—NTTS 2011, Brussels, Belgium.

Winker, P., Menold, N., & Porst, R. (2013). Interviewers’ deviations in surveys:

Impact, reasons, detection and prevention. Frankfurt, Germany: Peter Lang.

Matthew B. Fuller Matthew B. Fuller Fuller, Matthew B.

Family Educational Rights and Privacy Act Family educational rights and privacy act

665

667

Family Educational Rights and Privacy Act

The Family Educational Rights and Privacy Act (FERPA; 20 U.S.C. § 1232g; 34 CFR Part 99) is a U.S. federal law designed to protect students' privacy and personal access to educational records. FERPA applies to any educational agency or elementary, secondary, or postsecondary education institution to which federal funds have been made available under any program administered by the U.S. secretary of education. FERPA is one of the most often cited federal acts guiding educational governance. This entry further defines FERPA and its scope and then reviews the history of the act and its various amendments.

Scope of FERPA

FERPA establishes a framework for the disclosure of student records to various agencies or external requestors. The act provides parents of schoolchildren under the age of 18 years and students over the age of 18 with three rights regarding students' educational records: (1) the right to inspect and review the student's education records maintained by the school; (2) the right to request that a school corrects records that the parent or student believes to be inaccurate or misleading and, by extension, the right to file a formal compliant with the U.S. Department of Education's Family Policy Compliance Office (FPCO) if this request is denied; and (3) the right to provide written consent for institutions to disclose student records to requesting organizations. Parents and students' redress for FERPA grievances is to be directed to the Family Policy Compliance Office rather than judicial courts as indicated by the ruling in *Gonzaga University v. Doe* in 2002, as FERPA creates no personal rights to enforcement. Once students turn 18 or attend a school beyond high school, the rights held by their parents transfer to the students, making them *eligible students* under FERPA's definitions and holders of all FERPA rights for their educational records.

Schools must have written consent from the parent or eligible student prior to disclosing personally identifiable information to a requestor in order to release any information from a student's education record. Schools may disclose, without prior written consent, directory information such as a student's name, address, telephone number, e-mail, photograph, date and place of birth, honors and awards, and dates of attendance. However, schools must inform parents and eligible students about directory information and allow parents and eligible students a reasonable amount of time and provide a procedure to request that a school or institution not disclose directory information about the student. Schools must also inform parents and eligible students annually of their rights under FERPA.

However, FERPA (34 CFR § 99.31) allows schools to disclose certain records without consent to the following parties or under the following conditions: (a) school officials with legitimate educational interests, (b) other schools to which a student is transferring, (c) specified officials for audit or evaluation purposes, (d) appropriate parties in connection with financial aid to a student, (e) organizations conducting certain studies for or on behalf of the school, (f) accrediting organizations, (g) to comply with a judicial order or lawfully issued subpoena, (h) appropriate officials in cases of health and safety emergencies, and (i) state and local authorities, within a juvenile justice system, pursuant to specific state law. Moreover, 34 CFR § 99.31 gives parents of eligible students certain rights with respect to their children's education records. According to subsection 8, prior written consent is not required if "the disclosure is to parents, as defined in § 99.3, of a dependent student, as defined in section 152 of the Internal Revenue Code of 1986." Moreover, educators may disclose without prior written consent, firsthand observations of the health and safety of eligible students to parents or in situations of a health or safety emergency (34 CFR § 99.31, subsection 10). Nothing in FERPA prohibits an educator from disclosing to parents information based on the educators' firsthand knowledge or observation of the student's health or safety, provided this knowledge is not based on information contained in an education record.

History of FERPA

FERPA is also commonly known as the Education Amendments of 1974 or the

Buckley Amendment after its principal sponsor, Senator James Buckley of New York. The act was offered as an amendment on the senate floor to a reauthorization of the Elementary and Secondary Education Act of 1965 and signed into law by President Gerald Ford on August 21, 1974. The act was meant to address concerns that educational institutions were including secret information in student records and preventing students from gaining access to their records.

Immediately following enactment of FERPA, educators voiced concerns over the act, particularly about letters of recommendation for college admissions, written under auspices of confidentiality, but now open to student inspection under FERPA. These and other concerns led to significant amendments to FERPA in December 1974, giving postsecondary students the right to inspect and review their records. An August 6, 1979, amendment clarified the concept that states and education officials are allowed to view educational records during audits and evaluations. These two amendments were the first in a string of nine in total, with the next coming more than a decade later in 1990.

The passing of the Campus Security Act in 1990 led to FERPA amendments allowing institutions to disclose to victims of violent crimes results of institutional disciplinary proceedings against alleged perpetrators. In July 1992, the act was again amended to exclude institutions’ law enforcement records from the definition of educational records under FERPA. An October 1994 amendment extended students’ rights to inspect and review education records maintained by state education agencies and certification offices.

The 1998 Higher Education Amendments enhanced institutions’ abilities to disclose the final results of disciplinary hearings, in which students were found responsible for a crime of violence or nonforcible sexual offenses. Congress also added an amendment that allows postsecondary institutions to inform parents if their child has violated a law or school rule pertaining to use or possession of alcohol or illegal drugs. This amendment also included photographs and e-mail addresses as new student directory information that can be disclosed without student consent.

The early 2000s saw the last three amendments to FERPA, primarily in response to concerns over campus safety and terrorism. In October 2000, Congress clarified that FERPA does not prohibit educational institutions from disclosing information about registered sex offenders on their campus. Following the September 11 terror attacks, Congress enacted the USA PATRIOT Act, allowing

September 11 terror attacks, Congress enacted the USA PATRIOT Act, allowing the attorney general to request a court order to acquire educational records in investigations or prosecution of domestic or international terrorism. In January 2002, technical corrections were made to text of the act.

In 2008, following the U.S. Supreme Court decision in *Owasso Independent School Dist. No. I011 v. Falvo* (534 U.S. 426, 2002), FERPA was amended to ensure that peer-reviewed papers were not considered educational records. However, biometric data such as fingerprints and DNA were included, as educational records and educational agencies were permitted to disclose, without consent, educational records to “contractors, consultants, volunteers, and other outside parties providing institutional services and functions or otherwise acting for an agency or institution.” Another amendment to FERPA in December 2011 revised the act’s definition of directory information and clarified terms such as a*uthorized representative* and *education program*. The amendment also authorized educational agencies to publicly disclose student ID numbers that are displayed on individual cards or badges.

Matthew B. Fuller

See also Health Insurance Portability and Accountability Act

Further Readings

Cornell University Law School. (2015, July 1). 34 CFR Part 99—Family Educational Rights and Privacy. Retrieved from https://www.law.cornell.edu/cfr/text/34/part-99

Gonzaga University v. Doe, 536 U.S. 273 (2002).

McDonald, S. J. (2008). The Family Rights and Privacy Act: 7 myths—and the truth. Chronicle of Higher Education, 54(32), A53–A54.

Rooker, L., & Falkner, T. (2013). 2013 AACRAO FERPA Guide. Washington, DC: American Association of Collegiate Registrars and Admissions Officers.

U.S. Department of Education. (2014, February 11). Legislative history of major

FERPA provisions. Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/ferpa/leg-history.html

U.S. Department of Education. (2015, June 26). Disclosure of information from education records to parents of postsecondary students. Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/hottopics/ht-parents-postsecstudents.html

U.S. Department of Education. (2015, June 26). Family Educational Rights and Privacy Act (FERPA). Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

Joseph Calvin Gagnon Joseph Calvin Gagnon Gagnon, Joseph Calvin

Brian R. Barber Brian R. Barber Barber, Brian R.

Feasibility

Feasibility

667

668

Feasibility

Feasibility, as it relates to research, is the extent to which those who implement a research study or an intervention can practically do so within an identified authentic setting. Feasibility can be the central focus of developmental research, as in a *feasibility study*, or a component of a full-scale intervention trial when used to evaluate effectiveness under typical circumstances and during normal implementation. The importance of evaluating feasibility within education research was noted in 2013 within the *Common Guidelines for Education Research and Development*, developed via a joint effort of the National Science Foundation and the Institute of Educational Sciences, part of the U.S. Department of Education. The remainder of this entry explains the value of feasibility studies, looks at the difference between a feasibility study and a pilot study, and briefly reviews how feasibility data are collected.

Feasibility studies are used formatively to estimate important parameters needed to design a full-scale trial and to reduce threats to the validity of a study's outcomes. By conducting feasibility studies, a researcher is able to determine the appropriateness of further evaluation, given practical considerations related to (a) process, (b) resources, (c) management, and (d) scientific basis for a planned trial. As part of a feasibility study, a small-scale or pilot test may be implemented to provide initial evidence that assesses and/or compares several study components such as capacity, participant recruitment and retention strategies, or initial data trends. Additionally, a feasibility study may help the researcher identify necessary modifications to the intervention and study procedures and protocols. This information is particularly critical in cases in procedures and protocols. This information is particularly critical in cases in which there are unique attributes of or little previous research with the selected participants, within the setting, or that employs a specific procedures. Additionally, when researchers involve multiple agencies or groups of people in a study, feasibility data may assist in intervention coordination.

The terms *feasibility* and *pilot* are often used interchangeably. However, the National Institute for Health Research Evaluation, Trials and Studies Coordinating Center identifies feasibility studies as those that look at specific design aspects of the proposed full study, whereas pilot studies test whether the procedures of the full study are effective to produce unbiased investigation. Thus, randomization is not necessary in a feasibility study but would be required in a pilot for a randomized trial.

When conducted as one aspect within a study, feasibility is often measured retrospectively. The inclusion of feasibility as an aspect of evaluation in a larger trial may be used as evidence of necessary adaptations needed to produce effect within a new context or with a different population. Whether collected as part of a feasibility study or as part of a larger trial, feasibility data are typically collected from those implementing or overseeing the study and may include surveys, interviews, focus groups, cost analysis, direct observation, checklists, and self-reports. In the case of pilot studies conducted within feasibility-process evaluations, data may also be collected via small-scale studies that evaluate the effectiveness of the intervention and implementation fidelity.

Joseph Calvin Gagnon and Brian R. Barber

See also Evaluation Versus Research; National Science Foundation; Pilot Studies

Further Readings

Bowen, D. J., Kreuter, M., Spring, B., Cofta-Woerpel, L., Linnan, L., Weiner, & Fernandez, M. (2009). How we design feasibility studies. American Journal of Preventive Medicine, 36, 452–457.

Institute of Educational Sciences, Department of Education and the National Science Foundation. (2013, August). Common guidelines for education research and development. Washington, DC: Authors.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., & Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. BMC Medical Research Methodology, 10, Article 1. doi:10.1186/1471-2288-10-1

Julie Slayton Julie Slayton Slayton, Julie

Federally Sponsored Research and Programs Federally sponsored research and programs

668

671

Federally Sponsored Research and Programs

The federal government sponsors a significant amount of research that takes place in the United States, though the vast majority of those funds are directed toward research conducted at the university level. Funding for both research and programs comes from a range of different agencies within the federal government and can be obtained as grants, contracts, and programs. The federal government has been funding research since 1953, and though it has not risen every year, and has, at times, experienced slight declines, it has increased significantly since the federal government began funding research. As of 2013, federal funding for national research and development made up approximately one fourth of all funding available. In terms of university research and development funding, federal funding since 1990 has counted for the vast majority of funding provided. In fact, it accounts for well over half of all research and development funding received by universities. In terms of the distribution of federal research funds, by far, the largest amount of money is consistently spent on National Institutes of Health biomedical research, far exceeding the amount spent on the next subject, engineering.

Those seeking funding from the federal government need to be knowledgeable about how to discover and pursue the available funding. Some might say that seeking federal funds for research is not for the faint of heart. There are very few simple ways to discover what funding is available, the grant writing and review process can be time-consuming and arduous, and the chances of receiving funding are small. Moreover, one of the best ways to ensure that a grant is received is to have previously been the recipient of a major award from the federal government, and in some cases, specifically the agency from which one is seeking funding. All these realities are important to understand when deciding to seek federal funding. On the other hand, receiving a federal grant is essential to seek federal funding. On the other hand, receiving a federal grant is essential for the vast majority of researchers in the higher educational community, as these grants support work that allow researchers to contribute to the intellectual community and advance in their careers.

This entry provides an overview of federal sponsors of research and then provides insight into some of the challenges associated with seeking federal grants and contracts as well as a list of resources related to federal agencies that sponsor federal research and programs.

Federal Sponsors of Research

The federal government has 26 separate agencies that offer grant funding opportunities to researchers. These agencies range from the Agency for International Development to the Department of Justice to the Nuclear Regulatory Commission. Each of these agencies offers granting opportunities, some of them offering as few as one (e.g., Corporation for National and Community Service) and others with as many as 1,341 (e.g., the Department of Health and Human Services). Each grant is posted and disseminated separately with its own set of requirements, due dates, and funding terms. It is important for any researcher interested in federal funding opportunities to become familiar with the grants.gov website and the idiosyncrasies of each specific funding agency. In some cases, it is also important to become familiar with specific grant cycles, as some grants are issued on a yearly basis and the terms of the grant do not change or only change slightly from year to year. It is also important that one becomes familiar with the agencies that are most likely to fund research in one’s area but to not focus so exclusively on those agencies that one believes will be the “best funding source” because other agencies might put out a request for proposal (RFP) that aligns well with one’s research interest.

In addition to becoming familiar with the different agencies, one should become familiar with the different categories of research as another way to discover funding opportunities. In addition to identifying grants through agencies, one can explore grant opportunities by category. For example, while the Department of Education may only indicate that it has four RFPs available, a search by category would reveal that 492 grant opportunities are available across a range of agencies in the area of education. Moreover, someone in search of federal funding should be aware of the range of eligibility categories; there are 16 different eligibility categories. Grants may specifically target a given recipient type. These range from cities or township governments to Native American tribal governments to public-and state-controlled institutions of higher education. Thus, some grants are specifically intended for individuals or groups that the federal government has determined would particularly benefit from the receipt of federal funding for research.

The federal government does provide a number of tools for grant seekers to help them navigate the process. Those interested in funding may sign up to receive e-mail notifications about new funding opportunities. These e-mails inform the recipient of upcoming RFPs, changes to RFPs, “dear colleague” letters, and recent findings or results of research sponsored by the federal agency. The grants.gov website also provides training documents and videos, and individual agencies offer webinars for most if not all of the RFPs they post. In addition, specific agencies provide other resources to support grant seekers. For example, the National Science Foundation publishes a Proposal and Award Policies and Procedures Guide that details the specific requirements for those interested in seeking National Science Foundation funding. These guides are issued periodically, and therefore, anyone interested in abiding by the most current expectations should be on the lookout for updates to these materials, which may change on a yearly basis. In addition to these resources, each RFP provides a significant amount of detail regarding the specific requirements of the grant, including the number of awards anticipated, the format expectations for the proposal submission, and the name of a contact at the funding agency, and questions can be directed to those program officers.

Thus, those interested in seeking federal grants or contracts or participating in federally funded programs have many resources they might access to identify and target funds. As the next section suggests, having the information available does not guarantee that receiving funds is an easy endeavor.

Challenges of Federally Sponsored Research

Although federal funding is a powerful tool, the road to obtaining that funding is filled with challenges and littered with the proposals of those who have been unsuccessful in the effort. First, discovering exactly which of the thousands of grants might best match one’s research interest is a time-consuming exercise. The onus is often on the researcher to become familiar with each funding agency and the specific types of grants it offers. Moreover, aligning one’s interest to the call for proposal can be very challenging. Many RFPs are directed at very specific audiences, excluding a large percentage of those who might be seeking funding. For example, many of the RFPs issued by the National Institutes of Health or the National Science Foundation require that the researcher employs a randomized control trial and demonstrates alignment with the federal What Works Clearinghouse criteria for research. Although these are high standards, they are standards that have been criticized by many social scientists as unreasonable and inappropriate for much of the social science research that is conducted and are considered to be unnecessary constraints to the production of high-quality research that contributes to the generalizable knowledge base. Similarly, many of the program officers responsible for guiding the grant review process have very specific ideas about the types of projects that should be funded. If the researcher has not made an effort to investigate whether the researcher’s project is one that would be interesting to the program officer, it likely limits the researcher’s opportunity to receive funding.

In addition, the number of awards is often very small, making the likelihood of success equally small. In some cases, funding opportunities are offered and then withdrawn based on a lack of available funds from the federal government. Other challenges that exist when it comes to obtaining federal funding include the size of the awards that are offered and the extent to which those awards truly cover the costs of the research intended to be undertaken. Federal grants are awarded for a period of 1 to 5 years, each year requiring a certain proportion of the effort to be undertaken during that time. The current federal rate for research conducted on campus is 65%. This means that up to 65% of an award is directed toward university overhead and the actual amount of money available for the research must come from the remaining 35% of the funds provided through the grant. Thus, what on the surface may appear to be a substantial award quickly dwindles to a much more modest number. This fact creates tension in the grant writing process, as the researcher must demonstrate that the research scope requires the full amount of the award, when it is highly unlikely that the full award will be available to cover the costs of the actual research. Another challenge to receiving federal funds lies in the expectation that those who are seeking funds have rich and long track records with federally funded research efforts. This is understandable, as the federal government would like to have confidence that its money is being well spent. On the other hand, this fact also makes it much harder for those who are looking to enter the field and build their reputations as researchers to make good use of federal funds.

Although these challenges may seem impossible to overcome, an individual who is willing to invest the time and energy associated with finding the right grant is willing to invest the time and energy associated with finding the right grant opportunity ensures that the individual has written a grant that aligns with the RFP and the formatting expectations (as a grant can be rejected without being reviewed simply if it does not abide by the formatting expectations), reaches out to the program officer to seek guidance in relation to the proposal, and invests in the long-term likelihood of receiving a grant may very well find that it is worth the effort.

Julie Slayton

See also Institute of Education Sciences; National Science Foundation; Research Proposals

Further Readings

Grants.gov. Retrieved from http://www.grants.gov/web/grants/home.html

National Science Foundation. (2015). *National patterns of R…D resources* series. Constant-dollar conversions based on GDP deflators from Budget of the U.S. Government FY 2016. Washington, DC: AAAS. Retrieved from http://www.aaas.org/sites/default/files/USPerf1.jpg

National Science Foundation. (2016). Proposal and award policies and procedures guide (NSF16-1). VA: Author.

National Science Foundation, National Center for Science and Engineering Statistics. (2015). Higher education R…D series, based on national survey data. Includes Recovery Act Funding. Washington, DC: AAAS. Retrieved from http://www.aaas.org/sites/default/files/UniSource1.jpg

National Science Foundation, National Center for Science and Engineering Statistics. (2016). Federal funds for R…D series, based on national survey data. Washington, DC: AAAS. Retrieved from http://www.aaas.org/sites/default/files/USFund1.jpg

Angelo S. DeNisi Angelo S. DeNisi DeNisi, Angelo S.

Feedback Intervention Theory Feedback intervention theory

Feedback Intervention Theory

Feedback intervention theory (FIT), first proposed by Avraham Kluger and Angelo DeNisi in 1996, attempts to explain why feedback is not always effective in improving subsequent performance. Because feedback is an important component of many educational programs and interventions, it is critical to understand more about how feedback actually affects subsequent behavior. This entry first explains the origin of FIT and then examines the arguments critical to understanding feedback effects as well as how the three views of the self affect feedback. Finally, implications of the theory are considered.

FIT grew out of the results of a meta-analysis of over 600 effect sizes dealing with the relationship between feedback and subsequent performance. Traditionally, feedback was viewed as being an effective tool for changing behavior, but the results of this meta-analysis indicated that, in almost one third of the cases, feedback had a negative effect on subsequent performance. That is, individuals receiving performance feedback did more poorly on subsequent tasks than did individuals who received no feedback. Furthermore, these results were independent of the sign of the feedback; positive feedback had the same type of effect as did negative feedback. Interestingly, the review and meta-analysis revealed that there had always been evidence that feedback had such mixed effects, but that inconsistencies with the way some past reviews were conducted had obscured this fact. These results, and their implications, necessitated a proposed theory to explain when feedback would likely have the positive effects usually associated with it and when those effects might be negative, and this was termed feedback intervention theory.

The development of FIT begins with noting that the usual assumptions underlying studies of feedback effectiveness (i.e., that behavior was regulated by attempts to reduce the discrepancy between feedback and standards for attempts to reduce the discrepancy between feedback and standards for performance) were too simplistic, on their own, to fully explain feedback effects. Instead, building upon control theory, FIT argued that, in addition to the usual standards-discrepancy arguments, understanding feedback effects required two other arguments. The first was that not all feedback–standards gaps could receive attention and only those that did receive attention could be related to behavior regulation. The second, and more critical argument, was that feedback interventions changed the locus of attention and therefore affected behavior.

This second argument was central to understanding the inconsistent effects of feedback on performance. Attention can be directed to the self, to the task at hand, or to the details of the task at hand. Feedback interventions that direct attention to the task at hand are likely to produce positive effects, regardless of the sign of the feedback, just as has always been suggested. Feedback that directs attention to the details of the task at hand can have positive effects, but there is also a potential problem of feedback recipients focusing too much on details. Feedback that directs attention to the self, however, is the most problematic, as it diverts cognitive resources away from task performance and produces affective reactions that can interfere with task performance. Partial tests of FIT found that feedback interventions containing both praise and criticism produce lower positive and some negative effects on performance when they direct attention to the self.

But subsequent statements of FIT also drew upon distinctions among the three views of self: the actual, the ideal, and the “ought” selves. We view our actual self as what we believe ourselves to be, our ideal is what we wish to be, and our ought selves are what we should be. This work suggests that information about discrepancies from the ideal self focuses on promotion goals (possible gains) and focuses our efforts on trying to achieve that ideal. On the other hand, information about discrepancies from our ought self focuses our efforts on prevention focus (possible losses) and leads us to try to achieve socially prescribed standards. In this latter case, feedback will tend to push us toward the standard so that, when we receive superior feedback, this will be followed by performance decline, and when we receive poor feedback, this will push us toward performance increase. Thus, feedback interventions that direct out attention to the self will have much different effects on performance, depending upon whether they direct our attention to our ideal self or to our ought self.

Thus, FIT was proposed as a way to help explain the somewhat surprising inconsistency in the effects of feedback on performance. Therefore, there was more emphasis on the reasons why feedback might have a negative effect (not “no effect”) on subsequent performance and less emphasis on aspects of feedback interventions that would make it more likely that the intervention worked as intended. Nonetheless, FIT does include potential guidance on how feedback should be delivered to help ensure it has the desired positive effect on performance. Recommendations included the notion that any feedback intervention should be accompanied with some type of goal-setting program. It was also recommended that, where possible, feedback should include information about the “correct solution” or how to improve performance. There was also evidence that feedback that was given more frequently, and that showed changes from previous trials, should also enhance the effectiveness of feedback. Finally, there was some evidence that computer-generated feedback was more effective and that feedback was more effective with simple rather than complex tasks.

The greatest implication of FIT for research and practice in fields such as education is that researchers cannot simply assume that feedback is effective, but that they should actually test whether or not it is. Also, there are some basic considerations that can help ensure that feedback works as it was intended, and these should be considered in the design of any intervention related to feedback.

Angelo S. DeNisi

See also Active Learning; Behaviorism; Goals and Objectives; Learning Theories; Punishment; Reinforcement

Further Readings

Carver, C. S., & Scheier, M. F. (1981). Attention and self-regulation: A control theory of human behavior. New York, NY: Springer-Verlag.

Higgins, E. T. (1997). Beyond pleasure and pain. American Psychologist, 52, 1280–1300.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology, 64, 349–361.

Kluger, A. N., & DeNisi, A. S. (1996). The effects of feedback interventions on performance: Historical review, meta-analysis, a preliminary feedback intervention theory. Psychological Bulletin, 119, 254–284.

Kluger, A. N., & DeNisi, A. S. (1998). Feedback interventions: Toward the understanding of a double-edge sword. Current Directions in Psychological Science, 7, 67–72.

Sharon Brisolara Sharon Brisolara Brisolara, Sharon

Feminist Evaluation

Feminist evaluation

672

676

Feminist Evaluation

Feminist evaluation is an approach to program evaluation that emerged as a distinct model at the end of the 1990s. Feminist evaluation draws from rich literatures of feminist theory and feminist research methods in the social sciences, philosophy of science, liberal arts, and natural sciences. An initial impetus behind the articulation of a feminist evaluation model was a recognition of the negative consequences of lack of attention to gender and gender inequities in conceptualizing, designing, conducting, and analyzing data from program evaluations. These gender inequities that result in social injustice are placed at the center of the feminist evaluation model. The model development grew to encompass principles about evaluation, knowledge, discrimination, and ways of knowing. Feminist evaluation is different from, but shares values with, other evaluation models; the controversies that have arisen over the model have tended to focus on use of the term *feminist*, the range of possible roles for the evaluator, and whether one has to identify ideologically as a feminist to conduct a feminist evaluation. As feminist theory and feminism have garnered the attention of a new generation, the term *feminism* and its aims are being reformed, resulting in the continued development of feminist evaluation as a model. This entry examines the origins of the approach, how feminist evaluation is situated within the greater landscape of evaluation models, eight feminist evaluation principles, and the dimensions and controversies of feminist evaluation.

The Origins of Feminist Evaluation

Feminist evaluation draws from rich, engaged philosophical, methodological, and epistemological literatures emerging from diverse disciplines. Contributions to theory and methods have come in what are sometimes described as waves of feminism; within the world of social science research, such trends are further understood within schools of thought. Each of these schools have critiqued the dominant research and methodological paradigms of its periods, posing challenges to epistemology (the nature and scope of knowledge), methodology, and ontology (the nature of being, reality, or existence).

Feminist empiricism was an early form of feminist theory that adhered to many of the tenets of positivism but critiqued androcentric perspectives that led to biased results. Later, standpoint theory critiqued positivism and encouraged the use of multiple standpoints as critical to gendered insights. Critical theory focused attention on power and domination; feminist postmodern and poststructural theories urged reconsideration of the possibility of objectivity. Global and postcolonial theories contributed an investigation into Western assumptions and the effects of the colonial past. Queer and lesbian, Black feminist, Chicana, indigenous, and race-focused theories examined, in different ways, biases, privilege, multifaceted identities, and how to engage adherents in discourses surrounding action against oppression. Each school drew from rich bodies of work, each significantly influenced multiple fields, and each has had its own critics.

Feminist Evaluation as Evaluation Model

Program evaluation is a form of applied research used to examine the merit, worth, value, or state of development of social interventions (programs). Program evaluation is transdisciplinary and has been influenced by diverse fields including educational psychology, sociology, statistics, and anthropology. In the 1960s, the evaluation profession began to develop in the United States and the diversity of disciplines from which practitioners hailed led to a number of different evaluation models or approaches to evaluation. The term *model* implies preferred approaches to data collection, ideas about what constitutes knowledge, concepts regarding what constitutes credible evidence, and perspectives about what can be known. Models also provide explicit or implicit perspectives on the relationship of the evaluator to stakeholders, the appropriate role of the evaluator, and the use of findings. Models emerge from practice, experience, and theoretical developments and are often created in response to perceived socioeconomic needs or dynamics.

In the early decades of the program evaluation profession, it was commonly thought that practitioners should strive for consistency in paradigmatic positions implicit within a model and the methodological choices related to these. As the field developed, practitioners and theorists began to value what could be achieved by bringing different perspectives to bear upon the “object” of evaluation. For example, the use of mixed methods (qualitative and quantitative) stopped being the subject of heated debate and became good evaluation practice. Likewise, the field demonstrated greater acceptance of “mixing paradigms” and a recognition that using more than one model can aid in understanding diverse program aspects and dynamics.

Another contributing factor to the proliferation of evaluation models has been the growth of the field and an expansion of the need for and uses of evaluation in new settings. Evaluation practitioners now become skilled in a range of models and select the models that will guide their work based on the social, political, and organizational context of the programs themselves and their assessment of stakeholders’ needs. Feminist evaluation shares an affinity with many categories of evaluation models, including stakeholder-based evaluation models that attend to and address key stakeholder values, democratic evaluation models that value pluralism, and attend to power relationships and explicitly collaborative models including participatory evaluation, empowerment evaluation, and transformative evaluation. Like many of these models, feminist evaluation has made unique contributions to the field. First, it acknowledges and examines of the structural nature of inequities, beginning with gender as a point of departure. Second, it offers guidance in examining multiple and intersecting identities that include, but are not limited to, sex, race, class, and ability. Feminist evaluation has also contributed to critical conversations surrounding the evaluator’s role in addressing the social justice aims of interventions by offering examples of engagement and making a case for action as integral to the aims of the evaluative enterprise, given the violence and poverty that are often consequences of significant gender inequities.

Feminist Evaluation Principles

The feminist evaluation model is based on eight principles that have evolved over time. These include concepts related to the nature of knowledge, the nature of inquiry, and social justice.

1. *Knowledge is culturally, socially, and temporally contingent*. For feminist

evaluators, knowledge is deeply connected to a particular time, place, and social context; it is incumbent upon practitioners to recognize how this knowledge is situated. This both limits the evaluator's claims to generalizability and increases the evaluator's attention to specific social contexts in order to better understand factors shaping actors, relationships, and situations.

2. *Knowledge is a powerful resource that serves an explicit or implicit purpose*. Knowledge that is shared during or as a result of the evaluation is a significant resource. Those collecting data are gatekeepers of what is learned, who gets credit for knowledge, and information use. Feminist evaluators contend that this power brings responsibility.

3. *Evaluation is a political activity; practitioners' personal experiences, perspectives, and characteristics come from and lead to particular political stances*. The contexts in which evaluations operate are politicized and imbued with asymmetrical power relationships that influence everything from the funding of programs to daily decision making. Methods of inquiry are imbued with biases, reflecting the dominant ideologies within which they were created. The evaluation itself and the choices that lead to evaluations are political activities.

4. *Research methods, institutions, and practices are social constructs*. As social constructs, research and evaluation methods, institutions, and practices are products of their culture and time, including the dominant ideologies, theories, academic traditions, and perspectives that shape the world of inquiry.

5. *There are multiple ways of knowing*. Feminist theory suggests that particular ways of knowing, such as logic, are privileged over others by those with the power to sanction or privilege ways of knowing. Feminist evaluation advocates the use of diverse ways of knowing, such as the use of intuition, emotions, and love and other sources of insight into problems or dynamics.

6. *Gender inequity is one manifestation of social injustice. Discrimination cuts across race, class, and culture and is inextricably linked to all three*. An awareness of and attention to gender inequities is a point of departure for more deeply understanding the multiple effects of discrimination and existing power dynamics.

7. *Discrimination based on gender is systemic and structural*. Discrimination

based on gender (like other forms of discrimination) is perpetuated through social norms that shape and restrict possibilities through the policies, practices, and structures of social institutions. Discriminatory practices are so embedded within structures and systems that they are not easily recognized. Structural and systemic problems require structural and systemic solutions.

8. *Action and advocacy are considered to be morally and ethically appropriate responses of an engaged feminist evaluator*. Action and advocacy can take many forms, from strategic dissemination of findings to engaging in activities aimed at altering the balance of power. Decisions on the appropriate level of advocacy and action must occur in relationship with stakeholders.

Dimensions and Controversies

By the time a full-length volume on feminist evaluation was published, the field had already seen a number of feminist evaluation articles in a range of professional journals and new handbooks on feminist methods and gender responsive evaluations. However, approaches that challenge and critique dominant paradigms are typically accompanied by controversy. The three most common controversies related to feminist evaluation are the use of the term *feminist*, what constitutes acceptable evaluator roles, and whether one has to identify ideologically as a feminist to conduct feminist evaluation.

Some advocates for feminist evaluation point to the long struggle for gender equity embedded within and illuminated through decades of feminist theory as best being represented through the clear use of the term *feminist*. These advocates claim that to the extent that the model being used is feminist in nature, describing it as such reflects the transparency and honesty desired if more equitable balances of power are to be achieved. Those opposed to the use of the term *feminist* claim that some may believe in feminist evaluation principles and yet not feel comfortable being associated with the term because of its Western or political history. A common position is the concern that use of the term *feminist* will inhibit funders or more conservative stakeholders from accepting the evaluation approach or that discomfort with the word itself may detract from program or evaluation progress.

Discussions about the appropriate role of the evaluator exist and are not confined to the feminist evaluation model. The evaluation field allows for a greater diversity of roles than was true in the early days of the profession. What tends to be contentious with respect to the feminist evaluator role is discussion over the degree of advocacy or social action appropriate for the evaluation practitioner. Some view action on issues of social inequity as a moral and ethical imperative given the evaluator’s position of power and knowledge of a given situation. Others view advocacy as incompatible with the role of an evaluation practitioner and claim that any action is most appropriately addressed after the evaluation has been completed.

A third debate occurs over who can be a feminist evaluator. This debate is not focused on biological sex or gender identity but on the degree to which one identifies, publicly or even privately, as a feminist. Some practitioners state that agreeing with the basic principles of feminist evaluation and building skills in applying a feminist lens to program and evaluation design is criteria enough. Others counter that without identifying as a feminist and immersing oneself in feminist literature, applying a feminist lens is difficult if not impossible.

Feminist evaluation garnered increased attention in past years, concurrent with the proliferation of gender equity and gender responsive approaches to monitoring and evaluation within the field of international development. Feminist evaluation has tended to differ from gender responsive approaches in that it examines and seeks to address the social dynamics leading to gender equity and oppression; it challenges the factors that keep such dynamics in place rather than recording these as context. The need for such approaches has been heightened not only because of a recognition of the cost of insufficient attention to gendered dimensions of development projects, but because of the gendered assessment and design needs generated by the United Nation’s Millennium Development Goals. Feminist evaluation will continue to be shaped by more nuanced understandings of gender identity and expression; new scholarship in the fields of feminism, gender, and sexuality; and unforeseen social needs that would benefit from investigation through a feminist lens.

Sharon Brisolara

See also Ethical Issues in Evaluation; Evaluation, History of; Gender and Testing; Participatory Evaluation; Program Evaluation; Transformative Paradigm

Further Readings

Brisolara, S., & Seigart, D. (2012). Feminist evaluation research. In S. N. Hesse-Biber (Ed.), Handbook of feminist research: Theory and praxis (2nd ed., pp. 135–153). Los Angeles, CA: Sage.

Brisolara, S., Sengupta, S., & Seigart, D. (Eds.). (2014). Feminist evaluation and research: Theory and practice. New York, NY: The Guilford Press.

Engendering Policy through Evaluation. Retrieved from www.feministevaluation.org

Hesse-Biber, S. N. (Ed.). (2012). Handbook of feminist research: Theory and praxis (2nd ed.). Los Angeles, CA: Sage.

Hesse-Biber, S., & Leavy, P. L. (Eds.). (2007). Feminist research practice, a primer. Thousand Oaks, CA: Sage.

Mertens, D. (2009). Transformative evaluation and research. New York, NY: Guilford Press.

Seigart, D., & Brisolara, S. (Eds.). (2002). Feminist evaluation: Explorations and experiences. New Directions in Evaluation, (96), 1–114.

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

Field Notes

Field notes

676

677

Field Notes

Field notes, which are based on observation in one's research setting, allow researchers to see and record, firsthand, the activities in which research participants are engaged in the contexts of these activities. Observation is often used as a method of data triangulation—meaning the use of multiple data sources to achieve a range of contextual data—because the validity of self-reporting (such as in interviews and focus groups) often comes into question; therefore, observational field notes—how observations become data—validate information garnered from focus groups, interviews, questionnaires, and other methods of data collection. This entry examines three basic types of field notes, the importance of and skills required for recording field notes, and the sequential process of writing field notes.

Depending on a study's specific methodological frame and how it approaches field notes as part of a broader data set, field notes generated by observation can be descriptive, inferential, and/or evaluative. In descriptive field notes, researchers observe and describe what has been observed as neutrally as possible. This can be confusing (and misleading) because often what seems objective is in fact inference, so researchers must pay attention to their interpretive filters. Inferential field notes require that researchers understand that they are making inferences—interpretations and assumptions that extend beyond the data—about what is observed and the underlying motives, affect, and/or emotions of the events and behaviors observed. Evaluative field notes mean that researchers are consciously making inferences and judgments about the nature and motives of the behaviors or events observed. Understanding these various approaches to field notes, with a focus on the goals, roles, and differences between them, is vital. Broadly, field notes include descriptive as well as between them, is vital. Broadly, field notes include descriptive as well as inferential data. Although it is important to acknowledge the differences between these, the lines can be blurry, especially between description and inference. This makes systematic and structured reflexive engagement with field note data crucial for validity.

It is vital to understand that without recording observation through writing, there are no data. This is why observation and field notes are considered to be one method because it is essential to record observations through the careful and systematic process of writing field notes. There are many approaches to field note writing, with varying reasons and processes that relate to each choice. As Robert Emerson, Rachel Fretz, and Linda Shaw make clear in their book *Writing Ethnographic Fieldnotes* (2011), writing field notes requires that a researcher develops specific skills, including the following:

1. moving from theory (or a problem statement) to what is in focus/observed;
2. understanding the theoretical construction of the study’s focus and guiding questions;
3. learning to engage in a disciplined way to what one sees and hears and taking detailed notes while in the setting;
4. capturing social interactions in words (i.e., observing and writing about the order or sequences of action);
5. learning to write an analysis that is conscious of stylistic and representational choices; and
6. seeking the perspectives, language, and indigenous concepts of insiders in the setting.

There is a sequential process of writing field notes, starting with in-the-field “jottings,” which are contemporaneously written while at the research site. These jottings are turned into broader, more coherent written accounts of what is observed as the researcher turns them from jottings into field notes after leaving the field. The “real-time jottings” are an essential grounding and resource for writing the fuller field notes, which should be written shortly after leaving the field so that they are written close to the time of actual observations and are therefore more reliable.

Separating the real-time experience of the observation and jottings that happen in the field from the written field notes about this observation, as if they are objective and separate, confuses the meaning and goals of field “data” because it treats these data as objective information rather than as interpretive and specific to observer subjectivities. Situating the observational field notes within the subjective researcher interpretation and yet seeking to keep the notes as close to the events as possible is the goal of field notes.

Sharon M. Ravitch

See also Ethnography; Hawthorne Effect; Qualitative Data Analysis; Qualitative Research Methods; Reliability; Triangulation; Trustworthiness; Validity

Further Readings

Angrosino, M. (2007). Doing ethnographic and observational research: The Sage qualitative research kit. Los Angeles, CA: Sage.

Bernard, H. R., & Ryan, G. W. (2010). Analyzing qualitative data: Systematic approaches. Thousand Oaks, CA: Sage.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). Writing ethnographic fieldnotes (2nd ed.). Chicago, IL: University of Chicago Press.

Guest, G., Namey, E. E., & Mitchell, M. L. (2013). Collecting qualitative data: A field manual for applied research. Los Angeles, CA: Sage.

Hammersley, M., & Atkinson, P. (2007). Ethnography: Principles in practice (3rd ed.). Hoboken, NJ: Taylor … Francis.

Ravitch, S. M., & Carl, N. M. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: Sage.

Frank A. Bosco Frank A. Bosco Bosco, Frank A. Jr.

File Drawer Problem File drawer problem

File Drawer Problem

The *file drawer* is a metaphorical term referring to a storage location for nonpublished research. The *file drawer problem*, a term coined by Robert Rosenthal, refers to the possibility that nonpublished results differ systematically from published results. Systematic differences between published and nonpublished research are especially problematic for the field of education, where summaries of research through meta-analysis are increasingly relied upon to inform practice. This entry describes the nature, causes, and consequences of the file drawer problem as well as the methods for its detection and eradication.

Of all studies conducted by researchers, some become published and easily accessible to consumers. Other studies are said to be relegated to the file drawer. The file drawer problem is one type of publication bias, a broader phenomenon whereby published research is a nonrepresentative sample of all research. Reasons for a research manuscript not being accepted for publication are often linked to reviewer or editorial bias against null or nonsignificant results during the peer review process.

Consequences of publication bias became salient with the rise of meta-analysis. Indeed, meta-analytic inferences rest on the assumption that the included studies constitute an unbiased sample. In the modal case, the concern is with upward bias; that is, some studies with small or null effects are missing from the summary, resulting in unrealistically high meta-analytic estimates. As a second consequence, an unrepresentative sample of published research can provide unrealistically low estimates of the reproducibility of scientific research. Indeed, if only the “best-looking” findings are selected for publication, then replication attempts are increasingly likely to fail. Finally, publication bias has the potential to stymie attempts at evidence-based practice. Indeed, failures of evidence application should increase with the level of bias associated with the evidence application should increase with the level of bias associated with the evidence.

Meta-analysts have developed several techniques for detecting and correcting for the impact of publication bias. Indeed, Rosenthal’s seminal approach provides an estimate of the number of file drawer studies with null results that would need to exist in order to affect one’s meta-analytic conclusions. Newer approaches provide revised meta-analytic estimates after imputing studies assumed to be contained in the file drawer or by making other modifications to the distribution of effects.

To completely eradicate publication bias would be preferable to improving its detection or assessment. Given that one culprit for the existence of publication bias is found in the journal editorial process, several journals have adopted modified peer review processes, wherein authors first submit manuscripts *without* findings and conclusions. Then, after peer review for rigor and relevance has been completed, the results and discussion sections are submitted. As another culprit for publication bias, authors might simply abandon research projects without having submitted them for publication. There are now several mechanisms available to reduce this concern, such as those provided by the Center for Open Science, that allow researchers to upload and make available research data, manuscripts, and the like.

Frank A. Bosco Jr.

See also Effect Size; Meta-Analysis; Missing Data Analysis; Quantitative Research Methods; Threats to Research Validity; Type I Error

Further Readings

Kepes, S., Banks, G. C., McDaniel, M. A., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. Organizational Research Methods, 15, 624–662. doi:10.1177/1094428112452760

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? Industrial and Organizational Psychology: Perspectives on Science and Practice, 6, 252–268. doi:10.1111/iops.12045

Gail Tiemann Gail Tiemann Tiemann, Gail

Fill-in-the-Blank Items Fill-in-the-blank items

Fill-in-the-Blank Items

Fill-in-the-blank items are assessment questions in which test takers must hand enter a response, rather than select from a list of predetermined answer choices. Generally, a sentence or paragraph is presented to the examinee with key components replaced by blank spaces where the student must fill in the response that completes the sentence. Fill-in-the-blank items are also known as cloze items or completion questions. With this item type, correct answers are generally limited in length such as a single number, word, or phrase. After an example of a basic fill-in-the-blank item format, the rest of this entry examines the potential benefits and drawbacks of utilizing this specific assessment question, reviews a variation of the item type, and highlights best practices for constructing fill-in-the-blank items.

The following is an example of a fill-in-the-blank item:

> Stem: “Four score and seven years ago” is the famous opening phrase of Lincoln’s battlefield address given in the Pennsylvania town of _______.
>
> Answer: Gettysburg

Fill-in-the-blank items may offer more authentic assessment of a domain or construct than traditional multiple-choice items because test takers must construct an answer, rather than choose from or possibly guess from a finite list of possible answer choices. Fill-in-the-bank items of this type may also be easier to construct than multiple-choice questions, as the question writer does not need to create adequate distractors or incorrect answer choices. However, they may be more time-consuming to score because fill-in-the-blank items are often scored by hand using a scoring guide or answer key. However, many computer-based testing software applications can score this item type automatically, especially when the answer is straightforward and clear and if all the acceptable answer formats are fully specified in the software. Another potential drawback to fill-in-the-blank items is they may take more time for students to answer than selected-response item types. Additionally, fill-in-the blank items may be best for assessing lower order skills such as recall because student answers are usually short.

A fairly common format for fill-in-the-blank items, especially in elementary and middle school classrooms, combines that approach with a matching format in which the student may choose from a list of options or a “word bank.” Following is an example of that format:

> Question: “Four score and seven years ago” is the famous opening phrase of Lincoln’s battlefield address given in the Pennsylvania town of ________.
>
> Answer options: Appomattox
>
> Gettysburg
>
> Philadelphia
>
> Pittsburgh

Fill-in-the-blank items with one acceptable answer are best. One correct answer chosen a priori increases validity, makes the question fairer to students, and tremendously reduces the subjectivity required for a teacher to score the answer. Reduced subjectivity increases the reliability of fill-in-the-blank items. Another best practice has to do with the blank itself. There should be just one; questions with many blanks are called swiss cheese items. Furthermore, the single blank should go at the end of the statement. This allows for more efficient “search strategies,” as students search their knowledge base for the right answer.

Gail Tiemann

See also Essay Items; Matching Items; True–False Items

Further Readings

Further Readings

Cross, K. P., & Angelo, T. A. (1988). Classroom assessment techniques. In A handbook for faculty.

Frey, B. B. (2013). Modern classroom assessment. Thousand Oaks, CA: Sage.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. Teaching and Teacher Education, 21(4), 357–364.

Olga Korosteleva Olga Korosteleva Korosteleva, Olga

Fisher Exact Test

Fisher exact test

679

682

Fisher Exact Test

The Fisher exact test is an inferential statistical procedure to compare the number of people or things falling into different categories. It is applicable in two situations. The first one is when a sample is drawn from a population and two categorical variables are recorded for each element in the sample, for example, political affiliation (Democrat/Republican/Other) and potential vote on a certain proposal (in favor/opposed/abstained). In this case, researchers would be testing whether there is an association between these two variables (or, putting it more rigorously, whether the two variables are independent). The second situation arises when two or more samples are drawn from independent populations and measurements for one categorical variable are recorded for each sampled element. In this instance, the hypothesis of interest is whether proportions for each level of the categorical variable are equal across the samples. To illustrate, a sample of freshmen and a sample of seniors are drawn and students' employment status (unemployed/part-time/full-time) is recorded. Investigators would be interested in testing whether proportions of students in each category of the employment status differ between freshmen and seniors. After this entry further explores the fundamental attributes of the Fisher exact test, it examines how statistical hypotheses are formulated and the procedure for conducting the test. Next, examples of a Fisher exact test for independence and test for equality of proportions are provided. Finally, limitations of the Fisher exact test are discussed.

In preparation for conducting the Fisher exact test, observations are arranged in an *r* by *c* table called a *contingency table*. It may also be called a *two-way table* or *cross tabulation* or, simply, *cross tab*. In the former case, when a single sample is drawn from one population and two categorical variables with *r* and *c* levels, respectively, are observed for each unit in the sample, the *r* rows of the contingency table correspond to the levels of the first variable, whereas the *c* columns represent the levels of the second variable. In the latter situation, when *r* samples are drawn from independent populations and a categorical variable with *c* levels is observed for each sample element, in the contingency table, the *r* rows represent the samples, and the *c* columns contain frequencies of the *c* levels of the observed variable.

Each cell in the contingency table contains the frequency of observations in the corresponding level–level combination of the two observed variables (in the former situation) and in the corresponding sample at the certain level of the observed variable (in the latter situation). These frequencies are commonly referred to as *observed counts*.

In order to prepare the data for analysis, the *marginal totals* must be computed and added to the contingency table. They are defined as the total for each row and column. The row totals are put in an additional column on the right of the table, whereas the column totals go into the row added to the bottom of the table. As the name suggests, these totals are placed on the “margins” of the table. Next, the *grand total* is calculated and written below the column with row totals (or to the right of the row of column totals). The grand total is defined as the sum of all observed counts. It is also equal to the sum of row totals and, likewise, to the sum of column totals.

To validate the use of the Fisher exact test, one has to compute *expected counts* for each cell of the contingency table, which is defined as the product of the corresponding marginal totals divided by the grand total. If at least one expected count is below 5, then the Fisher exact test may be carried out. If all expected cell counts are larger than 5, then the chi-square test is applicable.

The Fisher exact test is, in a sense, a nonparametric alternative to the chi-square test. The Fisher exact test belongs to the class of nonparametric tests, as it does not assume a known algebraic form of the underlying distribution(s) of the observed variable(s) and/or the test statistic.

A famous English statistician, Sir Ronald Aylmer Fisher (1890–1962), introduced this test. In his seminal book titled *Statistical Methods for Research Workers*, which was published by Oliver and Boyd in 1925, he considered the case of a 2 × 2 contingency table. Later, in 1951, G. H. Freeman and J. H. Halton extended it to a general *r* by *c* contingency table and published in the journal *Biometrika*.

Statistical Hypotheses

The testing procedures are identical in either of the two situations (one sample and two categorical variables, or several samples and one categorical variable). The difference is in how the statistical hypotheses are formulated.

For a single sample, the null hypothesis states that there is no association between the two measured variables (i.e., these two variables are independent). The alternative hypothesis is that the association exists (or, equivalently, the two variables are not independent). This test is referred to as the test for *independence of variables*.

In the case of several samples and one measured variable, the null hypothesis is that the proportions in each column are equal across the rows (which represent samples), and the alternative hypothesis is that proportions in each column are not all equal. It is said that researchers are testing for *equality of proportions* in this case. Note that the alternative hypothesis states that in each column, not all proportions are equal, which means that some proportions may be equal, but some are different, and it is not specified which ones.

Testing Procedure

Prior to conducting testing, an *r* by *c* contingency table with marginal totals and the grand total is prepared. Let the contingency table contain observed cell counts x_{ij} where $i = 1,\ldots, r$ and $j = 1,\ldots, c$. Denote by $x_{i.}$ the *i*th row total, by $x_{.j}$ the *j*th column total, and by $x_{..}$ the grand total. As the first step, the probability of the observed table is computed. The *probability of a table* is defined as

$$P_{\text{table}} = \frac{x_{1.}!\,x_{2.}!\ldots x_{r.}!\,x_{.1}!\,x_{.2}!\ldots x_{.c}!}{x_{..}!\prod_{i=1}^{r}\prod_{j=1}^{c} x_{ij}!},$$

that is, it is equal to the product of factorials of all row and column totals divided by the product of the factorial of the grand total and factorials of individual cell counts. This formula can be derived based on hypergeometric distribution.

Next, all possible tables with the same marginal totals are listed and the probability of observing each table is calculated. Finally, the *p* value for the Fisher exact test is found as the sum of all probabilities that are less than or equal to the probability of the observed table.

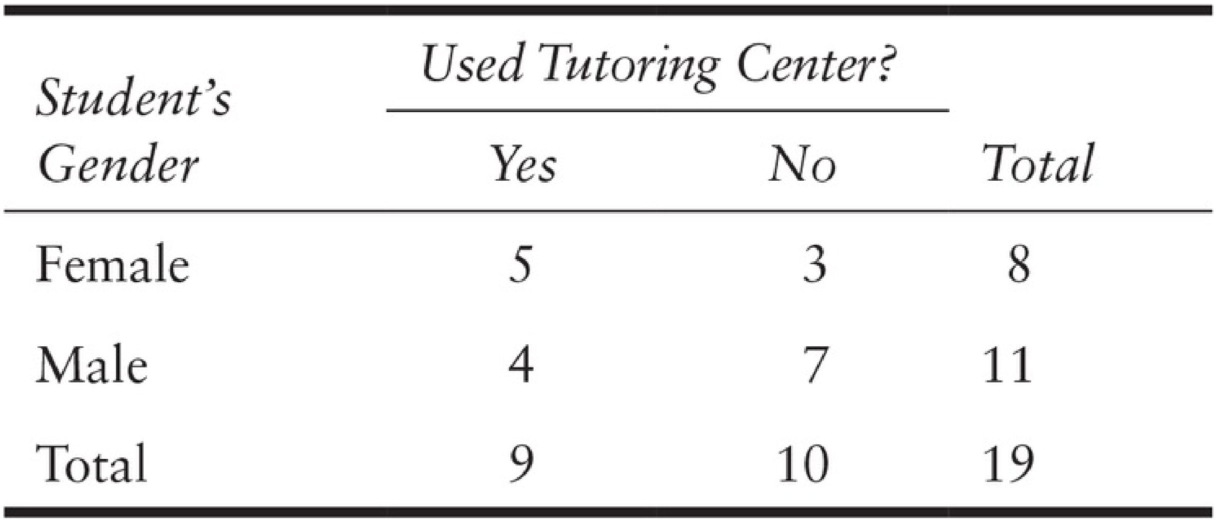
If the *p* value is in excess to a prefixed significance level α, it indicates that for the observed table, independence of variables (or equality of proportions, depending on the case) is likely to hold, and the null hypothesis cannot be rejected. On the other hand, if the *p* value is smaller than α, it signifies that the observed table is atypical under the null hypothesis (i.e., assuming that the null hypothesis is true), thus, the null hypothesis should be rejected.

It is noteworthy that the Fisher exact test is a nonparametric test based on permutations. It means that no assumption is made on the distribution of measured variables, and no test statistic is computed. Instead, all possibilities are enumerated and the *p* value is computed that reflects how unusual (unlikely) the data are if we assume that the null hypothesis is true.

As the reader might have guessed, enumerating all possibilities is generally a daunting task. That’s why statistical software packages are employed that output the *p* value of the test. Actual calculations are rarely carried out by hand, if only for illustrative purposes.

Example of Fisher Exact Test for Independence

For brevity, an example of a 2 × 2 contingency table is considered. Suppose a survey concerning the use of a tutoring center by students is conducted, and valid survey data are available for 19 students, eight of whom are female and 11 are male. Suppose that five female students said “yes” to the question whether they utilized the tutoring center in the past month, and three said “no” to that question. Of the 11 male students, four said “yes” and the remaining seven said “no.” These data may be summarized in a 2 × 2 contingency table as follows:



Student's	*Used Tutoring Center?*		
Gender	*Yes*	*No*	*Total*
Female	5	3	8
Male	4	7	11
Total	9	10	19

The null hypothesis in this setting is that there is no association between gender and utilization of tutoring center (or, with more rigor, gender and utilization of tutoring center are independent of each other). The alternative hypothesis is that there is an association between these two variables (or, equivalently, they are not independent). Assume that the test of hypothesis has to be carried out at the 5% level of significance, that is, $\alpha = .05$.

The primary step in performing the Fisher exact test would be to justify its implementation. To this end, expected counts are computed for each cell to see whether any of them are below 5. The expected count for female-yes is (8)(9)/19 = 3.79, for female-no is (8)(10)/19 = 4.21, for male-yes is (11)(9)/19 = 5.21, and for male-no is (11)(10)/19 = 5.79. For the ease of calculation, it might be noted that expected counts in each row must add up to the row total, and in each column, they must add up to the column total. So, in fact, here it is enough to calculate one expected count and figure the rest by subtraction. As can be seen, two expected cell counts are less than 5, and thus the use of the Fisher exact test is validated.

To carry out the test for independence, the probability of the observed table is computed as . Further, all possible tables with the same row and column totals are listed and their probabilities are determined. Following are the tables with the respective probabilities.

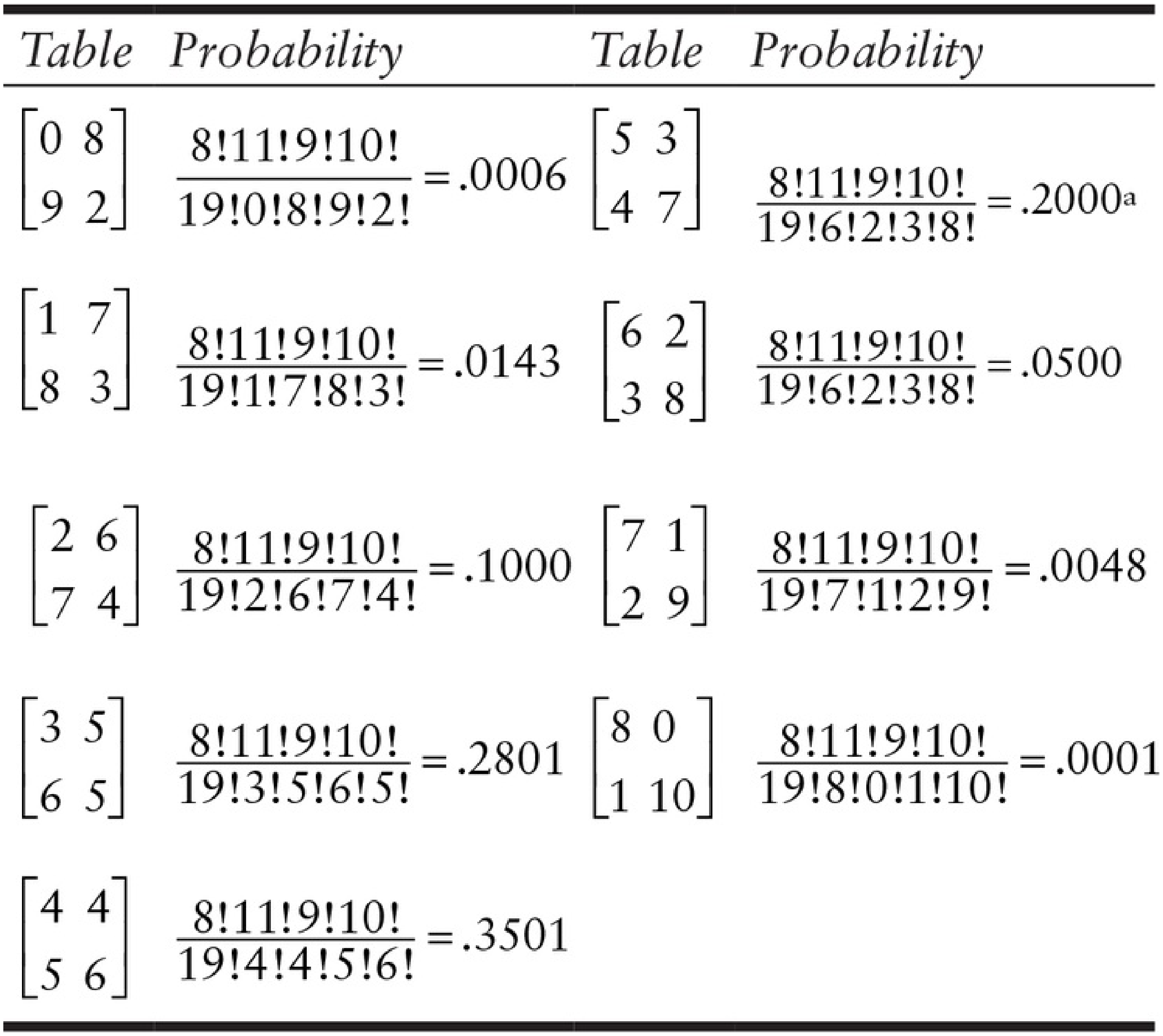


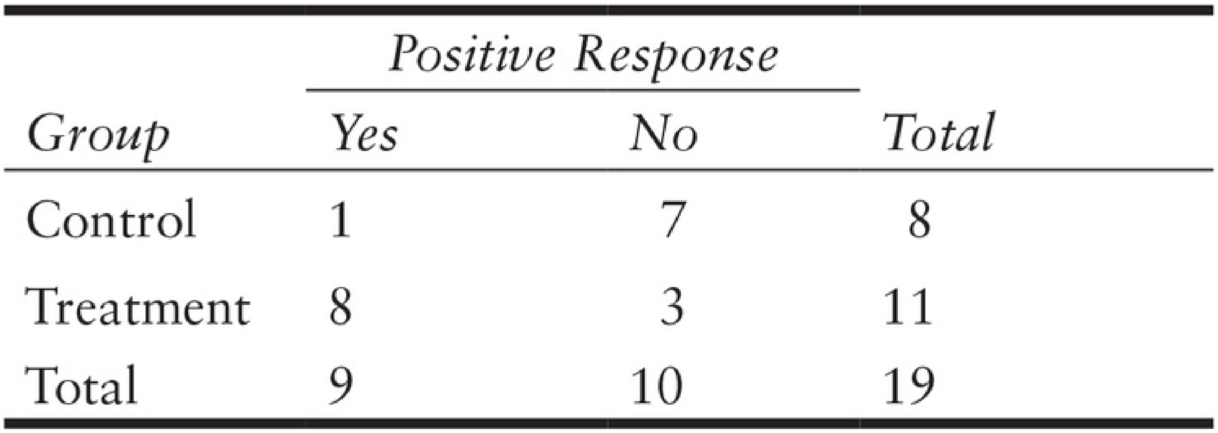
Table	*Probability*	*Table*	*Probability*
$\begin{bmatrix} 0 & 8 \\ 9 & 2 \end{bmatrix}$	$\frac{8!11!9!10!}{19!0!8!9!2!} = .0006$	$\begin{bmatrix} 5 & 3 \\ 4 & 7 \end{bmatrix}$	$\frac{8!11!9!10!}{19!6!2!3!8!} = .2000$[a]
$\begin{bmatrix} 1 & 7 \\ 8 & 3 \end{bmatrix}$	$\frac{8!11!9!10!}{19!1!7!8!3!} = .0143$	$\begin{bmatrix} 6 & 2 \\ 3 & 8 \end{bmatrix}$	$\frac{8!11!9!10!}{19!6!2!3!8!} = .0500$
$\begin{bmatrix} 2 & 6 \\ 7 & 4 \end{bmatrix}$	$\frac{8!11!9!10!}{19!2!6!7!4!} = .1000$	$\begin{bmatrix} 7 & 1 \\ 2 & 9 \end{bmatrix}$	$\frac{8!11!9!10!}{19!7!1!2!9!} = .0048$
$\begin{bmatrix} 3 & 5 \\ 6 & 5 \end{bmatrix}$	$\frac{8!11!9!10!}{19!3!5!6!5!} = .2801$	$\begin{bmatrix} 8 & 0 \\ 1 & 10 \end{bmatrix}$	$\frac{8!11!9!10!}{19!8!0!1!10!} = .0001$
$\begin{bmatrix} 4 & 4 \\ 5 & 6 \end{bmatrix}$	$\frac{8!11!9!10!}{19!4!4!5!6!} = .3501$		

[a] The observed table.

The next step is to compute the *p* value, which is the sum of all probabilities not exceeding 0.2000, which is the probability of the observed table. Hence, from the table displayed, the *p* value is .0006 + .0143 + .1000 + .2000 + .0500 + .0048 + .0001 = .3698. This *p* value is larger than the significance level of .05. Consequently, the researchers would fail to reject the null hypothesis and would conclude that utilization of the tutoring center is independent of students' gender.

Example of Fisher Exact Test for Equality of Proportions

During the pilot phase of a clinical trial for an innovative treatment for asthma, 22 people were randomly and equally assigned to the treatment and control groups. The treatment group received the innovative drug, whereas the control group was administered the best drug currently available on the market. Three people (called *subjects*) in the control group had to withdraw from the study for various reasons. Of the remaining eight subjects, one showed positive results, whereas the other seven did not. In the treatment group of 11 subjects, eight showed progress and the other three did not. The data are presented in the following 2 × 2 contingency table.



	Positive Response		
Group	*Yes*	*No*	*Total*
Control	1	7	8
Treatment	8	3	11
Total	9	10	19

In this setting, investigators would be interested in testing whether proportions of the subjects who responded positively to medication are the same in the treatment and control groups. This means that the hypotheses of interest are whether proportions for the two rows are the same in the first and second columns. In other words, researchers would be testing for column-wise equality of row proportions. The null hypothesis states that the proportions are equal, whereas the alternative hypothesis asserts that they are not equal. Assume that the researchers set the significance level at 5%.

For the ease of exposition, the observed table was chosen as one on the list in the previous example; thus, the use of the Fisher exact test is already justified. Moreover, the probability of this table is already computed as .0143. Consequently, the *p* value is equal to the sum .0006 + .0143 + .0048 + .0001 = .0198. Because the *p* value is smaller than α = .05, the null hypothesis is rejected in favor of the alternative, and the conclusion is that the proportions of subjects responding positively to the medications is not the same in both groups. Likewise, the proportions of subjects not responding to the medications are not equal in both groups.

Limitations of the Test

Some statisticians have argued that the Fisher exact test has a substantial flaw that has to do with *p*-value computation. Because the *p* value is a sum of a certain number of table probabilities, it increases discretely. The calculated *p* value is compared to a prefixed level of significance α, but in reality, the actual probability of Type I error is below the nominal level. This leads to the test being conservative.

To illustrate, suppose the null hypothesis is rejected for each contingency table for which the *p* value in the Fisher exact test is below .05. Because the set of tables is discrete, there may not be a table for which exact equality of *p* value to .05 is achieved. Then the actual significance level for this test is the largest *p* value not exceeding .05 and, for small sample sizes, the values might be much smaller than .05.

Olga Korosteleva

See also Chi-Square Test; Tests

Further Readings

Agresti, A. (2012). Categorical data analysis (3rd ed.). Wiley.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, UK: Oliver and Boyd.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika, 38, 141–149.

Carrie R. Houts Carrie R. Houts Houts, Carrie R.

Li Cai Li Cai Cai, Li

flexMIRT

flexMIRT

682

685

flexMIRT

flexMIRT is statistical software, authored by Li Cai and commercially distributed by Vector Psychometric Group, LLC, for item analysis and test scoring. Item analysis includes the estimation of item response theory (IRT) models and diagnostic classification models, both of which are widely used in educational research and measurement. The development of ever-more complex modeling frameworks and IRT models requires an adaptable and regularly updated software program capable of keeping pace with advancements in both computing and statistical/educational measurement theory; flexMIRT seeks to fulfill this need. This entry provides a broad overview of the capabilities of flexMIRT and briefly details licensing information.

First released in 2012, flexMIRT was initially published as a statistical software primarily for multidimensional, multiple group, multiple-level IRT model estimation, evaluation, and scoring within a confirmatory modeling framework using marginal maximum likelihood (or, optionally, modal Bayes) estimation. From its initial release, flexMIRT has also been able to simulate data from any model it is able to estimate. As of Version 3.0, released in the summer of 2015, updates to the program have included an alternate estimation routine better suited for truly high-dimensional models, intuitive syntax for the estimation of diagnostic classification models, expanded capabilities to estimate exploratory factor analysis models with analytic rotations, and an allowance for models that include covariates predicting the latent variables.

flexMIRT is a syntax-driven program written in C++, meaning that any system that is able to compile the language is capable of running the flexMIRT statistical engine. A graphical user interface has been created for computers running Windows, which allows for some point-and-click functionality. For operational and research situations, such as testing companies scoring thousands of respondents in the real-time or simulation studies that require a large of number of repeated analyses, flexMIRT is also able to call through the command-line interface (e.g., Windows Command Prompt), either for an individual analysis or to run a batch file that automates the running of an unlimited number of existing syntax files.

Capabilities and Features

Whether being used by a large-scale testing company or an individual user running a single analysis, all versions of flexMIRT are able to fit a wide variety of models, provide a large number of item-level and model-level fit statistics and diagnostics, produce a number of different IRT score types, and simulate data. There is no limit to the number of groups, individual observations, items, or item response categories that may be submitted for analysis, outside of the constraints of available memory and processing power of the computer on which flexMIRT is run.

The default estimation method of flexMIRT is marginal maximum likelihood via the Bock–Aiken expectation–maximum algorithm, which is the estimation method typically available in IRT software. Somewhat unique to flexMIRT is a generalized dimension reduction algorithm, which allows the program to estimate a certain subset of multidimensional IRT (MIRT) models with increased efficiency. Within the scope of models that contain more than a single dimension, flexMIRT is able to accommodate multilevel (sometimes called hierarchical) models. These models are often seen in educational research, as they allow researchers to properly account for nesting in data, such as students within the same classroom or teachers within a school; as of Version 3.0, flexMIRT is limited to models with two levels of nesting. Additionally, flexMIRT also has an alternate estimation routine called a Metropolis–Hastings Robbins–Monro algorithm, related to Markov chain Monte Carlo techniques, that is able to provide estimation for truly high-dimensional models that, historically, could not be estimated due to known computational issues associated with Bock–Aiken expectation–maximum estimation and high-dimensional models.

With respect to item types, flexMIRT is able to fit a wide variety of item models for dichotomous and polytomous items. Available item models for dichotomous items include both a model without a parameter to account for guessing (typically labeled a-and b-parameters only) and a model with a guessing parameter (typically labeled as a-, b-, and c-parameters). For polytomous items, a wide variety of item models are available, including models that assume ordered responses categories and others that do not. flexMIRT was designed primarily for the estimation of MIRT models and uses IRT item model parameterizations that are optimized for multidimensional, multilevel modeling situations. As these are likely less familiar to readers, we briefly detail and discuss the most commonly used models and implications for estimation in flexMIRT.

For all item models, we first define, for the *k*th row in η_{ijkg}, the linear predictor that is equal to , where is the set of *p* slopes on the between (Level 2)-latent variables and is the set of *q* slopes on the within (Level 1)-latent variables; if only a single-level model is fit, the Level 2 latent variables and slopes drop from the equation. For dichotomous items in which guessing is to be accounted for, flexMIRT has available a multilevel, multidimensional extension of the 3PL item model in which the probability of correct response/endorsement is defined as:

$$P_{\xi}(y_{ijkg} = 1 \mid \eta_{ijkg}) = \text{guess}_{kg} + ((1 - \text{guess}_{kg}) / 1 + \exp[-(c_{kg} + \eta_{ijkg})]),$$

where guess_{kg} is the item-specific pseudo-guessing probability and c_{kg} is the item-specific intercept. Consequently, P_{ξ} (y_{ijkg} = 0 | η_{ijkg}) = 1.0 − P_{ξ} (y_{ijkg} = 1| η_{ijkg}). As difficulty values/thresholds (b-parameters) do not have the intuitive meaning in MIRT models that they do in the unidimensional case, for each item a multidimensional intercept (labeled *c* as previously noted) is estimated and reported by flexMIRT.

For polytomous items with presumed ordered response categories, a multidimensional, multilevel extension of the graded response model is available. Suppose that item *k* has K-graded categories. Let the cumulative response probabilities be:

$$
\begin{aligned}
& P_{\xi}(y_{ijkg} \geq 0 \mid \eta_{ijkg}) = 1.0,\ P_{\xi}(y_{ijkg} \geq 1 \mid \eta_{ijkg}) \\
& \qquad = 1/(1 + \exp[-(c_{kg,1} + \eta_{ijkg})]), \\
& \ldots P_{\xi}(y_{ijkg} \geq \mathrm{K} - 1 \mid \eta_{ijkg}) \\
& \qquad = 1/(1 + \exp[-(c_{kg,\mathrm{K}-1} + \eta_{ijkg})]), \\
& P_{\xi}(y_{ijkg} \geq \mathrm{K} \mid \eta_{ijkg}) = 0.0,
\end{aligned}
$$

where the c_s are item intercepts and the boundary cases are defined for consistency. Then, the category response probabilities are $P_\xi(y_{ijkg} = l \mid \eta_{ijkg}) = P_\xi(y_{ijkg} \geq l \mid \eta_{ijkg}) - P_\xi(y_{ijkg} \geq l + 1 \mid \eta_{ijkg})$. As noted previously, difficulty values/thresholds (b-parameters) do not have the intuitive meaning in MIRT models that they do in the unidimensional case; for each item, multidimensional intercepts (labeled c_s as previously noted) are estimated and reported by flexMIRT rather than b-parameters users may expect. Users should also note that the 2PL item model for dichotomous data is a special case of the graded model in which the total number of ordered categories is K = 2.

For polytomous items with response categories that are not presumed to be ordered *a priori*, flexMIRT fits a multilevel extension of the reparameterized version of the nominal model detailed in David Thissen, Li Cai, and R. Darrell Bock’s chapter in the *Handbook of Polytomous Item Response Theory Models*. Through the application of appropriate constraints, several other popular IRT models (e.g., rating scale model, partial credit model, and generalized partial credit model) can be obtained as special cases of the reparameterized nominal model. For unidimensional models, flexMIRT will report parameters for both the reparameterized version of the nominal model and in the original nominal model metric.

By imposing constraints on item parameters, flexMIRT is able to fit a wide variety of item models consistent with the Rasch philosophy of measurement. Item types may be mixed within an analysis (e.g., 10 multiple-choice items and five ordered category items) without issue. Additionally, item models may be unidimensional, in which an item is only specified to measure a single construct unidimensional, in which an item is only specified to measure a single construct, or multidimensional, in which an item simultaneously measures two or more latent variables at once (e.g., a mathematical story problem that requires both math and reading abilities).

flexMIRT, upon request, will report a wide variety of fit statistics and diagnostics, including those related to convergence of the analysis, as well as for overall model fit (based on either full-information and limited-information techniques), individual item fit, scale information and reliability values, tests of the assumption of the local independence among items, and a test for the normality of the latent variable distribution. Associated particularly with multiple group analyses, flexMIRT is able to provide tests of differential item functioning, which allow researchers to assess the extent to which items behave differently, in terms of the estimated item parameters, across relevant defining groups such as sex, race, or income.

Additionally, flexMIRT produces values for item information functions and test information functions, which are often used to provide graphical representations of item and scale performance. Although flexMIRT does not have any built-in graphing functions, the distributors have made R code available that performs the read-in of flexMIRT-produced item parameters and item and test information values as well as the plotting of item characteristic curves/trace lines and item information functions and test information functions.

Also by providing item parameters, either from its own estimation routine or read-in from an existing source, such as previous results from a different program, flexMIRT can produce a wide variety of IRT scale scores. Available estimation methods for scale scores include maximum likelihood, maximum a posteriori, expected a posteriori, and multiple imputation estimation, which is only available when using Metropolis–Hastings Robbins–Monro estimation. Additionally, flexMIRT can produce summed score to IRT-expected a posteriori scale score conversion tables for unidimensional and certain MIRT models.

In addition to the wide variety of models, item models, fit statistics, and score types available in flexMIRT, the program has other noteworthy features. These include the ability for users to select their preferred standard error estimation method from a variety of choices, conduct exploratory factor analysis with analytic rotations, empirically estimate that shape of the latent variable distribution rather than assume it is distributed as standard normal, include observed predictors of the latent variables in the model, utilize a full-featured simulation module, and fit diagnostic classification models, a rapidly growing area of educational measurement inquiry, with relatively straightforward syntax.

Every installation of the flexMIRT graphical user interface comes with a PDF copy of the user's manual. In addition to an example-based user's manual, all syntax and data demonstrated in the manual are available on the flexMIRT support page, along with a frequently asked question page and other supporting information. flexMIRT has responsive technical support via an e-mail support desk.

Licensing

As noted earlier, flexMIRT is a commercially distributed software. New users are allowed a free 2-week trial period for evaluation purposes; the trial version is fully operational, without limitations or features made unavailable. The free trial is available for download only after registering for an account on the flexMIRT website.

Licenses are renewed yearly and are available for academic/research purposes as well as operational/commercial use. The yearly fee includes access to any updates or new versions of the program that are released during an account's valid license term. Each standard academic license of flexMIRT allows for the program to be installed and registered on up to three systems. Significant bulk discounts are available for instructors who wish to use flexMIRT in the classroom, and customized versions of the software are available upon request for large-scale clients.

Carrie R. Houts and Li Cai

See also Categorical Data Analysis; Confirmatory Factor Analysis; Diagnostic Classification Models; Exploratory Factor Analysis; Item Response Theory; Multidimensional Item Response Theory

Further Readings

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. Psychometrika, 75, 581–612.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. Journal of Educational and Behavioral Statistics, 35, 307–335.

Houts, C. R., & Cai, L. (2015). flexMIRT® user's manual version 3: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Reckase, M. D. (2009). Multidimensional item response theory. New York, NY: Springer.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: theory, methods, and applications. New York, NY: Guilford Press.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), Handbook of polytomous item response theory models (pp. 43–75). New York, NY: Routledge.

Thissen, D., & Wainer, H. (Eds.). (2001). Test scoring. New York, NY: Routledge.

Floor Level

See Basal Level and Ceiling Level

Daniel B. Hajovsky Daniel B. Hajovsky Hajovsky, Daniel B.

Flynn Effect

Flynn effect

Flynn Effect

The Flynn effect represents the secular increase in average scores on measures of intelligence. Richard Herrnstein and Charles Murray coined the term Flynn effect in *The Bell Curve* for James R. Flynn’s documentation and study of the tendency of intelligence quotient (IQ) scores to increase over time. IQ scores have increased by an average of 3 points per decade on conventional IQ tests since approximately the 1930s. Increases in IQ scores are observed by having different age cohorts take different normed versions of an intelligence test. For example, IQ scores are calibrated through standardization procedures using a sample of test takers to represent the general population. When IQ scores are normed, the average of the scores is typically scaled to a mean of 100. Approximately every 10 years, IQ tests are renormed using a younger age cohort to represent the general population. This younger age cohort typically scores higher, on average, on older versions of intelligence test batteries.

The Flynn effect has been observed across different intelligence test batteries (e.g., Raven’s Progressive Matrices, Stanford–Binet, Wechsler), with the most robust gains seen in tests of abstract reasoning such as the Raven’s Matrices or Wechsler Similarities tests. The Raven’s test is a measure of fluid reasoning, whereas the Similarities test is a measure of verbal reasoning and logical classification (e.g., “How are two things alike?”). After reviewing evidence supporting the Flynn effect, this entry examines why the Flynn effect happens and considers the implications of the phenomenon.

Evidence for the Flynn Effect

There is a lot of documented evidence to support the increasing trend in IQ scores both in industrialized and third-world countries. For example, the United States had a gain of 14 IQ points from 1932 to 1978, Estonia had a 12-point gain from 1933 to 2006, Japan had a 19-point gain from 1940 to 1965, and Argentina had a 21-point gain from 1964 to 1998. IQ gains have been robust across intelligence test batteries, ages, and ability levels. Although IQ gains have been found with an overall positive trend line, the rate in gains has varied by country, time period, and test type (e.g., scholastic vs. nonscholastic). Furthermore, the magnitude of gains has varied; and in some cases, the Flynn effect phenomenon has tapered off in a few developed nations, narrowing the gap in national IQ score differences between countries.

Why the Flynn Effect Happens

Numerous hypotheses have been advanced to explain the phenomenon of IQ score increases observed over time. Some explanations include improved nutrition, better education, greater environmental complexity, or even increases in test-taking skills. However, one drawback to the improvement in test-taking skills hypothesis is that those subtests most affected by improvements demonstrate the smallest gains. Researchers have noted that IQ score gains have occurred within too small of a time frame for genetic selection to be the cause. Another potential cause of IQ score gains over time may simply be due to artifacts of measurement differences. When comparing IQ test scores across different age cohorts, the measurement of the intended construct may vary by cohort, an issue of measurement invariance. In fact, researchers have found support to suggest that differences in measurement, not differences in true IQ, have explained the Flynn effect. Further, tests based on classical test theory versus item response theory may inhibit our understanding of gains in IQ scores. Specifically, tests based on classical test theory could show score differences in different normed versions of tests, but the differences are confounded by the fact that the difference may be due to a decreased level of difficulty in the test items, not a difference in true gains or in raw intelligence. This issue is typically resolved through the use of item response theory, as item properties can be taken into account. In general, researchers postulate that it may be some combination of factors at different times that have explained increases in IQ scores.

Why the Flynn Effect Matters

The empirical observation of IQ score increases over time, whether in true intelligence or not, has implications for psychologists, researchers, and the court of law. Psychologists and practitioners alike use a variety of IQ test batteries for the purposes of identifying disabilities, qualifying students for special education services, and implementing educational interventions. Consequently, the Flynn effect will represent an interpretive problem for those who administer IQ tests. For example, because the Flynn effect may cause published test norms to become less representative of one's intelligence as a function of time, one must be cautious in using test norms appropriately. Take the case of evaluating a student suspected of an intellectual disability, of which part of the diagnostic criteria is that an individual has a global IQ score below 70. In this case, changes in IQ norms due to the Flynn effect may affect the diagnosis—the identification and classification as someone who meets eligibility may be based more on the year the test norms were published and less on their latent level of global functioning. Furthermore, the Flynn effect may affect subtest norm interpretation, potentially confounding ipsative analyses. In the case of capital offenders, jurors or judges may hear conflicting diagnoses concerning whether an offender truly has an intellectual disability due to the Flynn effect. This has profound legal consequences, as an intellectual disability diagnosis would possibly influence the legal sentence. Lastly, those involved in empirical research may have to make decisions about whether to use the same test throughout a longitudinal study (10 years or more) or deal with score comparison issues involved in using different test batteries.

Daniel B. Hajovsky

See also Intelligence Quotient; Raven's Progressive Matrices; Stanford–Binet Intelligence Scales; Wechsler Intelligence Scales

Further Readings

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. Psychology, Public Policy, and Law, 12, 170–189.

Flynn, J. R. (2007). What is intelligence? Beyond the Flynn effect. New York, NY: Cambridge University Press.

Hiscock, M. (2007). The Flynn effect and its relevance to neuropsychology.

Journal of Clinical and Experimental Neuropsychology, 29, 514–529.

Williams, R. L. (2013). Overview of the Flynn effect. Intelligence, 41, 753–764.

David W. Stewart David W. Stewart Stewart, David W.

Focus Groups

Focus groups

687

692

Focus Groups

A focus group is a type of qualitative research that takes the form of a group discussion about a topic under the guidance of a trained group moderator. Focus group research is one of the most common research methods used by social scientists, marketing researchers, policy analysts, health and social services professionals, education researchers, political consultants, and other scientists and decision makers to gather information. Focus group research is a distinctive member of the qualitative research family, which also includes individual depth interviewing and ethnography, among others. Focus groups provide rich and detailed data about perceptions, thoughts, feelings, and impressions of group members in the members' own words. After expanding on the characteristics of focus groups, this entry introduces the emerging trend of virtual focus groups, reviews advantages and limitations of focus group research, and finally describes the process of designing, conducting, and analyzing focus group research.

Focus groups are a remarkably flexible research tool; they can be adapted to obtain information about almost any topic in a wide variety of settings and from very different types of individuals. The discussions that characterize focus group research may be very general or very specific and they may be highly structured or quite unstructured. Demonstrations, photographs, videos, products, samples, or other stimuli may be used to provide a focus for discussion and activities, such as role-playing, creative projects, and "show and tell," and can provide ways of obtaining data beyond simple discussion. The flexibility of focus group research makes it a particularly useful tool and explains its popularity.

A focus group involves a group discussion of a topic that is the "focus" of the conversation. The focus group interview generally involves 8–12 individuals who discuss a particular topic under the direction of a professional moderator who promotes interaction and assures that the discussion remains on the topic of interest. A typical focus group session will last from 1.5 to 2.5 hours. The most common objective of focus group research is to promote a deep and detailed discussion of a topic about which little is known. The group setting of the discussions that characterize focus group research is uniquely suited for quickly discovering qualitative similarities and differences among people with respect to perceptions, attitudes, beliefs, preferences, ways of doing things, and other characteristics. Focus groups also provide an efficient means for determining the language people use when thinking and talking about specific issues and objects and for suggesting a range of hypotheses about a topic. For this reason, focus groups are often used to inform the wording of surveys and to identify stimuli for subsequent quantitative research. While focus group research may be useful at virtually any point in a research program, they tend to be particularly helpful for exploratory research when little is known about the phenomenon of interest. As a result, focus groups are most often used very early in research projects and are often followed by other types of research that provide quantitative data from larger groups of respondents. Focus groups are also used following analyses of large-scale, quantitative surveys to facilitate interpretation of quantitative results and to add depth to the responses obtained in the more structured survey.

Focus groups, when properly designed and conducted, generate a rich body of data expressed in the respondents’ own words and expressions. The detail and variability in participants’ responses are high, unlike survey questionnaires that narrow responses to 5-point rating scales or other constrained response categories. In focus groups, participants can qualify their responses or identify important contingencies associated with their answers. Thus, responses have a certain ecological validity not found in traditional survey research. On the other hand, the data provided by focus groups may be idiosyncratic and unique to the group. Focus groups are particularly well suited for exploratory research that addresses broad, “grand tour” questions about “why,” “how,” “when,” “where,” and “what kind.” This is an important advantage of focus group research because it is impossible to answer quantitative questions efficiently—such as “how many,” “how much,” and “how often”—without first knowing, for example, “what kinds” to quantify.

Although focus groups can be conducted at a variety of sites, ranging from homes to offices, they are typically held in commercial facilities designed specifically for focus group interviewing. Such facilities provide one-way specifically for focus group interviewing. Such facilities provide one-way mirrors and viewing rooms where observers may unobtrusively observe an interview in progress. Focus group facilities may also include equipment for audio-or videotaping interviews and perhaps even small receivers for moderators to wear in their ears, so that observers may speak to them and thus provide input for interviews. Many focus group facilities are also equipped for “virtual” focus groups where the members may be broadly dispersed geographically and communicate through electronic media.

Virtual Focus Groups

Technology has made it possible to link people who are scattered across very broad geographic regions. This has made it possible to conduct interviews with highly specialized groups that might be difficult to assemble in a single location. The potential anonymity of virtual groups may also make participants more willing to participate when the topic is sensitive or potentially embarrassing. This latter advantage needs to be weighed against the prospect that group participants may not be who they represent themselves to be and the concern of some potential participants about sharing personal information with strangers in an electronic context. These latter issues are unlikely to be problems when respondents are prerecruited, identities are verified, and topics are not of a sensitive nature. Such circumstances would be typical of focus groups used in many marketing research situations and interviews with professionals but may be less typical in other applications of focus groups.

Use of virtual groups greatly expands the pool of potential participants and adds considerable flexibility to the process of scheduling an interview. Busy professionals and executives, who might otherwise be unavailable for a face-to-face meeting, can often be reached by means of information technologies. Virtual focus groups may be the only option for certain types of samples, but they are not without some costs relative to more traditional groups. The lack of face-to-face interaction often reduces the spontaneity of the group and eliminates the nonverbal communication that plays a key role in eliciting responses. Such nonverbal communication is often critical for determining when further questioning or probing will be useful, and it is often an important source of interplay among group members. Use of virtual groups also tends to reduce the intimacy of the group as well, making group members less likely to be open and spontaneous.

Advantages of Focus Group Research

There are several advantages of focus group research, including the following: 1. Focus groups can collect data from a group of people much more quickly and at less cost than would be the case if each individual were interviewed separately.

2. Focus groups allow researchers to interact directly with respondents. This provides opportunities for clarification and probing of responses as well as follow-up questions. Respondents can qualify responses or give contingent answers to questions. In addition, researchers can observe nonverbal responses, such as gestures, smiles, and frowns, that may carry information that supplements and, on occasion, even contradicts, verbal responses.

3. The open-response format of focus groups provides researchers the opportunity to obtain large and rich amounts of data in the respondents' own words. Researchers can determine deeper levels of meaning, make important connections, and identify nuances in expression and meaning.

4. Focus groups allow respondents to react to and build on the responses of other group members. This synergistic effect of the group setting may result in the production of data or ideas that might not be uncovered in individual interviews.

5. Focus groups are very flexible. They can be used to examine a wide range of topics with a variety of individuals and in a variety of settings. Focus groups may be useful for obtaining data from children or from individuals who are not particularly literate.

6. The results of focus group research are usually easy to understand. Researchers and decision makers can readily understand the verbal responses of most respondents.

7. Multiple individuals can view a focus group as it is conducted or review videotape or audio tape of the group session. This provides a useful vehicle for creating a common understanding of an issue or problem.

Limitations of Focus Group Research

Focus group research is not without limitations, though. Some limitations of focus group research include the following: 1. The small numbers of respondents who participate in a focus group, or even in several focus groups, limit generalization to larger populations. Indeed, persons who are willing to travel to a locale to participate in a 1-to 2-hour group discussion may be quite different from the population of interest. Such groups are rarely a “random” sample of a larger population. More often than not, focus groups are composed of judgmental samples, that is, people thought to be knowledgeable and willing to share their views about a topic.

2. The interaction of respondents with one another and with the moderator has two potentially undesirable effects. First, the responses from members of the group are not independent of one another; this restricts the generalizability of results. Second, the results obtained in a focus group may be biased by a very dominant or opinionated member.

3. The “live” and immediate nature of the interaction may lead a researcher or decision maker to place greater faith in the findings than is actually warranted.

4. The open-ended nature of responses obtained in focus groups often makes summarization and interpretation of results difficult. Statements by respondents are frequently characterized by qualifications and contingencies that make direct comparison of respondents’ opinions difficult.

5. A moderator, especially one who is unskilled or inexperienced, may bias results by knowingly or unknowingly providing cues about what types of responses and answers are desirable.

Designing, Conducting, and Analyzing Focus Group Research

As with any research method, decisions about research design are critical to the success of the method. The broad steps in the design of focus group research are similar to most other types of research in the social sciences.

Research Purpose and Data

A well-framed research purpose is critical to the success of a focus group. This purpose will guide the type of selection of respondents, the types of questions posed during the group session, and the types of analyses conducted following the group session. These issues also have implications for the number of focus groups that are fielded and group composition.

Group Composition

The characteristics of individuals who will participate in the focus group or groups are driven by the purpose of the research and consideration of the dynamics of individual within groups. Unlike survey research, where data are obtained from respondents whose answers are independent of one another, the design of focus group research must also include consideration of the likely dynamics that will be produced by any particular combination of individuals.

The Interview Guide

Although focus groups are relatively unstructured compared with the typical survey or other types of quantitative research, they are not completely without structure. The group's discussion needs to be guided and directed so that it remains focused on the topic of interest and the research questions of interest. The moderator plays an important role in maintaining this focus, but an important tool for creating the agenda for group discussion is the interview guide. The interview guide for a focus group discussion consists of a set of very general open-ended questions about the topic or issue of interest. It does not include all the questions that may be asked during the group discussion; rather, it serves to introduce broad areas for discussion and to assure that all the topics relevant to the research are included in the research. A typical interview guide for a 90-minute discussion includes no more than 10–12 questions. Generally, questions of a more general nature are raised first, and more specific issues are raised later in the guide.

The Focus Group Moderator

The moderator assures that the group discussion goes smoothly. The focus group moderator is generally a specialist who is well trained in group dynamics and interview skills. The amount of direction provided by the moderator influences the types and quality of the data obtained from the group. The moderator provides the agenda or structure for the discussion by virtue of the moderator's role in the group. The moderator must strike a balance between what is important to members of the group and what is important to the researcher. Less important to members of the group and what is important to the researcher. Less structured groups tend to pursue those issues and topics of greater importance, relevance, and interest to the members of the group. This is perfectly appropriate if the objective of the researcher is to learn about the things that are most important to the group. Often, however, the researcher has more specific information needs. Discussion of issues relevant to these needs may occur only when the moderator takes a more directive approach.

Analysis and Interpretation of Focus Group Research

The most common analyses of focus group results involve transcripts of the group interviews and discussion of the conclusions that can be drawn based on general themes of the discussion. There are occasions, however, when transcripts are unnecessary. When decisions must be made quickly, the conclusions of the research are rather straightforward, and all researchers or decision makers have had the opportunity to view the focus groups, a brief summary may be all that is necessary and justifiable. Apart from the occasions when only short summaries of the focus group discussions are required, all analytic techniques for focus group data require transcription of the interviews as a first step. Transcription provides a permanent written record of the interviews. However, nonverbal communication, gestures, and behavioral responses are not reflected in a transcript. Thus, the interviewer and observers may supplement the transcript with additional observational data, such as a videotape or notes by an observer.

Every effort to interpret a focus group discussion represents analysis of content. Some efforts are more formal than others. A quick and cost-effective method for analyzing a transcript is the cut-and-sort technique. This method involves the identification of topics or themes of interest and the identification of comments by the group that are relevant. This process can be carried out on any computer with a word processing program or with a hard copy transcript and scissors. Regardless of whether scissors or a personal computer is employed, this method yields a set of sorted materials that provides the basis for the development of a summary report. Each topic is treated, in turn, with a brief introduction. The various pieces of interview transcription are used as supporting materials and incorporated within an interpretative analysis. Although the cut-and-sort technique is useful, it tends to rely very heavily on the judgment of a single analyst. There is opportunity for subjectivity and bias in this approach. For this reason, it may be desirable to have two or more analysts independently code the focus group transcript. The use of multiple analysts provides an opportunity to assess the reliability of coding, at least with respect to major themes and issues.

assess the reliability of coding, at least with respect to major themes and issues.

More formal and rigorous approaches to the analysis of content emphasize the reliability and replicability of observations and subsequent interpretation. These approaches include a variety of specific methods and techniques that are collectively known as content analysis. Computer-assisted approaches to content analysis are increasingly being applied to focus group data because they maintain much of the rigor of traditional content analysis while greatly reducing the time and cost required to complete such analysis. It is important to note that in addition to verbal communication, there is a great deal of communication that takes place in a focus group discussion that is nonverbal and that is not captured in the written transcript. It is therefore desirable to videotape focus group sessions, so that the nonverbal behavior of participants can be recorded and coded. If videotaping is not possible, an observer may be asked to record nonverbal behavior.

The validity and utility of focus group research findings should be assessed relative to the research objectives and the degree to which the research design addresses these issues. This means that the issue of validity must be addressed throughout the focus group research process—from planning and data collection to data making, analysis, and interpretation. The execution of each step of this research process has the potential to influence the validity of focus group findings either positively or negatively. Finally, the results of focus group research must always be placed within the context of the research questions for which focus group research is appropriate.

David W. Stewart

See also Concept Mapping; Content Analysis; Epistemologies, Teacher and Student; Ethnography; Interviewer Bias; Interviews; Judgment Sampling; Market Research; Naturalistic Inquiry; Qualitative Data Analysis; Qualitative Research Methods; Threats to Research Validity

Further Readings

Jiles, T. (2013). The virtues of virtual focus groups. Greenbook. Retrieved from http://www.greenbook.org/marketing-research.cfm/virtues-of-virtual-focus-groups-04157

Kamberelis, G., & Dimitriadis, G. (2013). Focus groups: From structured interviews to collective conversations. New York, NY: Routledge.

Krippendorf, K. H. (2013). Content analysis: An introduction to its methodology (3rd ed.). Thousand Oaks, CA: Sage.

Liamputtong, P. (2011). Focus group methodology. Thousand Oaks, CA: Sage.

McCracken, G. (1988). The long interview. Thousand Oaks, CA: Sage.

Merriam, S. B., & Tisdell, E. J. (2016). Qualitative research: A guide to design and implementation (4th ed.). New York, NY: Wiley.

Seidman, I. (2013). Interviewing as qualitative research (4th ed.). New York, NY: Teachers College Press.

Stewart, C. J., & Cash, W. B. (2011). Interviewing: Principles and practices (13th ed.). New York, NY: McGraw-Hill.

Stewart, D. W., & Shamdasani, P. N. (2015). Focus groups: Theory and practice (3rd ed.). Thousand Oaks, CA: Sage.

Walden, G. R. (2008). Focus groups, Volume I: A selective annotated bibliography. Toronto, Canada: The Scarecrow Press.

Walden, G. R. (2009). Focus groups, Volume II: A selective annotated bibliography. Toronto, Canada: The Scarecrow Press.

Theodore J. Christ Theodore J. Christ Christ, Theodore J.

Allyson J. Kiss Allyson J. Kiss Kiss, Allyson J.

Formative Assessment

Formative assessment

692

696

Formative Assessment

The term *assessment* can reference either an instrument used to collect data or it can reference a process used to collect data. Assessment data may be either qualitative (e.g., narrative description of performance) or quantitative (e.g., score from a midterm exam). Either way, *formative assessment* occurs before or during program implementation. The formative data are used to improve program outcomes, which are often related to student development. As such, it is often referred to as *assessment for learning* or considered as *learner-centered assessment*. This entry defines formative assessment and evaluates its strengths and weaknesses. The entry also explains key components of formative assessment and provides an overview of its utility to improve student learning in the classroom.

In general, formative assessment is an effective tool to improve student learning. The ongoing assessment of student knowledge also enhances data-based decisions. It provides actionable evidence to guide instruction. Formative assessment provides information that allows teachers to determine who understands the lesson, what are student strengths and weaknesses, how should students be grouped, and what misconceptions or common errors need to be addressed. When effective tools are used, and results are used to inform instruction and provide explicit feedback to students, formative assessment is a powerful tool that can be incorporated into any classroom to facilitate increased student learning.

Formative Assessment in Education

Formative Assessment in Education

Both researchers and educators have placed a high emphasis on the utility of formative assessment; however, disagreements have emerged about what qualifies as formative assessment. In brief, there are at least two types of formative assessment, which are instruments and processes. Formative assessment as an *instrument* refers to established and common tools for data collection. Teachers and test publishers widely accepted this view of formative assessment. They typically refer to psychometric instruments with documented reliability and validity evidence to support their use to screen, diagnose, develop instructional plans, or monitor performance. The performance domains span academic achievement, student engagement, and social–emotional–behavioral development. Unlike summative assessments, formative assessments are often designed to guide short-term instructional decisions. The tests are often administered at multiple times a year to index and monitor instructional needs and instructional effects. Examples of standardized formative assessment include curriculum-based measurement, curriculum-based assessment, and informal reading inventories.

Teachers also widely accepted formative assessment as a *process*. It is not an assessment instrument but rather is an advanced instructional technique that incorporates a process of assessment to allow teachers to understand and evaluate student strengths and needs. Active observation, inquiry, and feedback by the teacher are considered formative assessment and instruction. Teachers act to collect data and provide input during and between instructional occasions. The data and feedback are often more qualitative in form and are used actively to guide instructional decisions to improve student learning. To the extreme of this perspective, formative assessment should not be associated with any sort of score and should instead be used to provide feedback to students, which is often provided informally.

Some take the stance that both of these definitions are key components of formative assessment. This stance would suggest that formative assessment is not solely a test instrument nor a process, but rather it is the combination of the two. With the use of high-quality formative assessments, educators can engage in the process of providing feedback and linking learning to explicit performance standards. Formative assessment is likely to be most effective when it encapsulates a combination of both the test instrument and the process of using data to guide instructional decisions.

Formative Versus Summative Assessment

Formative assessment is often differentiated from summative assessment. Although formative assessment is the “assessment *for* learning,” summative assessment can be described as the “assessment *of* learning.” Assessment that is formative involves an ongoing process to determine how learning is going and identify areas of improvement to enhance learning. Examples of formative assessments include rubrics or curriculum-based assessment. In contrast, summative assessment occurs at the end of instruction to collect data on the outcomes. Examples of summative assessment include end-of-the-year projects, final exams, course grades, or a statewide standardized achievement assessment. Generally, formative assessments do not produce critical consequences compared to summative assessments because summative assessments have the potential to influence high-stakes decisions such as determining final passing grades of a course and influencing teacher or system-level accountability.

It is not just the type of instrument or procedure that makes it formative or summative. It is also the manner in which the data are used. For example, an educator may use curriculum-based measures in a formative way to identify student strengths and weaknesses and use that information to inform intervention planning. Another educator may use the same curriculum-based measures with a summative approach to provide students with scores for their report cards or to rank the performance of students in their class. Alternatively, assessments that are typically described as summative, such as statewide tests, may be used in a formative way when teachers evaluate scores and identify areas of difficulty for their classes. In addition, the same assessment score can be used in both a formative and summative way. Some even recommend that summative assessment data should have a second formative use to enhance program planning.

Formative Assessment in the Classroom

The use of formative assessment to guide learning can be encapsulated by three questions: (1) Where am I now? (2) Where am I going? and (3) How can I close the gap and get there? These three questions take a similar theoretical approach to Lev Vygotsky’s zone of proximal development and sociocultural learning theory. Vygotsky’s framework suggests that to enhance learning or development, one must pinpoint where the learner is on a continuum of skill development. To do this, one must identify what leaners can do on their own development. To do this, one must identify what learners can do on their own (Where am I now?), what students are unable to do on their own (Where am I going?), and what learners can do with help. Following this framework, educators target instruction toward what the learner can do with help to close the gap or help the learner reach the expected standards or learning goals. Skills in the zone of proximal development are those that learners may not be able to do on their own but can be successfully performed if learners are provided with sufficient instruction, feedback, and encouragement. Thus, using formative assessment to determine where the student is now in terms of educational standards can help identify deficiencies so that instructors can adjust instructional techniques with the intent of and increasing the potential for learning.

Keeping these three key questions in mind helps establish an effective learning assessment process; however, there are a number of other factors that impact the effectiveness of formative assessment in the classroom. Research suggests a number of recommendations about the most effective components of formative assessment that are associated with increasing learning gains. These suggestions include (a) explicitly stating learning objectives and goals, (b) using explicit feedback during learning and providing opportunities to use feedback, (c) providing opportunities for students to respond and express their understanding of the material, prior to and after instruction, and (d) fostering the development of self-assessment and self-monitoring skills. A number of these key components are outlined in the following sections.

Content

To enhance learning, educators use assessment and instructional tasks that exemplify learning goals. One cannot use assessment in a formative way if it is unrelated to the purpose of instruction. With the current trend of standards-based reform in education, the emphasis on content is potentially even more important. In order to capture the more rigorous learning goals, there has been an emphasis on developing assessments that align with standards such as the Common Core State Standards. These increased standards and goals should be encapsulated not only through alignment with assessment tasks but also with instructional activities.

If the content of formative assessment aligns with important standards and learning, such expectations or goals should be communicated with students to improve learning. One flaw that often occurs during an assessment is an overemphasis of a single score. A score may inform students about their current level of performance (Where am I now?), but it does not inform the student about what the expected level of performance is (Where am I going?). Explicitly communicating expectations of performance with students can help them identify where they are going and provide them with feedback on what they need to do to get there. Many educators utilize rubrics to help make goals explicit and create a shared understanding of criteria. Once learning goals are explicit, older students may also engage in self-assessment and self-monitoring or graphing of performance to improve learning and motivation.

Prior Knowledge

Another component of effective formative assessment is examining students' prior knowledge. Assessing prior knowledge serves a number of purposes and facilitates the identification of a student's current level of performance (Where am I now?). First, cognitive research suggests that connecting or integrating new knowledge with prior knowledge is an effective learning strategy. Second, teachers can use assessments of prior knowledge to determine whether students have the prerequisite skills for learning more advanced skills. For example, a mathematics teacher may want to start a lesson in long division by assessing basic 2 × 1 division facts. If students do not demonstrate readiness for long division, teachers can modify their instruction and use reteaching or modeling techniques to prepare students for the more advanced skills. Third, assessing prior knowledge gives teachers the opportunity to address any misconceptions that students may have about a topic. This may be especially advantageous in the subject of science, in which there is a plethora of research that suggests students typically have misconceptions about topics.

There are a number of methods that can be used to assess prior knowledge. One that is often used in the classroom setting is the K-W-L technique. This technique is often used at the beginning of a new instructional topic or content area. Teachers instruct students to create two lists, one about what they already know about the topic and another about what they want to learn. After instruction, students revisit these lists and create a third, in which they write down what they learned and assess whether they learned things that they wanted to learn from the second list. This technique not only informs teachers about what prior knowledge students have but also provides them with information on what students are interested in learning. As such, teachers have the opportunity to modify their instruction to meet student needs and increase engagement.

Feedback

Another component of formative assessment, and arguably the most important, is feedback. Almost a century of research concludes that feedback facilitates learning. When reviewing the literature on formative assessment, one is unlikely to find a model or definition that does not include some component of feedback. Some even go as far as to say that formative assessment is essentially a method of receiving and providing feedback. Feedback informs students how to be effective learners by addressing any misconceptions about a topic and identifying frequent reoccurring errors to prevent them from occurring in the future. Feedback should be linked to explicit performance standards and provide guidance for students to achieve those standards. Feedback can be as informal as correcting a student’s work during whole class instruction or as formal as providing a rubric or graphing scores. To be most effective, feedback should be specific (e.g., by referencing learning goals, identifying patterns of errors), immediate (e.g., during the learning process), and positive (e.g., focus on positive learning while correcting errors). Feedback may also incorporate a component of modeling or reteaching to show students what they can do to improve.

Formative Assessment to Guide Instruction

Some researchers have set forth to develop tests that facilitate the process of formative assessment. Such tests should be designed or selected with a number of considerations in mind. First, these assessments should be psychometrically sound and follow measurement standards, such that they produce reliable (or stable) and valid scores of student learning. Second, formative assessment tools should be efficient so that teachers can easily use and quickly interpret these scores. Finally, these assessments should be instructionally relevant, such that they cover the content domain and address educational standards of interest.

An instructor may ask, “So, I have all of these data, what do I do next?” Although there are a number of existing effective tools, we have observed that many teachers do not have the tools needed to interpret these results and use them to adjust their teaching. This may be influenced by the fact that not all teachers have a strong understanding of the content domain in which they are teaching, especially at the elementary school level. As such, these teachers may not know which questions to ask or what errors to look for in student work. This makes it difficult to form hypotheses about student understanding and even more difficult to use that information to adjust instruction for all of the students in a class. Therefore, substantial work is needed to enhance professional development in formative assessment so teachers can efficiently use the plethora of informal and formal data they are collecting on a daily basis. Recently, many computer-based systems show promise to facilitate the formative assessment process (e.g., the Formative Assessment System for Teachers [FastBridge Learning] and the Cognitively Based Assessment, of, for, and as Learning [Educational Testing Service]). These assessment tools take much of the burden of interpreting results away from the teacher, thus providing more time to give students explicit feedback.

Learning progressions are one resource that has the ability to help teachers identify the next steps in modifying instruction. Learning progressions are carefully created outlines of the continua of learning for a particular skill or domain. These progressions outline the development of skills to help support and monitor learning. These progressions are often developed in terms of levels and outline skills that students need prior to moving toward the more advanced level of understanding. The progressions help teachers scaffold instruction and answer the question of “what do I teach next?” While most teachers have some ideas of how skills are developed, the formalized and often empirically based learning progressions can provide a more structured teaching tool. Although many learning progressions have been validated, it is important to note that not all students will develop skills in the same way, and the individual student needs should always be considered in designing instruction and intervention.

Limitations of Formative Assessment

Although the effectiveness of formative assessment is promising, a number of limitations do exist. First, as previously mentioned, formative assessment requires a deep understanding of the content domain, progressions of learning, and common errors. Substantial professional development is needed to master the techniques of using data to inform instruction; however, once trained to do so, the benefits are extensive. Research suggests that training increases learning outcomes for students. Second, the use of formative assessment demands teacher support and can be time-consuming. Selecting the most efficient assessment tools and incorporating assessment into current instructional practices can prevent this. A third limitation of formative assessment is disagreement over the definition. The existence of various definitions of formative assessment makes it difficult to judge the effectiveness of the practice, and some meta-analyses have displayed mixed results. Finally, there is limited research that identifies the active components that make formative assessment effective. More research is needed to determine exactly what about formative assessment improves student learning. Is it the use of explicit feedback? Is it the increased student–teacher interaction? Is it a function of standards-driven education? Is it a combination of all of these factors? These questions are currently left unanswered.

Theodore J. Christ and Allyson J. Kiss

See also Curriculum-Based Assessment; Curriculum-Based Measurement; Formative Evaluation; Psychometrics; Reliability; Summative Assessment; Validity

Further Readings

Bennett, R. E. (2011). Formative assessment: A critical review. Assessment in Education: Principles, Policy … Practice, 18, 5–25.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy … Practice, 5, 7–75.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? Educational Measurement: Issues and Practice, 28, 24–31.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. Educational Measurement: Issues and Practice, 30, 28–37.

Sadler, R. D. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18, 119–144.

Theodore J. Christ Theodore J. Christ Christ, Theodore J.

Jessie Kember Jessie Kember Kember, Jessie

Formative Evaluation

Formative evaluation

696

699

Formative Evaluation

Evaluation is the process of examining a program, procedure, or product to estimate its function, effect, and worth. There are two main functions of program evaluation in education. The first is to inform the development and implementation of the program. The second is to estimate the outcomes and program effects. Formative evaluation is the use of data before and/or during instruction or the implementation of an intervention. These data are specifically used to improve and inform curriculum planning, instructional design, and learning. The goal of formative evaluation is to meet the specific needs of students by identifying those objectives that have and have not been mastered by the student and determining what needs to be taught, individualizing educational programs for all students. Most importantly, formative evaluation is a cyclical process that includes planning, managing, delivering, and evaluating instruction, learning, programs, and interventions.

Formative evaluation allows for ongoing, real-time adaptations and modifications to aid in the development of empirically developed and empirically informed instruction or intervention practices. While formative evaluation aims to ensure that specific goals and objectives are being met, it also allows for improvements to be made. Formative evaluation can involve the use of both quantitative and qualitative data. For example, formative evaluation can rely on student performance scores on assessments or tests, and it can also rely on students’ perceptions about an intervention that has been implemented. This entry further expands on the definition of formative evaluation before detailing its use in education. Methods of formative evaluation are then reviewed.

its use in education. Methods of formative evaluation are then reviewed, followed by an examination of effective strategies and themes, and advantages and disadvantages of formative evaluation. The entry concludes with an example of instructional design.

Defining Formative Evaluation

There has been a clear and fundamental distinction between formative and summative evaluation since the 1960s. Summative evaluation specifically refers to evaluation completed at the end or summation of instruction, intervention, or program activities. In contrast, formative evaluation is intended to develop and improve a process, activity, or product in an ongoing manner, while the process, activity, or product is active. Formative evaluation and summative evaluation differ with regard to the goals and intended use of information. For example, summative evaluation provides information about the degree to which terminal outcomes have been successfully attained over the course of a class, activity, program, or intervention. In contrast, formative evaluation provides information about needs and progress during the time a program is implemented. The underlying purpose and expected uses of the information differ. Formative evaluation answers whether it is working; summative evaluation answers whether it worked. Finally, formative evaluation is not the same as formative assessment. Although all formative assessment is formative evaluation, not all formative evaluation is formative assessment. Assessment refers to the process of measuring information about a student or program to yield a source of information. Evaluation is the process of using the information that has been collected to make informed decisions. Put simply, assessment is the collection of information, while evaluation is the use of that information.

Uses of Formative Evaluation in Education

There are many potential uses of formative evaluation in an educational setting. For example, formative evaluation can serve as a needs assessment, examining whether a program or intervention is addressing a specific goal or objective. Formative evaluation specifically involves the use of data to identify individual student needs. These data are then used to plan, inform, and improve academic instruction. With increased attention toward accountability of improving outcomes for all students, the need for linking assessment to intervention practices is irrefutable. Formative evaluation can also be used to modify instruction, a program, or an intervention. Applying appropriate modifications during these processes allows instructors to increase the likelihood of success. Even further, formative evaluation may be used to determine the extent to which an intervention or program is implemented with fidelity and whether it has been implemented with consistency and quality. Thus, in some instances, formative evaluation can serve as a means for quality control. Finally, formative evaluation may be used to document progress on an ongoing basis in a standardized fashion, complementing summative evaluation methods. Although formative evaluation and summative evaluation can be successful independently, each is supplementary to the other and is more successful when used alongside the other. A comprehensive evaluation likely includes both summative and formative practices. When used in conjunction, these evaluation components can examine how an intervention or program was implemented, factors that both constrained and facilitated success and effectiveness.

Methods of Formative Evaluation

Formative evaluation may include either qualitative or quantitative data. Because formative evaluation may refer to the evaluation of activities, programs, curricula, or interventions, methods of formative evaluation are wide and varied and may include various assessment techniques (e.g., midsemester evaluations, curriculum-based measurement in reading), self-evaluation or self-assessments, surveys (i.e., open and close-ended questions), focus groups or expert review, or observation techniques. Methods employed will largely be determined by the purpose of the evaluation and the questions of interest.

Effective Strategies and Themes

For formative evaluation to be effective, there are several recommended strategies and underlying themes. The first of these strategies is defining a specific purpose of the evaluation. The purpose should be relevant. Even further, decisions about how the data will be used should be specified before the data are collected. In the realm of school psychology, there are four general types of decisions that can be made regarding individual student performance: screening, progress monitoring, analytic, and outcome. Formative evaluation is used primarily for progress-monitoring decisions. Progress-monitoring decisions refer to those decisions made to determine whether or not a student’s rate of progress is adequate. While this particular application is narrow, it can be generalized to the evaluation of activities, programs, curricula, and interventions. Regardless, there needs to be a clear and explicitly defined purpose for the program, intervention, or curriculum.

The second of these strategies is visual analysis. By evaluating the implementation of an intervention, data can be visually analyzed at multiple time points throughout the intervention to determine the effectiveness. There are four visual analysis criteria: change in mean, change in level, change in trend, and latency of change. Change in mean refers to the extent to which the average rate of performance during the intervention differs from the average rate of performance before the intervention. Change in level refers to the extent to which there is discontinuity of performance when comparing baseline data to data collected during the intervention. Change in trend refers to whether performance is increasing or decreasing throughout the intervention. Finally, latency of change refers to the amount of time that occurs before a change in performance is observed after implementing the intervention.

Visual analysis of data can be used to formatively evaluate student progress and to determine whether the predefined goals will be met. For example, if the student’s performance trend is flatter than the goal line, a decision might be made to change the implemented intervention. If the student’s performance trend is equivalent to the goal line, a decision might be made to continue the intervention. Finally, if the student’s performance trend is greater than the goal line, a decision might be made to increase the goal.

Third, effective formative evaluation rests on our ability to test hypotheses about instruction, learning, programs, or interventions. This iterative process involves examining student performance frequently, routinely, and in an ongoing manner. Evaluation is essentially founded on accurate inferences or logical conclusions derived from a given body of evidence.

Finally, formative evaluation is a dynamic process. Generally, the more frequently educators collect data, the better. While summative evaluation provides a static determination, decision, or diagnosis, formative evaluation provides a responsive, data-based problem-solving strategy. Information is specifically collected for the purpose of making decisions about instruction, learning, programs, or interventions. Formative evaluation allows for individualized educational programs based on student performance. This inductive and systematic approach to developing instruction and intervention allows educators to adjust the intensity, frequency, and content. Formative evaluation relies on follow through. A process or practice is formative to the evaluation relies on follow through. A process of practice is formative to the extent that evidence about student performance is elicited, documented, interpreted, and used whether it is used by the teacher, peers, or the learner.

Advantages and Disadvantages

Formative evaluation encourages ongoing, data-based decision making for the purpose of improving practices, processes, plans, and programs. The information obtained throughout this dynamic process increases the likelihood of success and allows for efficient resource allocation. In addition, formative evaluation provides educators with a strategy for refining practices and programs that takes a preventive approach because it takes place during the formation stage.

However, formative evaluation does come at a cost. Formative evaluation requires time and resources. Also, although formative evaluation increases the likelihood of intervention or program success, it also has the ability to distort our impressions of intervention or program effectiveness. More specifically, formative evaluation itself may be considered an intervention. Thus, it can be difficult to evaluate the independent impact of instruction, learning, a program, or an intervention. Finally, in some instances, formative evaluation may require making decisions and modifications with seemingly little evidence.

An Instructional Design Example

In regard to instruction, formative evaluation requires planning, managing, delivering, and evaluating. In planning instruction, one needs to assess the baseline skill level of students before instruction occurs or after preliminary instruction. Screening to collect student performance data can provide educators with an assessment of the instructional environment. Next, managing instruction involves adjusting the instructional level for individual students based on one’s assessment of the classroom environment. This includes identifying concepts and skills that need to be taught to certain groups of students. Effectively delivering instruction relies on continuous assessment of student mastery of the material and immediate and explicit feedback. Finally, in evaluating instruction, instructors assess student learning and set goals for future instruction.

Theodore J. Christ and Jessie Kember

See also Curriculum-Based Assessment; Curriculum-Based Measurement; Evaluation, History of; Summative Assessment

Further Readings

Bloom, B. S. (1971). Handbook on formative and summative evaluation of student learning. New York, NY: McGraw-Hill.

Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. School Psychology Review, 15, 358–374.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. Exceptional Children, 53(3), 199–208.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? School Psychology Review, 22, 27–27.

Scriven, M. (1967). *The methodology of evaluation*. In Perspectives of curriculum evaluation (AERA Monograph series on curriculum evaluation 1). New York, NY: Rand McNally.

Stiggins, R., & Chappuis, S. (2005). Putting testing in perspective: It’s for learning. Principal Leadership, 6(2), 16–20.

Dean R. Gerstein Dean R. Gerstein Gerstein, Dean R.

45 CFR Part 46

45 CFR Part 46

699

700

45 CFR Part 46

Studies on education often involve an intervention or interaction with human subjects either in person or through identifiable records, which brings such studies under the dominion of rules for protection of human participants in research, aka human subjects. At nearly all U.S. institutions under whose auspices educational research occurs, many studies involving human subjects must have their plans reviewed in advance by an institutional review board (IRB; often called “Committee for Human Research Protection” or the like). These entities are broadly governed by federal regulations that define ethical principles and procedures that researchers may be obliged to understand and follow.

The most widely cited regulation on human subjects research in the United States is referred to by the shorthand 45 CFR Part 46 (also 45CFR46), which includes the federal Common Rule applicable to all research populations and additional protections for special populations such as children. Many other countries have similar regulations. This entry discusses the history and elements of 45CFR46.

The acronym 45CFR46 refers to Code of Federal Regulations Title 45: Public Welfare, U.S. Department of Health and Human Services–Part 46: Protection of Human Subjects. There are five subparts to 45CFR46, which was issued originally in 1991, with later amendments. A revised version of 45CFR46 was published in January 2017 after 6 years of rulemaking involving extensive internal and public review and commentary. Subpart A, more than half the contents, is the Common Rule.

The 10,600-word Common Rule is the fundamental federal policy for protection of human subjects, applying to research at 18 federal departments and agencies ranging from the U.S. Department of Agriculture to the Central Intelligence Agency and to all of their respective grantees and contractors. The Common Rule and related regulations describe and mandate how institutions conducting studies funded or authorized by the federal government shall protect participants in biomedical and behavioral studies from research-related insult, harm, or injustice and provide assurance of these protections. The text of the Common Rule is reproduced (with trivial technical variations) in 15 Code of Federal Regulations titles. For example, 34CFR97 applies the Common Rule to the federal Department of Education, and 45CFR690 applies it to the National federal Department of Education and 45CFR690 applies it to the National Science Foundation.

Subparts B, C, and D of 45CFR46 define additional protections for three subpopulations: pregnant women, fetuses, and newborns (subpart B); prisoners (subpart C); and children (subpart D). Inclusion of these additional protections by individual federal departments and agencies is discretionary; the Department of Education regulations, for example, include subpart D only.

The final Subpart E provides additional details on the registration of IRBs, which departments, agencies, grantees, and contractors are required to establish in order to implement and monitor their protection of human research subjects. The Office for Human Research Protections in the U.S. Department of Health and Human Services serves as the federal government's central registrar of IRBs and of organizational assurances of compliance with the Common Rule and other subparts.

The Common Rule is based on a tripartite core of ethical principles and procedural applications set forth by the 1979 report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. These core principles and their corresponding applications are (a) respect for persons/informed consent, (b) beneficence/assessment of risks and benefits, and (c) justice/selection of subjects. Applications of the core principles are detailed in 24 sections of the Common Rule, uniformly numbered 101–124 in all pertinent parts of the CFR; for example, §46.111 "Criteria for IRB approval of research" in 45CFR46 corresponds to §97.111 in 34CFR97. A few minor sections of the Common Rule in various iterations are "reserved," that is, empty due to technicalities in how the regulations have been implemented across the government.

Common Rule sections cover definitions of key terms, the extent and limits of the rule's application, how IRBs are constituted, how they should operate and make decisions, and what information must be conveyed to assure informed consent to participate in research. The Common Rule is strictly applicable to studies performed with federal funding, but grantee and contractor institutions may and often do elect and pledge to apply the Common Rule (and other subparts) to all studies under their auspices, greatly multiplying its impact.

Research involving human subjects is defined as a systematic investigation designed to produce generalizable knowledge in the course of which there is interaction or intervention with living human subjects or their identifiable (to the research team) records. The fundamental principle of the Common Rule is that with certain defined exceptions (exemptions), the ethical acceptability (compliance with the Common Rule) of such investigations must be reviewed and determined in advance, by one or more members of the IRB under whose cognizance the research takes place.

Studies with human subjects may be deemed exempt from the Common Rule and IRB monitoring under any one of a series of conditions defined in 45CFR46, including that the research is “conducted in established or commonly accepted educational settings, that specifically involves normal educational practices that are not likely to adversely impact students’ opportunity to learn required educational content or the assessment of educators who provide instruction” (§46.104(d)(1)). Other exempt activities include behavioral research that does not subject participants to above-minimal risk of physical, psychological, or social harm; behavioral research on public officials, publicly available records, or records with no traceable identities; an official evaluation of public benefit or public service programs; or a study of taste for foods deemed wholesome by the Food and Drug Administration.

For nonexempt studies, a critical determination is whether such research involves more than minimal risk, where “risk” means the possibility of harm or discomfort, including criminal or civil liability or damage to a person’s financial standing, employability, or reputation; and “minimal” means not exceeding the level of risk encountered in everyday life, including routine physical or mental examinations or tests. A research plan (often called a protocol) that exceeds minimal risk cannot be implemented until it has been discussed and received the written approval of a convened IRB. Studies that pose minimal risk and are not exempt may be conducted after review and approval by a single designated member of an IRB who has suitable expertise (this is called “expedited” review).

Subpart D provides special provisions for research with children. Of principal relevance to educational researchers, an IRB reviewed behavioral studies in which children are identifiable and the data are sensitive, if not subject to the “normal educational practices” exemption (expedited or convened review), and both parental consent and children’s assent to participate in such research are usually required, with limited exceptions.

Dean R. Gerstein

See also Belmont Report; Human Subjects Protections; Human Subjects Research, Definition of; Institutional Review Boards

Further Readings

Department of Homeland Security and Other Agencies. (2017, January 19). Federal policy for the protection of human subjects; final rule. Federal Register, 82 (12), 7149–7274. Retrieved March 2, 2017, from https://www.gpo.gov/fdsys/pkg/FR-2017-01-19/html/2017-01058.htm

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Retrieved May 15, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Ludmila N. Praslova Ludmila N. Praslova Praslova, Ludmila N.

Four-Level Evaluation Model Four-level evaluation model

701

704

Four-Level Evaluation Model

The four-level model of training evaluation criteria is a classic framework originally developed by Donald Kirkpatrick for evaluation of organizational training but subsequently extended to many other educational contexts. The four levels in the model are *reaction, learning, behavior*, and *results* criteria. Each subsequent level of criteria provides increasingly valued information for evaluation of training effectiveness, yet the difficulty of obtaining information also increases with each subsequent level. In addition to traditional use in business and not-for-profit organizations, Kirkpatrick’s model has been applied to understanding educational effectiveness in schools, colleges, and universities and other contexts such as camps. This entry describes the model and methodological recommendations for evaluation on each level, its applications and modification in organizational contexts, and its applications to various educational contexts.

Kirkpatrick’s Four-Level Model of Training Evaluation

Kirkpatrick’s model for evaluating training programs was developed to provide actionable information to trainers and organizations to help decide whether to continue offering a particular training program and how to improve future programs. It is also used to validate the work of training professionals and organizational training enterprise overall. The four levels of evaluation allow for comprehensive evaluation of training, and it is recommended to proceed through all four levels without skipping, as all levels provide uniquely valuable information. The following are the four levels in Kirkpatrick’s model:

1. *reaction* criteria, or the participant’s feelings regarding the training;
2. *learning* criteria, or the participant’s knowledge and understanding of training content;
3. *behavior* criteria, or change in the participant’s behavior, sometimes referred to as transfer of learning; and
4. *results* criteria, or intended outcomes such as increased productivity.

Reaction and learning criteria focus on what occurs within the training program and thus are considered internal. Behavioral and results criteria focus on changes that occur outside and typically after the program and are seen as external criteria. Kirkpatrick noted that evaluation becomes more important and meaningful as it progresses from the reactions level to the results level, but at the same time, it becomes more difficult, complicated, and expensive. It is important to be reminded of the importance of unique information obtained from evaluation of the behavior and results levels when the difficulty and cost of such evaluation tempt professionals to rely solely on the less complex reaction and learning evaluation levels.

Reaction Criteria

Reaction criteria are trainees’ perceptions of training. Kirkpatrick defined evaluation of reactions as evaluation of trainees’ feelings of whether they liked the training program. One of the later modifications to the model proposed by George Alliger and his colleagues suggests distinguishing between trainees’ enjoyment of the training (affective reactions) and perceived amount of learning (utility judgments) within the reaction criteria. Suggestions for measuring reactions include clearly defining the goal of measurement and carefully aligning specific questions with that goal, developing standards for evaluation, obtaining quantifiable data, ensuring honest responses through participant anonymity, striving for a 100% response rate, and providing an opportunity for additional qualitative comments.

Kirkpatrick cautioned that although reaction criteria provide valuable information and are important to measure well, use, and communicate, positive reaction itself is not an indication of learning. It is even less of an indication of the behavioral change or results attributable to training. Meta-analysis of the relationship between the reaction criteria and the other levels of evaluation found no association between affective reactions and other levels and a very weak relationship between utility judgments and the other levels. Similarly, in educational contexts, student evaluation of teaching and self-perception of learning are found to be weakly—and in some studies negatively—related to objectively measured learning. Heavy reliance on reactions criteria in evaluation of teaching may even lead to diminished use of teaching methods that benefit long-term learning and transfer of learning, such as facilitating desirable difficulties or varying learning conditions, in favor of approaches that elicit positive reactions in the short term, yet do not result in lasting learning.

Despite the fact that Kirkpatrick and many others caution against the use of reactions alone and stress the importance of using learning, behavior, and results criteria, in practice the reaction-level criteria remain the most often evaluated, in part because of the apparent ease. However, most researchers also agree that evaluating learning, behavior, and results is essential for truly understanding the effects of learning. In education—most notably in higher education assessment—the term *indirect assessment* or *indirect evidence of learning* is often used to refer to reaction-level criteria, and the term *direct assessment* is typically used for evaluation of the other levels of criteria. Direct assessment is typically seen as essential for obtaining an accurate picture of program effectiveness. Much of the direct assessment evidence is obtained on the level of learning criteria, which is considered in the next section.

Learning Criteria

Kirkpatrick defined learning criteria as measures of the amount of a participant’s learning or knowledge and understanding of the facts that constitute the training content, skill improvement, or attitude change due to training. Appropriately measuring the learning for the purposes of training program evaluation can be achieved by obtaining quantitative results, using premeasurement and postmeasurement to ensure that learning can be credited to the program, using measurement as objective as possible and, when possible, adding a control group. Appropriate statistical analysis of the data helps ensure the quality of the evaluation. It is also important to make sure that evaluation results are used to inform the action.

Across business and education contexts, learning is typically assessed by various knowledge tests as well as skill demonstration and measures of performance in the training context, such as, for example, giving a persuasive speech or using the new software to accomplish tasks. In some cases, learning tasks embedded within training can also serve as evaluation techniques, which allows for efficient and authentic measurement. In other cases, “add-on” tests specifically serving evaluation purposes are used.

Assessing learning through pretraining and posttraining measurement is popular in educational evaluation, especially in K–12 education in the United States, which often relies on data from placement tests and prior years’ standardized assessment scores as pretests to evaluate educational gains in each subsequent grade. In higher education, much of the assessment of student learning with direct evidence relies on the learning-level criteria, although using behavioral criteria is also possible.

Behavioral Criteria

Behavioral criteria, as defined by Kirkpatrick, include measures of actual on-the-job behavior and can be used to identify the effects of training on work performance. Evaluation of training in terms of behavior is more difficult than evaluation of reaction and learning. Causal attribution of change in behavior to learning or the training program is difficult because many factors impact the behavior, including opportunities to put learning into practice, attitudes of others and group behavioral norms/group climate, availability of support, workload, among others. For example, if individuals are trained in safety behaviors but return to the workplace or an academic research facility that does not provide working protective gear or sufficient time to follow safety procedures while accomplishing assigned tasks, safety behavior will be lacking regardless of the quality of training program and individuals will have limited opportunities to practice safety skills.

Suggestions for evaluating the behavior include conducting a posttraining appraisal at least 3 months after training, so that trainees have an opportunity to put into practice what they learned and using subsequent evaluation to further add to validity of the study. Additional helpful practices include using a control group, using before-and-after evaluation, and including survey or interview appraisal of performance by as many of the following as possible: trainees themselves; trainees’ supervisors, subordinates, and peers; and others familiar with trainees’ relevant behavior or performance. Conducting statistical analyses of before-and-after performance is also important in determining causality of changes. Kirkpatrick recommended obtaining samples of 100 trainees/learners, or another sample appropriate for quality statistical analysis. Sufficient and careful sampling, as well as measuring and statistically controlling for possible intervening variables, might help alleviate concerns with establishing the link between training/education programs and behavior.

Educational contexts, specifically in higher education, provide some unique opportunities for evaluation on the level of behavior. Many degree programs introduce students to concepts and skills early in introductory courses, allow for additional practice in more advanced courses, and then facilitate internships, field practicum, or capstone experiences in which students are expected to apply learning developed during prior academic quarters, semesters, or years in a different context. Evaluation might also be conducted by multiple instructors and practicum supervisors, allowing for additional analysis of rater effects and inter-rater reliability. Careful analysis of student performance data as they progress from course-level learning to application in a different context provides unique insight for evaluation of program effectiveness and suggestions for program improvement.

Results Criteria

Evaluating results, such as reduction in costs, increase in quality, increase in production, lower rates of employee turnover and absenteeism, and higher levels of morale and engagement, is both extremely desirable and extremely difficult. Suggestions for evaluation of results are similar to evaluation of behavior and include allowing sufficient time for results to be achieved as well as repeating the measurement at additional appropriate times. Furthermore, quality of evaluation is strengthened by using a control group and measuring results-level criteria both before and after training, if feasible. Because evaluators can be deterred by the complexity and cost of evaluation at the results level, as well as by the time lag that often occurs between training and the results, it is important to consider the cost of evaluation versus the potential benefits. Evaluators also benefit from the ability to appropriately use the available evidence even if the “absolute proof” of results isn’t possible to attain.

Educational contexts and emerging “big data” methodologies provide unique opportunities for large-scale evaluation of programs and interventions on the level of results. Differences in educational approaches between nations, or between states within the United States, as well as educational interventions that have established implementation dates can be statistically analyzed in relation to population-level outcomes on various educational, civic, and economic indicators while controlling for a variety of available demographic data to provide “big picture” data on the results level.

Future Directions

The description of the four levels was originally published in 1959 and, with some further developments, the model remains widely used and continuously applied to new contexts. Although over the years there were challenges to the model and its assumptions, the model’s popularity in business and organizational settings remains high, and its popularity in educational settings continues to grow. As demands for evaluation data to substantiate educational effectiveness and value of programs continue to increase, the comprehensive nature of the four-level model coupled with its simplicity and a well-developed body of recommendations, examples, and case studies will likely result in further adaptations to help respond to new challenges of increasing accountability demands and the need to evaluate effectiveness of new delivery formats. Development of more sophisticated statistical techniques as well as availability and popularity of big data opens promising avenues for applying the model while overcoming limitations of data availability for large-scale evaluation of educational approaches on the previously elusive results level.

Ludmila N. Praslova

See also Accountability; Pretest-Posttest Designs; Program Evaluation; Transfer

Further Readings

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of relations among training criteria. Personnel Psychology, 50, 341–358.

Arthur, W., Jr., Bennett, W. J., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. Journal of Applied Psychology, 88, 234–245.

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Edens, P. S. (2003). Teaching effectiveness: The relationship between reaction and learning criteria. Educational Psychology, 23, 275–285.

Ewell, P. T. (2001). Accreditation and student learning outcomes: A proposed point of departure. Washington, DC: Council for Higher Education Accreditation.

Galloway, D. L. (2005). Evaluating distance delivery and e-learning: Is Kirkpatrick's model relevant? Performance Improvement, 44(4), 21–27.

Halpern, D. F., & Hakel, M. D. (2003, July/August). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. Change, 35, 2–13.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 13, 3–9.

Kirkpatrick, D. (1996). Great ideas revisited. Techniques for evaluating training programs. Revisiting Kirkpatrick's four-level model. Training and Development, 50(1), 54–59.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). Evaluating training programs. The four levels (3rd ed.). San Francisco, CA: Berrett-Koehler Publishers.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in higher education. Educational Assessment, Evaluation and Accountability, 22(3), 215–225.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. Annual Review of Psychology, 52, 471–499.

Nathan D. Jones Nathan D. Jones Jones, Nathan D.

Mary T. Brownell Mary T. Brownell Brownell, Mary T.

Framework for Teaching Framework for teaching

704

706

Framework for Teaching

The framework for teaching (FFT) is a classroom observation instrument used widely in teacher mentoring and professional development and, increasingly, in the evaluation of teachers (over 20 states along with hundreds of school districts either mandate or approve its use). The FFT generally receives broad support from educators because it is meant to reflect the complex professional responsibilities of classroom teaching. Rather than rely on checklists with those behaviors that are easy to measure, the FFT instead asks well-trained raters to make high-inference ratings of teachers' classroom instruction using detailed rubrics. This entry reviews the history and theoretical foundation of the FFT, the evolution of the instrument, research supporting its effectiveness, and how it is used in teacher evaluations.

History and Theoretical Foundations

In 1996, Charlotte Danielson first developed the FFT and published under the title *Enhancing Professional Practice: A Framework for Teaching*. The FFT represented an extension of the *Praxis III: Classroom Performance Assessments of the Praxis Series: Professional Assessments for Beginning Teachers*, an observation instrument developed through research conducted at the educational testing service and used to assess the teaching skills of first-year teachers.

The FFT is based on extensive empirical and theoretical literature and is intended to reflect teachers' instructional and noninstructional activities. It covers four broad domains of teaching: (1) planning and preparation, (2) the classroom environment, (3) instruction, and (4) professional responsibilities.

These four domains are composed of 22 components, which themselves are made up of 76 elements. Each element is scored on a 4-point Likert-type scale from *unsatisfactory* to *distinguished*, using rubrics with detailed descriptions at each scoring point.

The FFT reflects a constructivist approach to teaching and learning, in which students' development of knowledge is best promoted when they are doing intellectual work themselves. From this perspective, effective instruction involves the teacher designing activities that engage students in constructing their own knowledge, such as class discussions and other activities where students describe their own thinking. The constructivist perspective is threaded throughout the element-level rubrics in the FFT, with "distinguished" performance in many elements marked by students taking an active, central role in their learning. For example, in the element *activities and assignments*, evidence for distinguished performance is described as "All students are cognitively engaged in the activities and assignments in their exploration of content. Students initiate or adapt activities and projects to enhance their learning."

Changes to the Instrument

Since the FFT was released in 1996, it has been revised to refine and clarify language in the rubrics, and additional examples have been included to facilitate the use of the instrument with new teacher population. The 2007 edition of FFT included minor changes to elements' names, additional versions for nonclassroom specialists such as school nurses and librarians as well as additional information on the instrument's psychometric properties. In 2011 and 2013, revised versions of the instrument were released under the title *The Framework for Teaching Evaluation Instrument*. The goal of the most recent round of revisions was to support schools and districts in using the FFT for formally evaluating teachers; thus, updated versions have further clarified rubric language and included additional examples to support evaluators. The newest rubric is aligned with the Common Core State Standards.

The Research Base Supporting FFT

A handful of studies in the mid-2000s investigated the FFT's psychometric properties. These studies established correlations between FFT and student achievement that varied across grade levels and subjects taught. For instance, in 2003, Elizabeth Holtzapple found positive and significant correlations between FFT composite scores (a summary for the four domains) and student gains on state assessments that varied depending on the subject taught (e.g., 0.27 for reading and 0.38 for math) and the year in which the data were collected (e.g., for social studies 0.28 in 2000–2001 and 0.31 in 2001–2002). These studies did not account for the nested structure of the data; thus, estimates of relationships between FFT and student assessments may have been inflated.

More recently, researchers have used analyses that account for the nested structure of the data. Thomas Kane and colleagues examined relationships between teachers' value-added scores and their performance on the Cincinnati's Teacher Evaluation System, a rating system based on the FFT. They found that a 1-point increase in the Teacher Evaluation System was associated with a student achievement gain of one sixth of a standard deviation in math and one fifth in reading.

In the Measuring Teacher Effectiveness Project, researchers established significant, but somewhat smaller, relationships between the FFT and students' value-added scores on math (0.18) and English-language arts (0.11). These researchers were the first to study the FFT reliability. They found that it took approximately four observations of a teacher to obtain a more stable estimate of the teacher's performance. Additionally, these researchers found that when performance on the FFT was combined with teachers' valued student achievement scores and students' surveys of their teachers' instruction, a more predictive estimate of student achievement was obtained.

Applications to Teacher Evaluation

The FFT appears appropriate for evaluating general education teachers in mathematics, reading, social studies, and science, particularly combined with other measures of teacher effectiveness. However, there are a number of implementation changes that warrant further attention from researchers and policy makers. Notably, there is initial evidence that principals and other local administrators—the ones likely to conduct observations in practice—struggle to score reliably even with substantial training. The observation cycle, including preobservation and postobservation conferences in addition to the observation itself, requires a substantial investment of administrator time, likely at the expense of other responsibilities. Finally, little is known about how well the FFT captures effective teaching for teachers of special populations, such as students with disabilities or those who are culturally and linguistically diverse. Despite these concerns, information gleaned from the FFT might be used to improve teachers' instruction and classroom management skills, though little research has been conducted to demonstrate how the FFT can be used as a professional development tool. Clearly, more substantive research is needed on the use of the FFT as a tool for professional development and as a tool for assessing the impact of all teachers, including those who serve the most complex learners.

Nathan D. Jones and Mary T. Brownell

See also Common Core State Standards; Evaluation; Evaluation, History of; Teacher Evaluation

Further Readings

Danielson, C. (2013). The framework for teaching evaluation instrument. Princeton, NJ: The Danielson Group. Retrieved from https://www.danielsongroup.org/framework/

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. Journal of Personnel Evaluation in Education, 17, 207–219.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill … Melinda Gates Foundation. Retrieved from ERIC database (ED540960).

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. Journal of Human Resources, 46(3), 587–613.

Lisa Bradford Lisa Bradford Bradford, Lisa

Fraudulent and Misleading Data Fraudulent and misleading data

706

708

Fraudulent and Misleading Data

Researchers who fraudulently or misleadingly report data engage in behavior that at best is unprofessional and at worst is unethical and illegal. In this entry, fraudulent data are defined as made up and/or falsely reported data. Misleading data are data manipulated or otherwise modified so that the presentation misrepresents true research results. This entry offers an overview of fraudulent and misleading data, describes potential consequences of this practice, and identifies ways to minimize this form of research misconduct.

The Use of Fraudulent or Misleading Data

Researchers can tamper with data when they record, report, or use data for instructional purposes. According to federal guidelines for research misconduct, “current, make federal guidelines for research misconduct” the use of fraudulent or misleading data is in violation of U.S. federal laws when a researcher has (a) deviated from standard practices in the field, (b) intentionally deceived or engaged in reckless research practices, and (c) when there is sufficient evidence to support these accusations. Under federal guidelines, using fraudulent or misleading data can be classified as either falsification (when data or elements of the research process have been manipulated to improperly represent the actual data) or fabrication (data or results have been made up).

The use of fraudulent or misleading data is not limited to quantitative research activities. Researchers employing qualitative or rhetorical methods can also engage in this unprofessional activity. In the most egregious cases of research misconduct, researchers have intentionally falsified or fabricated data to achieve different results than their actual data show. For example, some researchers have made up data to inflate their results or omitted data that did not support their hypotheses. Other researchers have unintentionally reported misleading data. For example, some researchers have ignorantly created graphs that exaggerate their results.

There are many reasons why researchers would *intentionally* use fraudulent or misleading data. Environmental reasons may include financial pressure (e.g., pressure to win government grants), institutional demands (e.g., requirements and time constraints in the tenure process), competition (e.g., colleagues competing for resources), and public pressure (e.g., pressure to solve an important societal problem). Personal reasons may include desires for prestige, recognition by colleagues, and financial gain. Some scholars have criticized universities for perpetuating competitive and pressured environments that tempt researchers to engage in research misconduct. Although intriguing, this criticism does not release individual researchers from the ethical responsibility to truthfully present their research.

A case from educational measurement and intelligence testing involves British psychologist Cyril Burt, who published studies in the 1950s and 1960s, showing a strong correlation between the IQs of twins who had been raised in separate homes. The high correlation supports the view that intelligence is largely inherited. After Burt’s death, it was suggested that the reported results of different studies that involved growing numbers of pairs of twins were too similar statistically to be likely. It was also suggested that one could not locate and recruit so many twins raised in the conditions required by Burt. Some have defended Burt, however, believing that he did not fake any data, and among educational researchers the case is not closed.

Two reasons researchers *unintentionally* use misleading data are carelessness and naiveté. Examples of carelessness in research include improper recording of data or data gathering procedures; mistakes in transcribing, coding, or uploading data; and haphazard decisions about analysis procedures. Examples of naiveté include using inappropriate analysis procedures, presenting data improperly, misinterpreting the results, and failing to understand and report study limitations. Regardless of intent, there are serious consequences for using fraudulent or misleading data.

Consequences for Using Fraudulent and Misleading Data

The Eric Poehlman case exemplifies some of the personal, professional, and legal consequences of using fraudulent data. In an October 22, 2006, *New York Times* article, Jeneen Interlandi detailed the case and sentencing of Poehlman. Poehlman, a tenured faculty member, pled guilty in 2005 to research misconduct after a 5-year investigation during which he maintained his innocence and lied under oath. Poehlman, a medical researcher, had falsified data in his research on the link between obesity and aging. In the end, Poehlman apologized and admitted he obtained millions of federal research grant dollars and published several papers based on falsified data. Poehlman's case represents one of the most intensive investigations of research misconduct in U.S. history. Notably, Poehlman was also the first researcher sentenced to jail for research misconduct. He was sentenced to a jail term of 1 year and 1 day. According to Poehlman's misconduct case file on the Office of Research Integrity (ORI) website, he was also required to pay $180,000 restitution for grant fraud and all attorney fees, send retractions and corrections for 10 published articles, and was barred for life from participating in any federally funded research activities. In the end, Poehlman not only lost his career, damaged relationships with colleagues, and ruined his reputation but also damaged the reputation of the institutions where he had been employed while simultaneously reducing public trust in research.

The ORI is the U.S. federal office that provides research integrity oversight for all federally funded projects and for researchers at institutions that receive federal funding (i.e., universities). The ORI website explains that the office was established in 1992 and is under the Office of Public Health and Science within the Office of the Secretary of Health and Human Services. Between 1974 and 1981, 12 cases of research misconduct made national news. These cases gained Congressional attention and motivated hearings and legislation that ultimately led to the establishment of the ORI and the federal research misconduct policy, which contains the regulations for defining, detecting, investigating, punishing, and preventing research misconduct.

Results of a 2005 study published in *Nature* indicate that research misconduct is widespread. In one of the most comprehensive studies on research misconduct to date, Brian Martinson and his colleagues surveyed 3,247 early-or mid-career scientists about their research behaviors. Results indicated that 33% of the researchers self-reported that they had engaged in at least one of 10 research misconduct behaviors during the past 3 years. The behavior reported by the highest percentage of scientists (15.5%) was changing a study design, methods, or results to please a funding source. Six percent of the researchers indicated they had chosen not to report data that contradicted their previous research and 3% reported they had “cooked” or falsified their data. Because these data were based on self-reports, it is possible that these percentages provide very conservative estimates of the actual percentage of scientists who engage in unethical research behaviors related to data reporting.

Minimizing the Use of Fraudulent and Misleading Data

Each approach to research has different requirements for presenting data that accurately represent the research findings. Individual researchers are responsible for learning the skills necessary to do so. Researchers reporting quantitative data need to understand what methodological and statistical information should be reported so that other researchers can examine their reports and get an accurate picture of their data and analyses. For example, when these researchers report means, distribution shapes should be reported in addition to the central tendency and dispersion indicators so that readers can understand how data are distributed. They should also understand how to design graphs to present quantitative data accurately.

Qualitative researchers should provide detailed descriptions about the data gathering process and analysis that provide readers with the information they need to assess the validity of the research and the interpretation of the data. For example, when content analyses are reported with data examples to illustrate themes, the examples selected should provide an accurate and complete picture of the breadth and depth of the data that support those themes.

Ultimately, the greatest responsibility for preventing the use of fraudulent or misleading data lies with the individual researcher, but university departments can do much to encourage researchers to use and report data appropriately through better and continued research ethics education and mentoring. In graduate programs, students should be introduced to federal research regulations and research ethics. They need mentoring through research projects where best practices are demonstrated for gathering, recording, analyzing, and reporting data including specific instructions on accurate record keeping and data presentation. Students and faculty should be provided with regular opportunities to attend colloquiums that feature continuing education in data analysis, data presentation and results reporting, methods for detecting cases of data fraud, changes in federal regulations, and ethics education.

changes in federal regulations, and ethics education.

Lisa Bradford

Note: Adapted from Bradford, L. (2017). Fraudulent and misleading data. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods*. (Vol. 2, pp. 586–588). Thousand Oaks, CA: SAGE.

See also Ethical Issues in Educational Research; Ethical Issues in Evaluation; Falsified Data in Large-Scale Surveys

Further Readings

Interlandi, J. (2006, October 22). An unwelcome discovery. The New York Times. Retrieved from http://www.nytimes.com/2006/10/22/magazine/22sciencefraud.html?pagewanted=all…_r=0

Joynson, R. B. (1989). The Burt affair. New York, NY: Routledge.

Marco, C. A., & Larkin, G. L. (2000). Research ethics: Ethical issues of data reporting and the quest for authenticity. Academic Emergency Medicine, 7(6), 691–694.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. Nature, 435, 737–738.

Office of Research Integrity. (n.d.). Case summaries. Retrieved from https://ori.hhs.gov/case_summary

Office of Research Integrity. (n.d.). Historical background. Retrieved from https://ori.hhs.gov/historical-backgroun

Office of Science and Technology. (2005). PHS Policies on Research Misconduct 42 C.F.R. Parts 50 and 93. Retrieved from

https://ori.hhs.gov/sites/default/files/42_cfr_parts_50_and_93_2005.pdf

John T. E. Richardson John T. E. Richardson Richardson, John T. E.

Friedman Test

Friedman test

708

712

Friedman Test

A common design in quantitative research involves the repeated testing of participants in a number of (*k*) different treatments or conditions. A related design involves the random allocation of subgroups of *k* matched individuals to *k* different treatments or conditions. In both cases, the observations are matched across the *k* conditions, and the research question is whether there is any variation among the conditions on some criterion variable. Classically, this question is addressed using an analysis of variance (with repeated measures in the former design and randomized blocks in the latter design). However, this procedure assumes that the criterion variable in question (a) is measured on an interval or ratio scale, (b) is normally distributed, and (c) has the same variance in all of the conditions.

The Friedman two-way analysis of variance by ranks (to give the full name for the test) was developed for use in situations in which one or more of these assumptions is not met. (The “two-way” refers to the fact that the raw data are often couched in the form of a table in which the columns refer to the conditions and the rows refer to the individuals or subgroups of individuals who have participated.) This entry describes the original derivation of the Friedman test, provides a simple worked example, discusses the test’s power and power efficiency, and describes the relationship between the test statistic and Kendall’s coefficient of concordance.

Analysis of Variance by Ranks

In 1937, an American statistician and economist, Milton Friedman, suggested that the assumption of normality in the parametric analysis of variance could be circumvented by converting the data in question into ranks. If the data table contains *k* columns and *n* rows, the entries in each row are replaced by the numbers from 1 to *k*, where 1 refers to the smallest observation and *k* refers to the largest observation across the *k* conditions. (Friedman noted that it was immaterial whether the ranking was from the lowest to the highest or from the highest to the lowest.) Suppose that R_i is the sum of the ranks in the *i*th condition and that *R* is the sum of all the ranks across the *k* conditions. The deviation of the mean of the ranks in the *i*th condition is . Friedman defined a statistic that he denoted by the symbol (chi-r-square) as the sum of the squared standardized deviations across the *k* conditions.

However, this can be simplified computationally because the data in each row are simply the integers from 1 to *k*. Within each row, the sum of all the ranks is *k*(*k*+1)/2, the mean of all the ranks is (*k*+1)/2, and the variance of all the ranks is $(k^2-1)/12$. This enabled Friedman to express his test statistic in terms of the following formula:

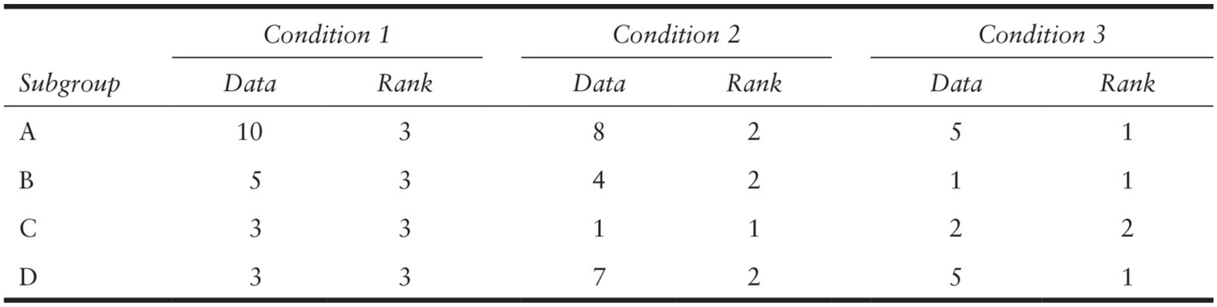
$$\chi_r^2 = \{12 / [nk(k+1)]\}\left[\sum\left(R_i^2\right)\right] - 3n\left(k+1\right),$$

where summation is carried out across the *k* conditions. For *k* = 2, Friedman noted that his test was formally equivalent to the sign test, whose properties had already been well-documented. For *k* = 3 and values of *n* between 2 and 9, and for *k* = 4 and values of *n* between 2 and 4, he presented tables showing the exact probability of obtaining a value of or a higher value under the null hypothesis that the observations in the different conditions were drawn from the same population. For larger values of *k* and *n*, Friedman proposed that the means of the ranks would be normally distributed under the null hypothesis; would therefore be distributed as chi-square (χ^2) with (*k* − 1) degrees of freedom, and it could be evaluated using existing tables of χ^2.

Friedman’s procedure assumes that the original observations are measured on at least an ordinal scale and that the observations in each row of the data set are independent of those in other rows. However, it does not make any assumptions about the parameters of the populations from which the data are drawn, and so it is an example of a nonparametric statistical test. András Vargha and Harold D. Delaney noted that, strictly speaking, Friedman’s statistic was measuring a tendency for observations in one of the conditions to be larger (or smaller) when paired with observations in all of the other conditions, a situation that they called *stochastic heterogeneity*. On the basis of results that they had obtained for the Kruskal–Wallis test, Vargha and Delaney argued that the Friedman test was a valid test of the hypothesis of stochastic *homogeneity* only if the variance of the ranks was the same across the *k* conditions. If this assumption was violated, Vargha and Delaney recommended the use of a robust parametric test on the ranks instead.

A Worked Example

A researcher is interested in whether particular interventions will influence performance on a test that is known to be influenced by a variety of demographic factors. The researcher identifies four subgroups of participants (A, B, C, and D) who are matched on the relevant demographic factors. In each subgroup, the participants are randomly assigned to receive one of the three different interventions. Table 1 shows the scores obtained by the participants on the relevant test, together with the ranks of the scores within each subgroup from 1 to 3. For these data, the sums of the ranks for Conditions 1, 2, and 3 are 12, 7, and 5; the value of is 144 + 49 + 25 = 218; and the value of is 6.50. Friedman reported that the exact probability of obtaining a value of 6.50 or greater under the null hypothesis that the observations in the different conditions were drawn from the same population was .05.



	Condition 1		*Condition 2*		*Condition 3*	
Subgroup	*Data*	*Rank*	*Data*	*Rank*	*Data*	*Rank*
A	10	3	8	2	5	1
B	5	3	4	2	1	1
C	3	3	1	1	2	2
D	3	3	7	2	5	1

The example in Table 1 was deliberately chosen to avoid tied observations. Friedman proposed that tied values should be assigned the mean of the ranks in question. For instance, the values obtained by the members of Group A were 10, 8, and 5, for which they were assigned the ranks of 3, 2, and 1, respectively. If the relevant values had been 10, 8, and 8, they would have been assigned the ranks of 3, 1.5, and 1.5, respectively. Friedman claimed that this would have little effect upon the test's validity. However, a number of researchers have provided techniques for adjusting the value of for ties, and these are implemented in modern statistical packages.

If the Friedman test yields a statistically significant result, this implies that at least one of the k conditions is different from the other conditions. However, in itself it does not indicate where such differences may have arisen. (There are, of course, a number of procedures for carrying out post hoc tests in the context of a parametric analysis of variance.) Thomas P. Hettmansperger described a procedure for carrying out $k(k-1)/2$ pairwise comparisons among the k groups that incorporated a Bonferroni adjustment to maintain the overall Type I error rate.

Power and Power Efficiency

The *power* of a statistical test is the probability of rejecting the null hypothesis when it is false. (Its complement is the probability of *not* rejecting the null hypothesis when it is false, in other words the probability of making a Type II error.) In general, nonparametric tests tend to be less powerful than the corresponding parametric test because they use less of the information that is contained in the data. (For instance, the Friedman test only uses the ranks of the observations, whereas the parametric analysis of variance uses the actual values of the observations.) The power of two different statistical tests in the same research design can be compared using the notion of *power efficiency*. This notion relies upon the fact that the power of a test in a particular situation depends (other things being equal) on the sample size. Suppose that Test 1 is the most powerful statistical test when used in a particular research design with data that meet its underlying assumptions. Test 2 is a less powerful test in the same design, in that it would need to be used with a sample of N_2 cases to match the power that is achieved by Test 1 with N_1 cases (where $N_2 \geq N_1$). The power efficiency of Test 2 is N_1/N_2, often expressed as a percentage.

It had been shown that the power efficiency of the sign test was only 63.7%, and Friedman inferred that the same would be true of his own test with $k = 2$. He did not report the power efficiency of his test for $k > 2$, but he claimed that it was likely to be higher than this. In fact, Erich L. Lehmann showed that it gradually approached a value of $3/\pi$ or 95.5% as the number of conditions increased. Friedman also compared the results of his test with those of the parametric analysis of variance for 56 sets of data in which the underlying assumptions of the latter were met. The results were in fact remarkably similar, and in 45 of the

56 cases, the significant levels yielded by the two procedures were essentially the same. Accordingly, the Friedman test can be recommended as an acceptable distribution-free test.

Kendall’s Coefficient of Concordance

The schema underlying the Friedman test can be used to raise other research questions. Suppose that *k* individuals or objects have been ranked from 1 to *k* by each of *n* independent judges. Two British researchers, Maurice G. Kendall and Bernard Babington Smith, discussed how one might judge the degree of consistency among the judges in their rankings. Using the notation described earlier rather than Kendall and Babington Smith’s notation, R_i is the sum of the ranks awarded to the *i*th object, *R* is the sum of all of the ranks across the *k* objects, and *S* is the sum of the squared deviations, . The minimum value of *S* is 0, when the sums of the ranks are all identical, reflecting no agreement among the judges whatsoever. The maximum value of *S* is $n^2(k^3-k)/12$, reflecting complete agreement among the judges. The statistic *W* is defined as ; it varies between 0 and 1 and is known as Kendall’s coefficient of concordance. It should not be confused with Kendall’s τ, which is a correlation coefficient.

Kendall and Babington Smith proved that *W* was directly related to the average of the Spearman rank correlation coefficients among all possible pairs of the *n* rankings. It was also related to Friedman’s statistic by the equation ; equivalently, . An American statistician, W. Allen Wallis, independently arrived at Kendall and Babington Smith’s statistic; he characterized it as the rank correlation ratio. In the parametric analysis of variance, the correlation ratio measures the proportion of the total variance that is explained by the independent variable and is denoted by the symbol η^2. Wallis denoted the rank correlation ratio by the symbol . Kendall and Babington Smith extended Friedman’s tables showing the exact probability of obtaining a value of or a higher value under the null hypothesis that the observations in the different conditions were drawn from the same population. They also provided a normal approximation that could be used with moderate values of *k* and *n*.

For the rankings shown in Table 1, the sums of the ranks are 12, 7, and 5; the mean of all the ranks is 24/3 = 8; and the sum of the squared deviations, *S*, is . Consequently, $W = (12\times26)/[4\times4\times(27-3)]=312/384=.8125$. From these calculations, , the figure that was calculated previously.

John T. E. Richardson

See also Analysis of Variance; Binomial Test; Bonferroni Procedure; Power; Rankings; Type I Error; Type II Error

Further Readings

Friedman, M. (1937). The use of ranks to avoid the assumption of normality inherent in the analysis of variance. Journal of the American Statistical Association, 32, 675–701. doi:10.2307/2279372

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of *m* rankings. Annals of Mathematical Statistics, 11, 86–92. doi:10.1214/aoms/1177731944

Hettmansperger, T. P. (1984). Statistical inference based on ranks. New York, NY: Wiley.

Kendall, M. G. (1970). Rank correlation methods (4th ed.). London, UK: Griffin.

Kendall, M. G., & Babington Smith, B. (1939). The problem of *m* rankings. Annals of Mathematical Statistics, 10, 275–287. doi:10.1214/aoms/1177732186

Lehmann, E. L., & D’Abrera, H. J. M. (2006). Nonparametrics: Statistical methods based on ranks (rev. ed.). New York, NY: Springer.

Siegel, S., & Castellan, N. J., Jr. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York, NY: McGraw-Hill.

Vargha, A., & Delaney, H. D. (1998). The Kruskal–Wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics, 23, 170–192. doi:10.3102/10769986023002170

Wallis, W. A. (1939). The correlation ratio for ranked data. Journal of the American Statistical Association, 34, 533–538. doi:10.2307/2279486

G

Daniel B. Hajovsky Daniel B. Hajovsky Hajovsky, Daniel B.

Matthew R. Reynolds Matthew R. Reynolds Reynolds, Matthew R.

g Theory of Intelligence *g* theory of intelligence

713

715

g Theory of Intelligence

In psychology, human intelligence is one of the most researched constructs, with its theoretical development spanning more than 100 years. Although there is no agreed upon verbal definition of intelligence, researchers and psychologists generally agree that it involves abstract reasoning and thinking and the ability to understand complex ideas and acquire knowledge. One question that arises is why do people differ in intelligence? Is it mostly due to one thing or is it mostly due to many things (e.g., memory, numerical ability, and verbal ability)? The *g* theory of intelligence describes differences in intelligence as mostly due to one thing. That "thing," often referred to as the "*g* factor," is commonly used to refer to general mental ability, general intelligence, or psychometric *g*. This entry describes the construct of *g* and then discusses the historical development of the *g* theory of intelligence.

The *g* Factor

The *g* factor is the most important variable in a *g* theory of intelligence. In psychology, the *g* factor is used to explain one of the most remarkable facts—scores on all mental ability tasks correlate positively. These mental ability tasks can include deciphering the logical progression of a complex pattern, recalling numbers backward, and defining words.

Despite the diversity in the universe of mental ability tasks, someone who performs well on one task, on average, also performs well on the other tasks. For example, individuals who perform very well on a task that requires them to manipulate shapes to make a pattern, also on average, perform well on a task that requires them to verbally define words. Although the tasks vary considerably, the performance across the tasks is related. The reason the performance is related is due to the *g* factor.

The *g* factor represents what is common across an infinite universe of mental ability tasks. That is, performance across all mental tasks shares a common influence, namely, the *g* factor. Through these mental ability tasks *g* is called forth, which is manifested in test score performance. Because the *g* factor represents general intelligence, the measurement of *g* does not rely on any specific mental ability task. Instead, the *g* factor is best represented by performance across a breadth of diverse mental ability tasks.

Today there are various intelligence tests that are available to measure intelligence. These tests and their constituent tasks (i.e., subtests) may vary widely in their appearance or theory underlying them, but a composite derived from the tasks, called an IQ, for the most part measures *g* the same across the tests. In 1904, Charles Spearman coined the phrase “the indifference of the indicator” to describe this very phenomenon—the surface characteristics of the specific task are unimportant in measuring the *g* factor.

Although IQs across a variety of intelligence tests tend to measure the same amount of *g*, not every mental ability task within each intelligence test is an equal measure of *g*. For example, reasoning-or vocabulary-type tasks are better measures of *g* than tasks that require a person to repeat back a set of numbers. The more complexity involved in a mental ability task, the more of *g* is measured. Complex tasks generally require more mental manipulation or information processing. A composite summarizing scores from a variety of mental ability tasks tends to measure a similar—though not identical—amount of *g* as does another composite summarizing scores from a different set of mental ability tasks. But a single mental ability task may vary quite substantially from another in how much *g* is measured in that task. All mental ability tasks measure *g* to some extent, but the IQs are the scores that mostly reflect *g*.

Although an IQ score represents the *g* factor, the *g* factor is not the same as an IQ—IQs are vehicles for the measurement of *g* much like thermometers are vehicles for the measurement of temperature. The *g* factor can be derived, however, through the factor analysis of mental ability scores. Factor analysis is a statistical technique used to uncover latent variables that produce correlations among variables. Because mental ability tasks are all positively correlated (this is sometimes referred to as positive manifold), a general factor or *g* is extracted to account for a large portion of the shared variance in the mental ability tasks; the shared variance is expressed in those correlations. According to the most basic *g* theory of intelligence, the general factor accounts for all of the shared variance among mental ability tasks, and two abilities produce individual differences in each mental ability task, the general ability and an ability that is specific to each test. (In reality though, other factors are also extracted, and usually hierarchical factor analysis is used, but that is beyond the scope of this entry.) The *g* factor is an unobservable psychological trait or a construct. It is often interpreted as a causal psychological variable that produces individual differences in all mental ability tasks. The *g* factor not only is related to performance on mental ability tasks but also has well-established relations with a number of nonpsychometric variables. For example, individual differences exist in the latent level of *g* and that variability explains differences in numerous academic and life outcomes. The *g* factor is often the best predictor of school performance and on-the-job performance, particularly as complexity increases in these performance areas. The *g* factor is also related to a number of biological and physical variables. Although genes and the environment both play a role in the formation and development of *g*, the heritability of *g* tends to increase with age, with genes accounting for 40%–80% of the variation in *g* by late adolescence.

Although the *g* factor represents what is common across mental ability tasks, it is specifically the pattern of positive correlations between mental ability tasks that supports the existence of a latent *g* factor. One interesting finding related to the patterns of positive correlations is the patterns of correlations between mental ability tasks is stronger in magnitude among individuals at the lower end of the distribution of intelligence than in those at the higher end. This pattern was first discovered by Spearman in the early 20th century and later “rediscovered” by Doug Detterman and Mark Daniel in 1989. Spearman likened this phenomenon to the law of diminishing returns in the field of economics. That is, *g*’s influence on test scores decreases as levels of *g* increases.

Why does the *g* theory of intelligence matter? One of the most important reasons is because performance on tasks that measure the *g* factor has strong, positive relation with performance on learning tasks. That is, the *g* factor may represent a mental capacity for learning in a wide variety of situations. The relation between *g* and learning may become apparent in classrooms where different individuals acquire qualitatively different information and skills at different rates. For example, children display differences in their ability to infer the meaning of words used in the classroom, to engage in trial-and-error problem solving, or to apply learned skills to new problems. These differences are even more apparent when learning tasks involves no prior learning or skills. With that said, the *g* factor is a good predictor of the aggregate learning across a wide variety of learning tasks.

Factor Analytic Studies of Intelligence

The discovery and subsequent development of the *g* theory of intelligence dates back to the late 1800s to early 1900s. It was Francis Galton, the founder of the field of individual differences, who later influenced the English psychologist Spearman to study individual differences. Spearman was able to study Galton's hypothesis that there is a general mental ability that enters into all mental activity. Spearman discovered that there was a tendency for all tests of mental ability to positively correlate.

Spearman subsequently theorized that an underlying common cause, namely, *g*, was the primary influence on test performance. Thus, in 1904, Spearman's two-factor theory of intelligence was born and used to explain individual differences in test performance. The two-factor theory consisted of a general or *g* factor, which is common to all mental ability tests, and a specific factor unique to each individual mental ability test. In today's parlance, the two-factor theory would be described as a one-factor theory. Thus, Spearman conceived intelligence as the result of "one" thing, not many.

One researcher who is associated with a conception of intelligence as a multi-dimensional construct was L. L. Thurstone. In 1938, Thurstone was developing and conducting multiple factor analysis on mental ability test data. He theorized that the structure of intelligence was based on seven to nine primary mental abilities that were independent of each other (e.g., reasoning, verbal, and numerical), so there was not a general factor. The reason for a lack of general factor, however, was due to the early factor analytic techniques he developed, and later, he found that the seven to nine primary abilities correlated.

Eventually, Thurstone changed his stance and thought the correlations among primary abilities were likely due to a general factor. Spearman also acknowledged the likely presence of group factors (similar to the primary abilities) beyond the general factor. The two researchers differed in their emphasis, however. Spearman clearly emphasized the general factor, whereas Thurstone emphasized the primary abilities.

The contemporary researcher often associated with the *g* theory of intelligence is the late Arthur Jensen. In 1998, Jensen penned one of the most thorough and empirically documented accounts of the existence, relevance, and distribution of *g* factor in a book called *The g Factor*. Jensen's work documented in painstaking detail the scientific study of mental ability, with a focus on the *g* factor. Although it was not without controversy, *The g Factor* is regarded as a seminal piece in the scientific study of individual differences in intelligence. That book, along with Spearman's original work, is considered the definitive resource for the *g* theory of intelligence.

Daniel B. Hajovsky and Matthew R. Reynolds

See also Cattell–Horn–Carroll Theory of Intelligence; Intelligence Quotient; Intelligence Tests; Multiple Intelligences, Theory of

Further Readings

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor analytic studies. New York, NY: Cambridge University Press.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. Intelligence, 24, 79–132.

Gottfredson, L. S. (2008). Of what value is intelligence? In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), WISC-IV clinical assessment and intervention (2nd ed., pp. 545–563). Amsterdam, the Netherlands: Elsevier.

Jensen, A. R. (1989). The relationship between learning and intelligence. Learning and Individual Differences, 1, 37–62.

Jensen, A. R. (1998). The g factor. Westport, CT: Prager.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project:

Standing on the shoulders of the giants of psychometric intelligence research. Intelligence, 37, 1–10.

Reynolds, M. R. (2013). Interpreting the *g* loadings of intelligence test composite scores in light of Spearman’s law of diminishing returns. School Psychology Quarterly, 28(1), 63.

Schneider, J. W., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (3rd ed., pp. 99–144). New York, NY: Guilford Press.

Spearman, C. E. (1927). The abilities of man: Their nature and measurement. New York, NY: Macmillan.

Immanuel Williams Immanuel Williams Williams, Immanuel

Steven Andrew Culpepper Steven Andrew Culpepper Culpepper, Steven Andrew

Gain Scores, Analysis of Gain scores, analysis of

715

717

Gain Scores, Analysis of

The analysis of gain scores is the evaluation of the difference between pretest and posttest scores in terms of treatment versus control design. In other words, this technique is used to determine the effect of a treatment on the difference between pre-and posttest scores compared to a control group. This process involves using analysis of variance (ANOVA) to determine whether the difference between pretest and posttest scores can be predicted by group membership (treatment *vs*. group). This design is important because it gives the researcher the ability to assess whether a protocol or method improves performance compared to the status quo. This entry discusses the history of analysis of gain scores, when and how gain scores are analyzed, and the limitations of this analysis.

The analysis of pretest and posttest scores in education dates back to 1956 when Frederic Lord discussed the difference between these test scores in terms of performance differences caused by summer vacation. Lord extended the methodology by evaluating the influence of specific diet changes in weight for a sample of men and women. Lord’s methodology for assessing change served as an early example for researchers on how to analyze the effectiveness of a treatment in terms of pre-and postmeasures.

The decision to analyze gain scores depends on the research question and the design of the experiment. Gain scores should be employed whenever the goal is to determine the influence of a treatment or intervention on the change from pre-to posttest. The null hypothesis is typically taken as the absence of improvement between the treatment and control groups and evidence for or against the null hypothesis can be tested via ANOVA on the gain scores using treatment (treatment *vs*. control) as a between subjects factor. If the treatment main effect is significant, then we reject the null hypothesis and state that there is a difference between treatment and control in terms of gain scores. An application of gain score analysis in education is the use of value-added models to evaluate teachers, schools, and district policy by analyzing differences in students' test scores over time to track changes in these students' academic performance.

The correlation between pretest and posttest scores within the treatment and control groups is a measure of the consistency of the treatment and control effects within both populations. A large correlation between tests implies consistency in performance, while a small correlation signifies nonconsistency in performance.

The steps in gain score analysis can be understood by considering its use to determine whether a new teaching method (Method A) improves students' performance compared to an existing method (Method B). One could administer a pretest to students and then randomly assign the students to either Method A or Method B. After each group of students is taught by the respective method, the researcher would administer a posttest. The gain score is the difference between the posttest and pretest scores. A positive difference denotes a positive gain in performance, whereas a negative score suggests a decline in performance.

A one-way ANOVA can be used to determine whether gains in student performance differ between the two methods of instruction. Another statistical technique that will provide a similar result is the repeated measures analysis of variance within a 2 × 2 design.

It is tempting to think the *t* test and ANOVA can be implemented to evaluate the difference between pretest and posttest scores with respect to a specific treatment in terms of an observational study. Both techniques are designed to determine whether there are gains. However, the interpretation of results is dependent upon the experimental design. For example, in nonexperimental settings, pretest scores relate to both experimental and control group membership and to the gain score, which is related to the familiar omitted variable bias problem. Consequently, researchers must be careful when analyzing gain scores in nonexperimental studies to include as many relevant covariates when making group comparisons. An example is with value-added modeling where students are not randomly assigned to classrooms and researchers are unable to estimate the causal effect of teachers on student growth. In contrast, random assignment into treatment and control groups as found in experimental studies alleviates the problem of omitted variable bias, so that researchers can make causal inferences regarding group differences.

Blocking and matching are methods of analyzing the variance between groups and within groups with respect to pretest and posttest score, respectively. Blocking can be used if it is known that a natural group that performs differently is found in both treatment and control groups. For example, if it is known that performance levels differ between schools or classrooms, it would be ideal to control for this factor and use a block design to analyze the difference in performance. Matching on the other hand can be used to match individuals who received the treatment to individuals who did not based on similar characteristics. This technique is a method to reduce the differences that may be inherent in performance within a group that is receiving the treatment.

One limitation to analyzing gain scores in the social sciences relates to the issue of measurement error. In particular, Lee Cronbach and Lita Furby noted that the difference between pre-and posttest scores tends to be less reliable in cases where the pre-and posttest scores are measured with error. One consequence is that additional measurement error may impact the power to detect effects. More recent research has employed latent variable models, such as structural equation model, to measure change.

Although the analysis of gain scores controls for individual differences in pretest scores by measuring the posttest score relative to each person’s pretest score, gain score analysis does not control for the differences in pretest scores between the two groups. That is, if one group’s pretest scores are significantly different from the other, the analysis of the effect of a specific treatment with respect to one of the groups is not reliable. Another issue that needs to be addressed is that the gain scores are less reliable compared to the pretest and posttest measures. One strategy to rectify this issue is to implement structural equation modeling. Structural equation modeling is a method that can be used to account for the measurement error found in both measures.

Immanuel Williams and Steven Andrew Culpepper

See also Analysis of Covariance; Matching Items; Repeated Measures Analysis of Variance; *t* Tests, Value-Added Models

Further Readings

Cronbach, L., & Furby, L. (1970). How we should measure “change”—Or should we? Psychological Bulletin, 74, 68–80.

Culpepper, S. A. (2014). The reliability of linear gain scores at the classroom level in the presence of measurement bias and student tracking. Applied Psychological Measurement, 38, 503–517.

Lord, F. M. (1956). The measurement of growth. Educational and Psychological Measurement, 16, 421–437.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29, 67–101.

Raudenbush, S. W. (2015). Value added: A case study in the mismatch between education research and policy. Educational Researcher, 44(2), 138–141.

Vautier, S., Steyer, R., & Boomsma, A. (2008). The true-change model with individual method effects: Reliability issues. British Journal of Mathematical and Statistical Psychology, 61, 379–399.

Yoon Jeon Kim Yoon Jeon Kim Kim, Yoon Jeon

Game-Based Assessment Game-based assessment

717

719

Game-Based Assessment

Game-based assessment (GBA) refers to the use of games, both video games and other types of games, to assess learners' various competencies—skills, knowledge, and dispositions. Practical implementations of GBA can vary from context to context. Different GBA models have varying levels of assessment capacity, and the foci depend on how much the role of assessment is emphasized in the process of designing the game. For example, if both assessment and learning designers are involved in the early design stages, it is more likely that assessment and game mechanics will be seamlessly integrated (i.e., stealth assessment). However, even if a game is developed without explicitly supporting assessment, educators can still use it to create GBA activities.

Well-designed GBA can (a) provide engaging and authentic contexts, (b) elicit evidence for the competency of interest, and (c) motivate learners to continuously adjust their actions, which can lead to learning. The underlying assumption of GBA is that when the learners attempt various problems in GBA, their interactions with the game provide evidence for underlying competency, and the gameplay simultaneously provides immediate feedback in response, motivating them to continuously modify their actions and strategies. This entry first discusses the advantages of GBA, then describes the use of evidence-centered design (ECD) for GBA. It concludes by discussing practical challenges faced in the use of GBA.

Advantages of GBA

The educational assessment community has recognized the needs for new kinds of assessment that (a) are based on modern theories of learning, (b) provide authentic real-world problems, (c) require application of multiple competencies, and (d) provide teachers and students with actionable information. Games, particularly video games, can be used as a vehicle for such assessments. Game and assessment design share similar principles of learning and employ compatible design processes. That is, game design focuses on creating mechanics that can continuously monitor and quantify players' interactions with the game and provide feedback to the players or summarize their performance in relation to other players' skills or resources. Similarly, educational assessment is the activity of observing what students say, do, or make, and of quantifying these observations in a meaningful way, to make more general inferences about their skills and knowledge.

Video games have great affordances for educational assessment for several reasons. First, playing video games is an integral part of daily life for many children and teens. A nationwide survey in 2015 found that 8-to 18-year-olds on average spend around 80 minutes each day playing video games, including games played on console and handheld video game players, computer games, and mobile games. Second, large amounts of data generated from gameplay can be rapidly collected without interrupting the learners' engagement in video games, which means assessment can be seamlessly embedded in their daily activity. Third, this ability to extract data can yield rich, comprehensive student models, which can be used to diagnose students' learning needs, provide formative feedback, and change gameplay to maximize learning according to the player's ability level. Fourth, GBA employs challenging problems involving the types of complex situations necessary to evaluate the application of 21st-century competencies that are often underemphasized in conventional school education. Finally, when people are engaged and motivated with a given task, more accurate inferences can be made about them.

ECD for GBA

To create a well-designed GBA, game and assessment designers need a common language to align game and assessment mechanics. ECD, an assessment design framework, has been widely adopted by the educational game community to develop assessment models for GBA, which in turn support the balance between game design and assessment design. The central principle of ECD is that educational assessment is an evidentiary argument. ECD guides the design and implementation of assessment as a principled process by formalizing the assessment structure to systematically align students' in-game actions with the specific competencies about which the assessors wish to make inferences. ECD is a process of addressing three questions that should be asked in any assessment design: *what*, *where*, and *how* are we measuring. This process leads to several design objects including competency, task, and evidence models, as follows: A *competency model* (CM) directly reflects the types of claims that the assessor wishes to make about students at the end of the assessment. Typically, one CM is used for a given assessment, but ECD explicitly assumes multidimensionality of CMs. In GBA, CMs represent students' skills, knowledge, and other traits for which a given game can provide evidence.

A *task model* (TM) involves tasks, which are individual units of activity attempted by the student. The student's interactions with the task produce a *work product* that is then scored and used to update the assessor's inferences about the student's competency. A work product is an object that students produce as they respond to or interact with the assessment. A work product can be as simple as a response to a multiple-choice item in conventional assessment or as complex as a series of actions and choices in interactive environments such as video games. A TM is a collection of the task features (i.e., TM variables) that the assessment designer must consider when engineering the contexts necessary to elicit evidence of the targeted aspects of competency. Each TM must have different levels of evidentiary strength or focus. Therefore, each TM variable has a range of possible values and provides one or more functions that influence the argument structure of the assessment.

An *evidence model* bridges CMs and TMs by specifying the student work products and associated scoring rules and by using statistical models that send the collected information to the CM. An evidence model includes two processes. The first process is an evaluation component that considers the salient features and values of work products for an evaluative outcome. This process involves evidence rules, which are comparable to scoring rubrics. In GBA, evidence rules can be specifications of players' observable behaviors in the game, and how these behaviors afford evidence with different levels of strength. The second process is the statistical component that analyzes how the obtained new evidence relates to CM variables in probabilistic terms.

Practical Challenges

Increasingly, educational game researchers and practitioners are emphasizing the importance of aligning students' learning with what students do in games.

However, it is often unclear how these researchers and practitioners leverage assessment to conceptualize game design around the competency of interest, even if they claim to use ECD. Therefore, there is a need for greater communication between design teams and the broader community. This communication must address the diverse methods and processes by which design teams, which often include learning scientists, subject-matter experts, and game designers, can seamlessly integrate design thinking and the formalization of assessment models. Some specific challenges that researchers and practitioners might face include the following:

How can assessment models be formalized?
How can formalized assessment models be translated into game design elements?
At what point(s) in the game design process does this translation occur most effectively?
How can CMs be transformed into interesting, engaging game mechanics?
How can psychometric qualities be ensured without being too prescriptive?

Furthermore, because GBA design requires the satisfaction of both psychometric and entertainment criteria, it is based on the assumption that GBA can offer a “sweet spot” that simultaneously meets these two different sets of criteria. However, little is known regarding how game and assessment designers can balance the design considerations of games versus assessment to maximize the effectiveness of GBA without losing game-like characteristics such as fun and engagement. Both researchers and practitioners need to develop an archive of design patterns and design principles that are specific to GBA.

Yoon Jeon Kim

See also Computer-Based Testing; Computerized Adaptive Testing; Evidence-Centered Design

Further Readings

Common Sense Media. (2015). The common sense census: Media use by tweens and teens. Retrieved from https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens

Gee, J. P. (2003). What video games have to teach us about literacy and learning. New York, NY: Palgrave Macmillan.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. Computers … Education, 87, 340–356.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), Assessment in game-based learning: Foundations, innovations, and perspectives (pp. 59–81). New York, NY: Springer.

Mislevy, R. J., Oranje, A., Bauer, M., von Davier, A. A., Hao, J., Corrigan, S., *et al.* (2014). Psychometric considerations in game-based assessment. New York, NY: Institute of Play.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), Computer games and instruction (pp. 503–524). Charlotte, NC: Information Age.

Marc Hallin Marc Hallin Hallin, Marc

Gauss–Markov Theorem

Gauss–markov theorem

720

723

Gauss–Markov Theorem

The Gauss–Markov theorem, named after German mathematician Carl Friedrich Gauss and Russian mathematician Andrey Markov, states that, under very general conditions, which do not include Gaussian assumptions, the ordinary least squares (OLS) method in linear regression models provides best linear unbiased estimators (BLUEs), a property that constitutes the theoretical justification for that widespread estimation method. This entry begins with a brief historical account of the Gauss–Markov theorem and then offers a review of least squares before undertaking an exploration of the Gauss–Markov theorem, including extensions of the theory as well as some of its limitations.

Historical Account

Gauss is often credited with laying the bases of the method of least squares in 1795, at the age of 18 years. French mathematician Andrian-Marie Legendre, however, was the first to publish them, in 1806, in a nonstochastic curve-fitting context. In 1809, then again in his 1823 work *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, Gauss develops the method in the statistical context considered here and provides a first proof of the key result on BLUE. In 1900, Markov rediscovers the same result and includes it in his 1912 book on probability theory, translated into German as *Wahrscheinlichkeitsrechnung*. Still in 1912, English mathematician R. A. Fisher turns least squares into a general estimation method. The term *Markov theorem* was favored by Polish mathematician Jerzy Neyman, which eventually led to the now-standard *Gauss–Markov* appellation.

Least Squares

An observed random vector **Y** = (Y_1, …, Y_n)′ is said to satisfy the classical assumptions of the general linear model of full rank if there exists an $n \times k$ matrix:

$$= \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix},$$

with rank $k < n$, of real constants (the regressors or covariates), a $k \times 1$ vector β = (β_1, …, β_k)′ of real parameters (the regression coefficients), and an unobservable $n \times 1$ random vector **e** = (e_1, …, e_n)′ (the errors) such that

1. **Y** = **X**β + **e**,
2. E[**e**] = **0** (hence E[**Y**] = **X**β), and
3. the covariance matrix of **e**, which coincides with the covariance matrix of **Y**, is of the spherical form Var(**e**)(= Var(**Y**)) = σ^2**I** for some unspecified $\sigma > 0$.

Under these assumptions, it is well known that the OLS of β is

$$n$$

$$\hat{\beta} := \operatorname{argmin}_{\mathrm{b} \in Rk} X(Y_i - (x_{i1}, \ldots, x_{ik})\mathbf{b})^2$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

$$i = 1$$

with expectation (unbiasedness) and full-rank covariance matrix . Note that is a linear transformation (with matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$) of the vector of observations **Y**; invertibility of **X**′**X** follows from the assumption that **X** has full rank k.

If (ii)–(iii) are reinforced into the assumption that **e** is multinormal, with mean **0** and covariance $\sigma^2\mathbf{I}$, with unspecified $\sigma > 0$, then moreover is the (Gaussian) maximum likelihood estimator (MLE) of β, and, being a linear function of the Gaussian vector **Y**, is itself multinormal with mean β and covariance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Furthermore, it can be shown that (under the above Gaussian assumptions) is the (almost surely unique) uniformly minimum variance unbiased estimator of β. More precisely, among all unbiased estimators of β (i.e., among all estimators such that for all β), has, irrespective of the actual value of β, the smallest variance—in the sense that the difference is positive semidefinite irrespective of the actual value of β.

Gaussian assumptions, however, are unrealistic in most applications; when they do not hold, uniformly minimum variance unbiased estimators typically do not exist. The OLS estimator nevertheless still enjoys a weaker form of optimality, the nature of which is described by the Gauss–Markov theorem.

Gauss–Markov Theorem

In its traditional version, the Gauss–Markov theorem states that, under the aforementioned Assumptions (i)–(iii), the least squares estimator given in (1) is a BLUE, in the sense that, for any estimator that is linear (of the form for some nonrandom $k \times n$ matrix **A**) and unbiased (such that irrespective of the actual value of β), the difference is positive semidefinite.

That result implies that, for any linear combination $\mathbf{x}'\beta$, $\mathbf{x} \in \mathrm{R}^k$, of $\beta_1, \ldots, \beta_k$, a uniformly minimum variance linear unbiased estimator is (sometimes called best linear unbiased predictor or BLUP). This is the case, for instance, for the expected value $(x_n+_{1,1}, \ldots, x_n+_{1,k})\beta$ of an additional observation Y_{n+1} to be made under covariate values $\mathbf{x}' = (x_{n+1,}\ 1, \ldots, x_{n+1,k})$. In particular (letting $\mathbf{x} = \mathbf{u}_j$, the jth unit vector in the canonical basis of R^k), the variance of is smaller, for any $1 \leq j \leq k$, than that of any other linear unbiased estimator of β_j.

Extensions and Limitations

Several extensions of the Gauss–Markov theory have been proposed in the literature, mainly in econometrics and the social sciences, which relax Assumptions (i)–(iii). Several of them are briefly discussed in the following sections.

Stochastic Regressors

The traditional assumptions considered so far are treating **X** (often called the design matrix) as a matrix of constants. Such an assumption, in general, is fine in experimental sciences, where designs are under the experimenter's control. The situation is different in social sciences and in econometrics, where the matrix **X** of regressors also has to be considered random, and **X** and **e** moreover need not be mutually independent. Assumptions (i)–(iii) then can be replaced with (i)′ $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, where **X** is a random matrix with unspecified distribution, but for the fact that the probability that it has full rank k (with $k < n$) is one; (ii)′ conditionally on **X**, the error **e** has expectation almost surely zero, that is, $\mathrm{E}[\mathbf{e} \mid \mathbf{X} = \mathbf{x}] = \mathbf{0}$, except perhaps for a set of **x** values of probability zero; and (iii)′ the covariance matrix of **e** conditional on **X**, which coincides with the covariance matrix of **Y** conditional on **X**, is (**X** almost surely) of the spherical form $\mathrm{Var}(\mathbf{e} \mid \mathbf{X})(= \mathrm{Var}(\mathbf{Y} \mid \mathbf{X})) = \sigma^2\mathbf{I}$ for some unspecified $\sigma > 0$.

In this context, an estimator of β is (conditionally) linear if it is of the form $\mathbf{C}(\mathbf{X})\mathbf{Y}$, where $\mathbf{C}(\mathbf{X})$ is a $k \times n$ matrix of functions of **X**. The existence of BLUEs for β can be examined either for conditional unbiasedness, or for the much weaker concept of unconditional unbiasedness. An estimator is conditionally unbiased if , **X** almost surely; conditional unbiasedness clearly implies, but is not implied by, unconditional unbiasedness, which only requires . It is easy to see that, under Assumptions (i)′–(iii)′, the OLS estimator (1) is best linear conditionally unbiased, in the sense that, for any linear conditionally unbiased , the difference is **X** almost surely positive semidefinite. It immediately follows, of course, that is also positive semidefinite. The Gauss–Markov property of OLS estimators thus essentially survives under stochastic regressors and Assumptions (i)′–(iii)′, within the class of conditionally unbiased estimators. The case of best linear unconditionally unbiased estimators is more delicate, and we refer to Juliet Popper Shaffer's article, in 1991, in *The American Statistician*, "The Gauss-Markov Theorem and Random Regressors," for a complete treatment.

Nonspherical Errors

Whether unconditional or conditional, Assumptions (iii) and (iii)′ both treat errors as spherical, that is, the same (conditional) variance for all e_i's, $i = 1, \ldots, n$ (conditional or unconditional homoscedasticity) and no (conditional) covariance between e_i and e_j for all $i = 6 = j = 1, \ldots, n$ (conditionally or unconditionally uncorrelated errors); note that i.i.d.-ness of the e_i's is not required. Though classical in econometrics, the terminology used here is slightly improper; strictly speaking, a *spherical distribution* should be rotation invariant, whereas we only require a spherical covariance structure here.

That sphericity assumption also can be relaxed, but the OLS estimator, which treats all observations equally, then should be replaced by the so-called weighted least squares (WLS) estimator. Let Assumption (iii) be replaced by: (iii)″ the covariance matrix of **e** is of the form Var(**e**) = σ^2**H**, where **H** is a known positive definite matrix, and $\sigma > 0$ remains unspecified.

The Gauss–Markov Theorem then holds, ne varietur, for the WLS estimator:

$$\hat{\beta}H := (X'H - 1X) - 1X'\ H - 1Y.$$

In practice, however, **H** is seldom known and depends on some unspecified parameter θ to be estimated. Generalized least squares methods have been proposed for a variety of cases, including heteroscedasticity, equicorrelation, autoregressive errors, seemingly unrelated regression models, and so on. For an extensive treatment of those cases, refer to *Generalized Least Squares*, authored by Takeaki Kariya and Hiroshi Kurata in 2004.

Endogeneity

Violations of Assumption (ii)′, in the case of random regressors, are by far more serious than those of Assumption (iii)′. Such violations occur when the regressor is correlated with the errors—a situation that econometricians describe as endogeneity. When the regressors, or some of them, are endogenous—this may occur for various reasons such as omitted (possibly, unobservable) regressors, so-called "reverse causality" effects, selection bias, or measurement errors on the regressors—traditional estimators are no longer unbiased or consistent. Instrumental variable regression and two-stage least squares methods provide a way to handle such problems—though with the delicate problem of choosing the right instruments.

Multivariate Response (Multiple-Output Regression)

All previous developments extend, mutatis mutandis, to the case of a p variate response ($p \geq 1$), with Assumptions (i)–(iii) replaced by:

1. ‴$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)' = \mathbf{XB} + \mathbf{E}$, where $\mathbf{Y}$ is an observed $n \times p$ matrix, $\mathbf{X}$ (the design matrix) still is a full-rank $n \times k$ matrix of constants, the parameter $\mathbf{B}$ is a $k \times p$ matrix,
2. ‴the rows of the error matrix $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)'$ are mutually orthogonal, with mean $\mathbf{0}_p$; and
3. ‴covariance $\sigma^2\mathbf{H}$, for some unspecified $\sigma > 0$.

A Gauss–Markov theorem then holds for the WLS estimator:

$$\hat{B}H := (X' H - 1X) - 1X' H - 1Y$$

(a $k \times p$ matrix; the covariance matrices to be considered in the BLUE property here are those of the vectorized estimators resulting from stacking the columns of on top of each other).

Putting the Theorem Into Context

The Gauss–Markov theorem is traditionally invoked as an optimality property justifying the application of ordinary or WLS estimation methods in linear models under possibly non-Gaussian conditions, and we have reviewed some of its extensions. One should not, however, overemphasize its importance. The fact that least squares are optimal within the class of linear unbiased estimators indeed is a consequence of the severity of the restrictions imposed on that class of estimators, at least as much as a reflection of the good performances of least squares. Linear unbiased estimators indeed are but weighted averages of the observations. When the observations all have the same variance, it is not overly surprising that the smallest variance is obtained by putting equal weights on all (squared) deviations.

Marc Hallin

See also Analysis of Covariance; Estimation Bias; Matrix Algebra; Variance

Further Readings

Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist’s companion. Princeton, NJ: Princeton University Press.

Green, W. H. (1997). Econometric analysis (3rd ed.). New York, NY: Prentice Hall.

Hayashi, F. (2000) Econometrics. Princeton, NJ: Princeton University Press.

Judge, G., Hill, C., Griffiths, W., & Lee, T. (1985). The theory and practice of econometrics. New York, NY: Wiley.

Kariya, T., & Kurata, H. (2004). Generalized least squares. New York, NY: Wiley.

Lehmann, E. L., & Casella, G. (1998). Theory of point estimation. New York, NY: Springer.

Shaffer, J. P. (1991). The Gauss-Markov theorem and random regressors. The American Statistician, 45, 269–273.

Stigler, S. M. (1981). Gauss and the invention of least squares. Annals of Statistics, 9, 465–474.

Stigler, S. M. (1986). The history of statistics: The measurement of uncertainty before 1900. Cambridge, MA: Belknap Press of Harvard University Press.

Stock, J. H., & Watson, M. M. (2011). Introduction to econometric (3rd ed.). Harlow, UK: Pearson Education.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. Cambridge, MA: MIT Press.

Patricia A. Jenkins Patricia A. Jenkins Jenkins, Patricia A.

Gender and Testing

Gender and testing

723

724

Gender and Testing

Analysis of standardized test results has revealed consistent differences in average levels of performance between different groups, including differences in the performance of males and females. These are referred to as achievement gaps when one group outperforms another and the difference in average scores for the two groups is statistically significant; the differences between the scores of males and females are often referred to as the gender gap. This entry describes differences between average scores of males and females on certain standardized tests and discusses some of the possible reasons for those differences.

A 2013 report on the results of the National Assessment of Educational Progress long-term trend assessments of 9-, 13-, and 17-year-old students found that female students scored higher in reading than male students at all three ages. The National Assessment of Educational Progress long-term trend assessments show some narrowing of the gender gap over time, however. While 9 year olds overall had higher scores in 2012 than their counterparts in 1971, 9-year-old boys made larger score gains than girls, leading to a narrowing of the gender gap at age 9. In mathematics, 17-year-old male students scored higher than 17-year-old female students did. However, the gender gap in math for students at that age narrowed between 1971 and 2012 because 17-year-old female students made gains in math during that period while 17-year-old male students did not. National Assessment of Educational Progress data indicate that achievement gaps based on income and race are larger than the achievement gap between males and females.

Another area where gender differences have traditionally been found is in standardized tests that are meant to predict college performance. On the SAT, for example, males have typically scored around a third of a standard deviation example, males have typically scored around a third of a standard deviation higher than females on the mathematics portion. Females, on the other hand, often are found to score higher on verbal tests than males. In recent years, these differences have been found more consistently.

There is a variety of theories to explain gender differences in performance on standardized tests. Because females tend to get higher grades at all levels of education than males, it seems likely that one explanation may lie in the tests or the testing contexts themselves. Those who explore gender differences in attitudes toward science, technology, engineering, and mathematics find that by high school, girls are less likely to be interested or motivated in math and related areas. It is reasonable, then, to assume that girls will have different math backgrounds and levels of interests in mathematics by the time they take college admissions tests. Another possible explanation is that there may be gender-specific ways of thinking or cognitive ability differences; however, in terms of testing, educational researchers generally have not claimed that score differences are caused by actual ability differences.

The impact of gender gaps crosses several issues, such as dropout rates, graduation rates, higher education admissions, and earned degrees. In general, there is a higher dropout rate and a lower graduation rate among males than among females. In addition, females surpass males in college enrollment, especially among Hispanics and Blacks. As of 2016, 57% of students in U.S. degree-granting postsecondary institutions were female. The disparities between males and females on these issues are wider among those in certain ethnic groups and socioeconomic levels.

Some researchers claim that brain differences account for differences between males and females in learning, but this research is often disputed and some theories are not supported by evidence from brain studies. For instance, a 1982 article published in *Science* indicated that the corpus callosum, which links the right and left hemispheres of the brain, is proportionately larger in females than males. This finding has been cited in discussions of differences in boys' and girls' learning; however, multiple subsequent studies have found no significant difference in the size of the corpus callosum between males and females either in children or in adults.

Differences in early brain development between boys and girls are thought to have educational implications, but brain development is influenced by the environment. This makes it difficult to separate the roles of biology and socialization in differences found between how male and female students behave socialization in differences found between how male and female students behave in the classroom and perform on academic tasks. For example, there is some research indicating that females perform more efficiently than males when switching rapidly between tasks. In addition, male students tend to overestimate their academic abilities while female students underestimate theirs. Differences such as these that affect classroom interactions and learning in turn could have an influence on test performance.

Patricia A. Jenkins

See also African Americans and Testing; Asian Americans and Testing; Dropouts; Latinos and Testing; National Assessment of Educational Progress; SAT; Standardized Tests; STEM Education

Further Readings

Blatchford, P., Pellegrini, A. D., & Baines, E. (2016). The child at school: Interactions with peers and teachers (2nd ed.). East Sussex, UK: Routledge.

Eliot, L. (2009). Pink brain, blue brain: How small differences grow into troublesome gaps—and what we can do about it. New York, NY: Houghton Mifflin Harcourt.

Fine, C. (2010). Delusions of gender: How our minds, society, and neurosexism create difference. New York, NY: W. W. Norton.

Fine, C., Jordan-Young, R., Kaiser, A. & Rippon, G. (2013). Plasticity, plasticity, plasticity … and the rigid problem of sex. Trends in Cognitive Sciences, 17(11), 550–551. Retrieved from http://dx.doi.org/10.1016/j.tics.2013.08.010

King, K., & Gurian, M. (2006). With boys in mind: Teaching to the minds of boys. Educational Leadership, 64(1), 56–61. Retrieved from http://www.ascd.org/publications/educational-leadership/sept06/vol64/num01/Teaching-to-the-Minds-of-Boys.aspx

Lopez, M., & Barrera, A. (2014). Women's college enrollment gains leave men behind. Pew Research Center. Retrieved from http://www.pewresearch.org/fact-tank/2014/03/06/womens-college-enrollment-gains-leave-men-behind/

National Center for Education Statistics. (2013, June). The nation's report card: Trends in academic progress 2012. Retrieved from https://nces.ed.gov/nationsreportcard/pubs/main2012/2013456.aspx

Ready, D. D., LoGerfo, L. F., Lee, V. E., & Burkam, D. T. (2005). Explaining girls' advantage in kindergarten literacy learning: Do classroom behaviors make a difference? Elementary School Journal, 106(1), 21–38.

Stoet, G., O'Connor, D. B., Conner, M., & Laws, K. R. (2013). Are women better than men at multi-tasking? BMC Psychology, 1, 18. doi:10.1186/2050–7283–1-18

Whitmire, R., & Bailey, S. M. (2010). Gender gap: Are boys being shortchanged in K–12 schooling? [Special Section]. Education Next, 10(2). Retrieved from http://educationnext.org/gender-gap/

Bruce B. Frey Bruce B. Frey Frey, Bruce B.

Generalizability

Generalizability

724

725

Generalizability

Generalizability is the degree to which the results of a research study reflect what the results would be “in the real world,” with another sample of participants or with the variables operationalized in other ways. In other words, research results are generalizable when the findings are true generally speaking in most contexts with most people most of the time.

In the classic quantitative research framework of experimental design, researcher design theorists such as Thomas Cook and Donald Campbell have emphasized *external validity* as a necessary criterion for concluding that research results are generalizable. Threats to external validity include how a sample was selected from the broader target population to which one wishes to generalize, the situational specifics of the experimental manipulations, and the measurement choices made when assessing the independent and dependent variables. Generalizability is optimized when samples are chosen randomly, the research environment and researcher behaviors are carefully controlled so as not to affect the outcome, and constructs are defined and measured in ways that validly and reliably represent the broad ways that variables operate.

In the qualitative research framework, there is a somewhat different understanding of generalizability. Although some qualitative researchers argue that it is inappropriate to assume that generalizability is even an appropriate goal of social science research, there are some generally accepted generalizability criteria if one wishes to understand research results in a wider context. However, qualitative researchers are often more interested in vertical generalization, the extent to which research findings add to building or understanding theory, than they are interested in horizontal generalization, the more traditionally they are interested in horizontal generalization, the more traditionally quantitative wish to conclude that there would be similar results with another sample drawn from the same population. The qualitative framework known as grounded theory, for example, is more focused on whether theory that has been induced from the data collected is a fair representation of the data than whether a sample of participants is a fair representation of some abstract population.

Bruce B. Frey

See also Grounded Theory; Qualitative Research Methods; Quantitative Research Methods; Random Selection; Threats to Research Validity

Further Readings

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design … analysis issues for field settings. Chicago, IL: Rand McNally College.

Morse, J. M. (1997). Completing a qualitative project: Details and dialogue. Thousand Oaks, CA: Sage.

John D. Hathcoat John D. Hathcoat Hathcoat, John D.

Oksana Naumenko Oksana Naumenko Naumenko, Oksana

Generalizability Theory Generalizability theory

725

730

Generalizability Theory

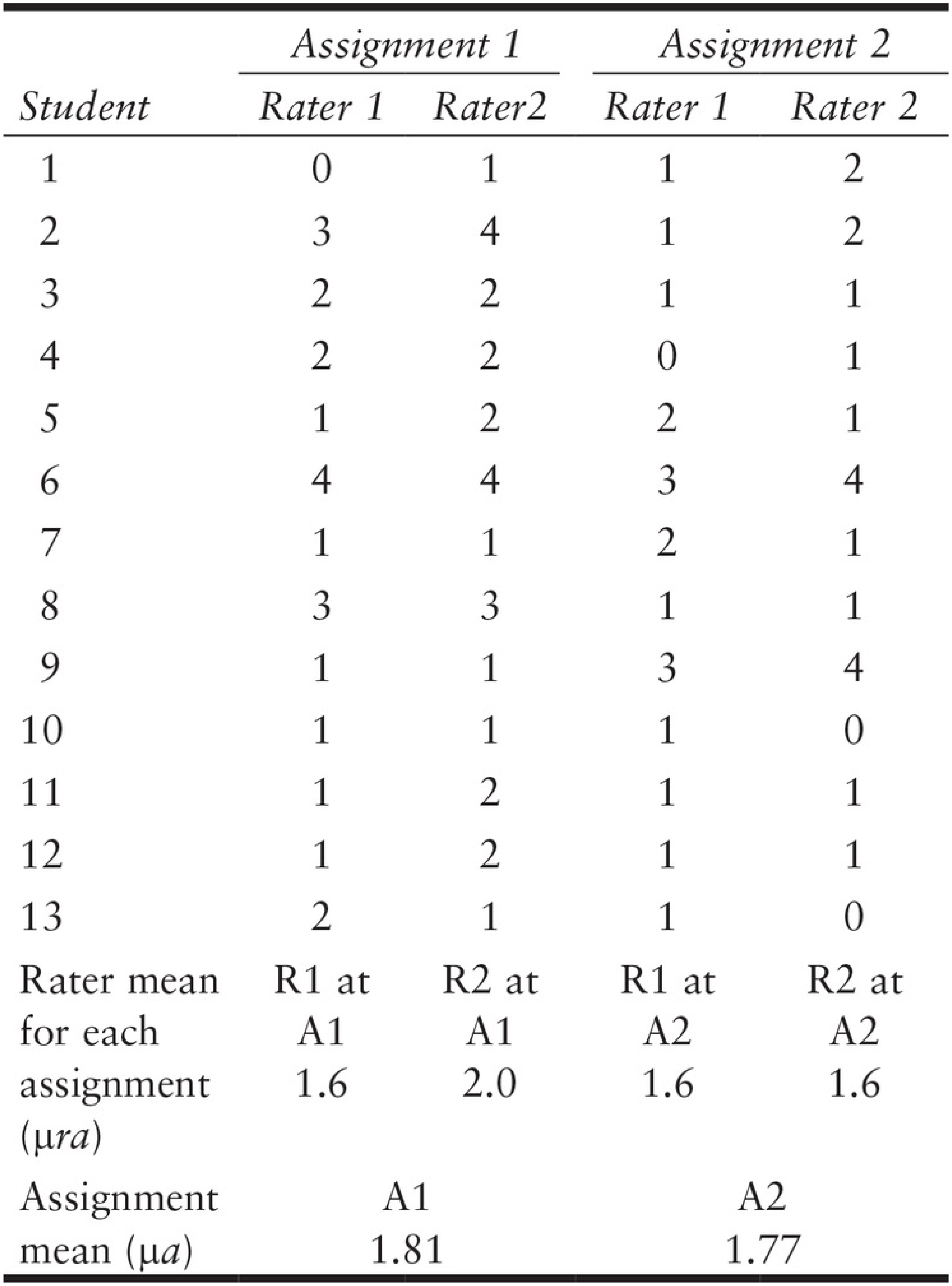
Educational researchers are often interested in making inferences from what may be considered observable to that which is unobserved. Responses to items on a multiple-choice test, an argumentative essay, and other overt behaviors (e.g., number of times a child raises her hand in a classroom) are *observable*. *Unobserved* variables on the other hand are used by educational researchers to explain patterns in observations. Intelligence, personality, aptitude, and critical thinking cannot, strictly speaking, be directly observed. Such variables refer to theoretical attributes that are at best indirectly investigated. For example, a researcher may hypothesize that differences in critical thinking (i.e., unobserved) account for why some students have higher scores than others on an assignment (i.e., observable). The extent to which it is reasonable to conclude that observed scores reflect critical thinking is a validity issue. Measurement error constrains the validity of score-based interpretations.

Measurement may be defined as the systematic assignment of numerals according to a set of rules. Measurements may distinguish mutually exclusive categories (e.g., ethnic groups), rank-order observations (e.g., high school class rank), or indicate differences in magnitude (e.g., degrees in Fahrenheit). Reliability assessment, traditionally conceived, aims to quantify the consistency of scores in a population whereas measurement error reflects random inconsistencies. Generalizability theory—hereafter referred to as G theory—provides a framework for investigating the extent to which distinct sources of error influence the precision of scores obtained from a measurement procedure.

The basic concepts of G theory, such as variance decomposition, universe scores, and facets of measurement, are introduced in this entry using a scores, and facets of measurement, are introduced in this entry using a hypothetical example. This is followed by discussing simple extensions in measurement design employed within G theory, such as whether a facet is treated as fixed or random. Finally, the entry concludes by summarizing the strengths and limitations of G theory when compared to traditional approaches for assessing reliability.

Universe Scores and Facets of Measurement Error

Assume a researcher sampled thirteen students to assess their critical thinking. Each student has submitted two assignments with each assignment scored by the same two raters. Possible scores range from 0 to 4 with higher values indicating greater critical thinking (see Table 1). Students are considered the *object of measurement* because the researcher aims to use this procedure to differentiate students according to their level of critical thinking. Raters and assignments are sources of imprecision or error. For example, it is unlikely that each rater will provide the same score to a student for a single assignment. Even if raters perfectly agreed about scores for one assignment, it is unlikely that the student would receive the same critical thinking score across multiple assignments. Given such possibilities, what would be the best estimate of a student’s critical thinking?



	Assignment 1		*Assignment 2*	
Student	*Rater 1*	*Rater2*	*Rater 1*	*Rater 2*
1	0	1	1	2
2	3	4	1	2
3	2	2	1	1
4	2	2	0	1
5	1	2	2	1
6	4	4	3	4
7	1	1	2	1
8	3	3	1	1
9	1	1	3	4
10	1	1	1	0
11	1	2	1	1
12	1	2	1	1
13	2	1	1	0
Rater mean for each assignment (μ*ra*)	R1 at A1 1.6	R2 at A1 2.0	R1 at A2 1.6	R2 at A2 1.6
Assignment mean (μ*a*)	A1 1.81		A2 1.77	

Note: Critical thinking ranges from 0 to 4 with higher scores indicating more critical thinking. Values indicate *i*th person's score provided by *r*th rater on the *a*th Assignment. R1 at A1 = mean of rater 1 on Assignment 1; R2 at A2 = mean of rater 2 on Assignment 2; R1 at A2 = mean of rater 1 on Assignment 2; R2 at A2 = mean of rater 2 at Assignment 2. A1 = overall mean for Assignment 1; A2 = overall mean for Assignment 2.

According to G theory, our best estimate of a student’s score is something like an average, or expected value, across all possible raters and assignments. This expected value is referred to as the *universe score*. Although it is not possible to observe all possible raters and assignments, we can attempt to generalize from observed scores to universe scores. G theory provides a framework to that end.

Measurement error hinders our ability to generalize from observed scores to universe scores. Error may arise from multiple sources such as measurement occasion, test form, and/or raters. Each source of error is referred to as a *facet of measurement* and serves to frame a *universe of admissible observations*. In this example, we may choose to view raters as admissible if they hold a bachelor degree in a given discipline, whereas assignments may be viewed as admissible if they have specific characteristics, such as a minimum page length and being consistent with a specific genre. With respect to our universe of admissible observations, we are willing to accept any rater being paired with any possible assignment that meets these definitions as an acceptable measurement condition. Decomposing observed score variance into distinct sources of error allows us to identify problematic facets of measurement error in this situation.

Decomposing Scores Into Sources of Variation

G theory employs analysis of variance concepts to partition variation in observed scores into different sources of measurement error. Facets of measurement are therefore akin to “factors,” or independent variables, in analysis of variance designs. A comprehensive examination of variance decomposition is beyond the scope of this entry; however, a conceptual overview of how this is accomplished is provided in Table 2. Each score can be perfectly reproduced by decomposing it, or “breaking it down,” according to different sources of variation.

Observed Score	*Effect*	*Interpretation of Variance Component*	*% of Total*
$X_{pra} =$	μ	Grand mean	–
	$+ \mu_p - \mu$	Differences in universe scores from grand mean	23.50%
	$+ \mu_r - \mu$	Differences in rater stringency/severity	2.20%
	$+ \mu_a - \mu$	Differences in assignment difficulty	0.50%
	$+ \mu_{pr} - \mu_p - \mu_r + \mu$	Extent to which rater severity differs across students	4.30%
	$+ \mu_{pa} - \mu_p - \mu_o + \mu$	Extent to which rank order of students changes across assignment	51.40%
	$+ \mu_{ra} - \mu_p - \mu_o + \mu$	Extent to which rater severity differs across raters	<0.00%
	$+ X_{pra} - \mu_p - \mu_r - \mu_o + \mu$	Three-way interaction between persons, raters, and occasion, confounded other sources of error	18.1%

Note: X_{pra} = person's score provided by a rater on a single occasion; μ_p = person mean; μ_r = rater mean; μ_a = assignment mean; μ_{pr} = mean of person for a rater; μ_{pa} = mean of person on assignment; μ_{ra} = rater mean on each assignment; % of total is based on data in Table 1.

Consider the score of 0 assigned to the first student for the initial assignment depicted in Table 1. Given the design of the study, this score can be reproduced by partitioning it into three main effects (i.e., persons, raters, and assignments) and all possible interactions (i.e., Person × Rater, Person × Assignment, Rater × Assignment, and the three-way interaction between each variable which is confounded with unidentified sources of error).

G theory capitalizes upon this reasoning to estimate the magnitude of variance components reflecting each source of error as well their interactions. Stated differently, variance in our object of measurement is partly due to students having different universe scores; however, this variation may also reflect measurement error. Estimating variance components by decomposing scores into distinct sources allows us to examine potentially problematic sources of error. Once identified, strategies can be employed in subsequent research to minimize the influence of particular sources of error in a measurement procedure.

G Studies, D Studies, and Types of Decisions

Two types of studies are discussed within G theory: (1) generalizability studies (i.e., G studies) and (2) decision studies (i.e., D studies). Although each type of study has a different aim, in practice, both studies are usually conducted using the same data. G studies estimate the magnitude of variance components attributable to different sources of error. D studies, as the name implies, use information from G studies to estimate reliability-like coefficients and to make decisions about optimal measurement designs in subsequent research. Stated differently, D studies use information from G studies to make inferences back to a *universe of generalization*. A universe of generalization is defined as the set of measurement conditions across which a researcher aims to generalize. In the example given earlier, the universe of generalization will be defined by both facets of error because we aim to generalize to unobserved, though theoretically exchangeable, samples of two raters and assignments. In this G study, 23.5% of the total variance is attributed to persons (Table 2). Because persons constitute our object of measurement, we wish for this value to be relatively large. A lack of person variance would be a cause for concern because it suggests that the measurement procedure is, at least for all practical purposes, aiming to detect miniscule differences in our object of measurement.

Potentially problematic facets of measurement error can be identified by examining their magnitude relative to the total variance in scores. The Person × Assignment interaction constitutes the largest source of variance in this example, consisting of approximately 28% of the total variance (Table 2). This interaction indicates that judgments about which students have higher critical thinking scores tends to be inconsistent across each assignment. For example, when examining the first assignment, raters may believe that three students have the highest critical thinking scores, yet come to radically different conclusions when examining the second assignment. Raters are not as problematic as assignments given that only 4.3% of the total variance is attributable to the Rater × Person interaction. The rank ordering of students according to critical thinking is therefore fairly consistent across each rater.

The D study uses the variance components obtained in our G study to determine the extent to which they hinder our ability to make inferences from our observations to a universe of generalization. Two reliability-like coefficients can be estimated as part of a D study, though each coefficient assumes that a different type of decision is being made about students. A G coefficient is estimated when a researcher is interested in making relative decisions, whereas a dependability, or ϕ coefficient, is estimated when making absolute decisions.

Relative decisions pertain to rank ordering an object of measurement (e.g., some students demonstrate higher levels of critical thinking than others). Absolute decisions aim to locate an object of measurement on a scale irrespective of relative standing (e.g., an individual has a score of 3 on critical thinking). Whether one examines a relative or absolute coefficient largely depends on the purposes of measurement. Absolute decisions may be more important than relative decisions in situations where students must achieve a particular standard regardless of how other students perform. If students were administered a driving exam, for example, absolute decisions are more critical than relative decisions (e.g., we do not care if a person can parallel park better than most applicants if they crash into other vehicles).

Both coefficients estimate the proportion of variance that can be attributed to differences in universe scores—they depart, however, in which sources of error are used to calculate the total or observed variance. As applied in this example, relative decisions are impacted by sources of error that influence rank ordering of participants (e.g., interaction between persons and each facet of measurement). Absolute decisions are more difficult to make than relative decisions and are influenced by each source of error in this example. The G coefficient was .42, and the ϕ coefficient was .41. Both coefficients have a theoretical range from 0 to 1, with higher values indicating that a greater proportion of variance can be attributed to universe score variance. Caution should be used when applying rules-of-thumb to interpret the magnitude of these coefficients because what is viewed as acceptable depends upon the context of a study. With this being said, our estimates would be a concern in most situations. Not only does our measurement procedure have difficulties placing individuals on a scale (i.e., absolute), but it also has problems detecting which students are doing better than others in critical thinking (i.e., relative).

Although not shown here, the Person × Assignment interaction was a large source of error when making either relative or absolute decisions. This implies that an efficient use of resources may focus on improving assignments as opposed to rater training. Once reliability-like coefficients are estimated, additional D studies can be conducted to investigate how these coefficients would change given a different number of raters and/or assignments. In support of our interpretation, a G coefficient of ≈.80 may be obtained by changing our universe of generalization to include samples of 14–16 assignments per student. Equivalent increases in the number of raters failed to provide similar improvements to the G coefficient (e.g., the G coefficient increased from .42 to .47 when including 16 raters).

Basic Extensions in Measurement Design

Broadly speaking, there are two ways in which the hypothetical example presented in this entry could be extended. First, G theory makes a distinction between fixed and random facets of measurement; second, facets of measurement may either be crossed or nested within other facets of measurement. A facet may be treated as *random* when one obtains a random sample of each level of a facet from a defined universe. A facet may also be treated as random when a researcher is willing to treat each level of a facet as exchangeable with other possible levels. In this example, raters are a facet of measurement that consist of two observed levels—Rater 1 and Rater 2.

Raters may be viewed as a random facet if we are willing to treat our raters as a sample of possible raters with similar characteristics. Treating a facet as random implies that we want to generalize from our observed raters to a universe of possible raters. Raters, however, may also be viewed as a *fixed* facet. In this case, we treat our observation of both levels analogous to that of observing a population—we are only interested in making inferences about two raters and both raters have been observed. Fixing a facet usually results in higher reliability-like coefficients because it reduces the universe of generalization though this comes at the cost of being capable of making inferences beyond what has been observed.

A facet of measurement may either be crossed with, or nested within, one or more facets of measurement. Two facets are *crossed* if each level of one facet is observed with each level of a second facet. In this example, assignments are crossed with raters because both assignments were scored by the same two raters. One facet is *nested* in a second facet if levels of one facet are observed in only one level of another facet. For example, assume we had four raters assigned to two groups. One group of raters was assigned to score the first assignment, whereas the second group provided scores for the other assignment. In this case, raters are nested in assignments since each rater is observed in only one assignment.

A benefit of crossed designs is that it is possible to disentangle potential sources of error; conversely, some sources of error cannot be investigated when using a nested design. For example, if raters are nested in assignments then it is not possible to distinguish differences in assignment means from differences between groups of raters. Do the differences in assignment means simply occur because we used different groups of raters? Nested designs, however, have some advantages over crossed designs because they tend to result in higher reliability-like coefficients and in some cases may be needed to control for other methodological issues (e.g., practice effects and fatigue). For example, it may be unfeasible to have raters score 100 assignments for 13 students because it could introduce systematic forms of error attributed to fatigue or other issues.

Strengths and Limitations of G Theory

The strengths and limitations of G theory are addressed by comparing this approach to classical test theory (CTT). Under CTT, observed scores are the composite of true scores (i.e., defined as an expected value across repeated replications) and random error. CTT provides one reliability coefficient for each source of error, most commonly reflecting test or item characteristics (i.e., parallel forms and internal consistency), measurement occasions (i.e., test–retest), and interrater reliability. G theory extends these concepts in a way that overcomes many of the limitations of CTT. CTT not only employs more restrictive assumptions than G theory (e.g., parallel forms in CTT assumes equal means, variances, and covariances, whereas G theory assumes observations are randomly parallel), but CTT may oversimplify the complexity of educational measurement by failing to isolate how distinct sources of error, as well their interaction, differentially contribute to imprecision. An advantage of G theory resides in its capacity to address such complexities. Ironically, the complexity of G theory has arguably served to limit its use among educational researchers.

In specific circumstances, some reliability-like coefficients obtained from G theory and CTT will be equivalent. For example, an α coefficient, which is an indication of internal consistency, will be equal to a G coefficient in measurement designs focusing on a single facet of error (e.g., items) that is treated as random as opposed to fixed. Despite such circumstances, CTT fails to integrate the nuances of more complex designs in reliability calculations. With respect to the example in this entry, CTT would minimally estimate three separate reliability-like coefficients: (1) parallel forms by correlating student scores across each assignment, (2) an interrater reliability coefficient for the raters on Assignment 1, and (3) a third interrater reliability coefficient for Assignment 2.

Unlike G theory, which provides diagnostic information about sources of error, CTT fails to consider how each source of error may combine to impact imprecision, thus making it difficult for researchers to decide how to modify subsequent measurement procedures. CTT also assumes that the researcher is solely interested in relative, as opposed to absolute, decisions. In other words, CTT focuses on issues with rank-ordering objects of measurement, whereas G theory allows the researcher to examine how sources of error influence our capacity to place students on a scale irrespective of their relative standing.

As previously mentioned, the flexibility of G theory allows one to incorporate the complexities of a measurement procedure within an investigation of error. The researchers are free to construct their universe of generalization, determine whether facets are nested in others, decide which facets should be fixed or random, and specify what types of decisions are most relevant to the situation.

This flexibility provides a researcher with many options, though in some sense, this may have also resulted in making G theory less accessible to professionals outside of the measurement community. Issues with accessibility partly derive from the fact that the equations used to investigate intricate measurement designs are cumbersome for many people, which is further exacerbated by a relative dearth of user-friendly software. The recent development of free statistical programs, such as R packages and EduG, may partly rectify this issue by providing user-friendly software to researchers interested in capitalizing upon the many benefits of G theory.

John D. Hathcoat and Oksana Naumenko

See also Analysis of Variance; Classical Test Theory; InterRater Reliability

Further Readings

Brennan, R. L. (2001). Generalizability theory. New York, NY: Springer.

Cardinet, J., Johnson, S., & Pini, G. (2011). Applying generalizability theory using EduG. Abingdon, UK: Taylor … Francis.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability scores and profiles. New York, NY: Wiley.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer (Vol. 1). Thousand Oaks, CA: Sage.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: overview. Encyclopedia of Social Measurement, 2, 99–105.

Dan He Dan He He, Dan

Hongling Lao Hongling Lao Lao, Hongling

Generalized Linear Mixed Models

Generalized linear mixed models

730

734

Generalized Linear Mixed Models

The generalized linear mixed model (GLMM) is a statistical framework that broadens the traditional general linear model to include variables that are not normally distributed, relationships that are not strictly linear, and data that have dependency. The general linear model is the foundational statistical structure that includes almost all parametric statistical procedures such as linear regression and analysis of variance. Two of this model’s offshoots are the generalized linear model and the linear mixed model. This entry describes these three models and then explores their most flexible of offspring, the GLMM.

General Linear Model

The general linear model is useful in answering research questions about the impact of one or multiple predictor variables on an outcome variable. A predictor variable is a variable that is being manipulated (ideally) in an experiment to observe its impact on the outcome variable, whereas the outcome variable is a variable that is hypothesized to be changed by the predictor variable(s). Alternative names for the predictor variable are explanatory variable or independent variable. Likewise, the outcome variables are sometimes referred to as response variables or dependent variables.

Statistically, a general linear model represents conventional linear regression models with a continuous outcome variable predicted by one or more continuous and/or categorical variables. A general linear model includes a simple linear model, a multiple linear model, as well as the analysis of variance and the analysis of covariance. In a general linear model, the model can be expressed in the following equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_j x_{ji} + e_i.$$

The outcome variable y_i is modeled by a linear function of the predictor variables x_{ji}, plus an error term (e_i). The subscript i indicates the ith observation, whereas the subscript j indicates the jth predictor variable.

The word *linear* in a general linear model implies that a combination of parameters $\beta_0 \ldots \beta_j$ could predict the observed values of the outcome variable. The word *general* refers to the fact that the outcome variable is dependent on potentially more than one predictor variable and it is normally distributed.

The previous equation can be rewritten with a shortcut as:

$$y = \mathrm{X}\beta + \mathrm{e}.$$

Each bold letter is a matrix. The matrix $\boldsymbol{y}$ represents the values of the outcome variable for all N observations, with a size of $N \times 1$. The matrix $\boldsymbol{X}$ stores the values of all j predictor variables for all N observations, with a size of $N \times j$. The matrix β indicates the estimated values of the general linear model parameters, with a size of $j \times 1$. The matrix $\boldsymbol{e}$ shows the difference between the predicted and the observed value of the outcome variable (i.e., residual/error) for all N observations, with a size of $N \times 1$.

A general linear model is the most widely used statistical model, while it has to meet certain assumptions: (a) linearity, (b) data independency, and (c) the residuals are independent of each other and normally distributed, $e_i \sim N(0, \sigma^2)$. The linearity assumption means there is a linear relationship between the predictor variables and the outcome variable. In other words, a one-unit change in the predictor variable is expected to bring about the same amount of change in the outcome variable for all observations. The data independency assumption means each observation is independent of another.

Generalized Linear Model

A generalized linear model is one step rebellious from its parent, the general linear model. While keeping every other aspect the same, a generalized linear

linear model. While keeping every other aspect the same, a generalized linear model allows for more flexible sample space in the outcome variable, which is a violation of the linearity assumption in the general linear model.

A common problem in modeling is the mismatch of the sample space between the linear predictor and the outcome variable. A *sample space* is the range of all possible values of a variable. The linear predictor can take on any value from negative infinity to positive infinity. Yet, the outcome variable may not be so. For example, a probability of success as the outcome variable ranges from 0 to 1.

In the general linear model, the observed values of the outcome variable are predicted by the linear model. The observed values and the model-predicted values for the outcome variable are on the same infinity scale. However, in the generalized linear model, we have two scales on our hands: One is from our observed outcome variable, which is a nonnormal distribution and has boundary restriction, while the other is from the linear model, which predicts continuous values that range between (−∞, ∞). How can one incorporate two different scales into the model? The key is to first transform the observed outcome variable *y*'s true values into a new outcome variable *y*′ using a link function. Then one builds a model to predict the values of the transformed outcome variable *y*′.

Conceptually, a link function is any mathematical rule that specifies one type of data transformation. A link function is represented as *g*(·). It rescales the outcome variable *y* to match the scale of the linear combination of predictor variables, so that the model can predict the transformed outcome variable *y*′ linearly. On the contrary, an inverse link function *h* can convert the transformed *y*′ back into the original *y* scale.

$$gy = y' = X\beta + e,$$

$$hy' = y = h(X\beta + e).$$

For example, a link function could transform a dichotomous or discrete outcome variable with a restricted data scale into a continuous data scale ranging from −∞ to ∞. For example, a logit link function could transform data of binomial distribution ranging between (0, 1) to data of normal distribution ranging between (−∞, ∞). A logit function is the natural log of the odds of the outcome variable *y* equal to one category. The odds is the ratio of the probability of observing *y* equal to one category over the probability of observing *y* equal to the other category.

$$gy = y' = \log p1 - p,$$

$$hy' = y = e(p)1 + e(p),$$

$$p = X\beta + e.$$

Likewise, for an outcome variable that follows a Poisson distribution, a link function to transform the data from 0, ∞ to −∞, ∞ is needed, such as a logarithm link function. The exponential function serves as the inverse link function, transforming *y*′ back into *y*.

$$gy = y' = \log(X\beta + e),$$

$$hy' = y = e(X\beta + e).$$

Linear Mixed Model

The other parent of the GLMM is the linear mixed model. Similarly with the generalized linear model, the linear mixed model is one step more flexible than its parent, the general linear model. The new step in the linear mixed model is allowing data dependency, which is a violation of the data independency assumption in the general linear model.

The source of the dependence in the data can be modeled via a random effect. For example, in a common educational research situation, imagine that test scores of students from the same classroom are correlated due to sharing the same teachers and the same school environment. In other words, the student test scores are dependent and such dependence can be modeled in a mixed model.

A hypothesis usually specifies a speculated process that generates the outcome variable. It can contain two parts: a deterministic part and a stochastic part. The deterministic part, which is called the fixed effect, is the data generating process shared among the whole population. The stochastic part, which is called the random effect, is a data generating process that applies only to a specific subset random effect, is a data generating process that applies only to a specific subset of the population. The fixed effect speculates the general property in the population, serving as the population baseline. A fixed effect is sufficient for a statistical model. A random effect is an optional and additional layer of guessing. It speculates the extent of deviation of individuals away from population baseline.

A model with a fixed effect only is called a fixed model. A model with both fixed effects and random effects is called a mixed model. A fixed model is a special case of the mixed model, when the random effect equals zero. The combination of fixed effects and random effects makes up the linear predictor of the linear mixed model. It expresses the speculated process that generates the outcome variable in a mathematical format.

A linear mixed model can be represented with the following equation:

$$y = X\beta + Z\gamma + e.$$

The ***y*** is the column vector for the outcome variable, ***X*** is a matrix of the predictor variables, β is the column vector of the fixed effect coefficients, *Z* is the design matrix for the random effects, γ is the column vector of the random effects, and ***e*** is the column vector of the residuals.

Effect here means the impact of a predictor variable on the outcome variable. A general linear model estimates a specific parameter β_j for a predictor variable X_j to indicate this effect. The predictor variable as a factor might be grouped into different values or levels. The estimated β_j represents the common mean effect on outcome variable for all subgroups of this predictor variable. This is what we call fixed effect. Furthermore, if we care about influences/effects differences of the subgroups of X_j on the outcome variable and if we believe that the subgroup means deviate from the big group mean in a way that is not arbitrary but are sampled from a larger population and follow a distribution (usually Gaussian), it is where we start to consider adding random effect into the model. Once a model has both fixed effect and random effect, it becomes a mixed effect model.

The motivation of incorporating random effect into a model is to take into account both group-level and subgroup-level variation in estimating group-level effects. For repeated measures (a single unit is measured multiple times), a random effect model could appropriately model the correlation of data under each unit; for a partial grouping sample with few data points in certain groups, a random effect model allows us to “borrow” information from other subgroups random effect model allows us to "borrow" information from other subgroups having more data points to determine the appropriate coefficient for low-sample subgroups.

Take a two-level mixed model as an example. To keep it conceptually simple, we do not specify any predictor variables in the model.

$$y_j = \mu + \alpha_{j[i]} + e_i ,$$

where μ indicates the overall mean of the outcome variable, representing the overall mean. This is the fixed effect component of the mixed model. The overall mean could be predicted by a linear combination of predictor variables. The subscript *j* indicates subgroups and the subscript *i* indicates the *i*th observation in the sample.

This model includes two levels of variation: at the overall group level, $e_i \sim N$ (0, σ^2); and at the subgroups level, $\alpha_j \sim N$ (0, $\sigma\alpha^2$). The random coefficients are e_i and α_j. The $\alpha_{j[i]}$ is the random effect component of the mixed model, representing the subgroup mean deviation from the overall group mean. The random effect could be predicted by a separate linear combination, different from those in the fixed effect component. The *ei* is the residual, indicating the individual deviation from the subgroup mean. A mixed effect model does not estimate the individual value of the random coefficients for each observation from a model parsimony consideration. Instead, the variance of the random coefficients (i.e., σ^2 and $\sigma\alpha^2$) are estimated, which greatly reduce the number of parameters to be estimated. We could find the most likely values of these parameters through an approach of maximum likelihood.

GLMM

The GLMM inherits good genes from its parents. It can flexibly handle both nonnormal outcome variables and dependent data. For example, the outcome variable we care about might be binary data or categorical data, which follow a different distribution besides Gaussian. The range of the outcome variable will not be from negative infinity to positive infinity (−∞, ∞) like continuous data with a normal distribution but from zero to one (0, 1) or from zero to positive infinity (0, ∞). Sometimes the data are not independent. The observations might be clustered in different groups, like classroom, school, and district. They might also be repeated measures from one person. Residuals of the model are not random and independent of each other but dependent on into which group the observation falls.

The GLMM can be further developed as an extension for more complicated situations. The word “generalized” indicates the outcome variable in the model is a nonnormal distribution (such as binominal or Poisson distribution). “Mixed” refers to random effects as an addition to the fixed effects in a linear model to account for the dependency in the data. A GLMM could be expressed as the following equation:

$$gy = X\beta + Z\gamma + e,$$

where *g*(+) is the link function for the outcome variable; *y* is the outcome variable, an *N* × 1 vector of observations; *X* is *p* predictor variables of fixed effects, an *N* × *p* matrix; β is the parameters of fixed effects, a *p* × 1 vector; *Z* is *q* predictor variables of random effects, an *N* × *q* matrix; γ is parameters of random effects, a *q* × 1 vector; and *e* is residuals that cannot be explained by the model, an *N* × 1 vector.

Applications of the GLMM

Generally speaking, there are a few steps in applying any statistical model. Based on the data type and research question, we first select a model. Then we specify the model either in mathematical terms for conceptual understanding or in programming syntax, as the modeling requires the power of computers for estimation. If the model converges at the estimation step, we can further evaluate whether the model fits well with the data. If so, we can advance to interpret the results and make inferences. If not, we have to go back to check for reasons and start over.

Issues in Making Inferences

Model Scale Versus Data Scale

This issue applies to models with non-Gaussian (nonnormal) outcome variables, the generalized models. For models only with Gaussian outcome variables, they are called general models. General models are special cases of the generalized models. A link function is used in generalized models to solve the sample space mismatch issue, as previously described.

mismatch issue, as previously described.

Two related concepts need to be clarified, which are *model scale* and *data scale*. The model scale is in the sample space of the linear predictor. This is because the model specification and estimation are on the same scale of the linear predictor. The data scale is in the sample space of the outcome variable. For parameter interpretation, the results need to be transformed from the model scale back into the data scale, in order to be meaningful. A link function is used to extend the sample space of the outcome variable (data scale) so as to match that of the linear predictor (model scale). On the contrary, an inverse link function is used to transform the estimated parameters from the model scale back to the data scale. For example, a logit is a common link function that extends limited sample space of the binary outcome variable to match that of the infinity sample space of the linear predictor. A logit is the log odds of a probability.

If the two model scales are the same as the data scale, no transformation is needed. This is a special case of the link function. In such cases, an identity link function is enough, which makes no transformation at all. There is a common misunderstanding due to the existence of two sample spaces while making inferences. The model scale and the data scale may not be the same. It is important to keep them separate and choose the appropriate scale for different purposes.

Broad Inference Versus Narrow Inference

This issue occurs only for mixed models with random effects. Making inferences based on the fixed effect only is called the broad inference. It is a general inference on the population level. Making inferences based on both the fixed effect and random effect is called the narrow inference. It is a specific inference on the subset level of the population.

Marginal Distribution Versus Conditional Distribution

This issue occurs for a non-Gaussian outcome variable in mixed models with random effects. This happens when the marginal distribution of the outcome variable is different from the conditional distribution of the outcome variable.

Dan He and Hongling Lao

See also Analysis of Variance; Mixed Model Analysis of Variance; Multiple Linear Regression

Further Readings

Stroup, W. W. (2013). Generalized linear mixed models: Model concepts, methods, and applications. Boca Raton, FL: CRC Press.

Jonathan Wai Jonathan Wai Wai, Jonathan

Harrison J. Kell Harrison J. Kell Kell, Harrison J.

Giftedness

Giftedness

734

735

Giftedness

What the term *giftedness* “really” means has been discussed for centuries. There are many domains in which one can be considered gifted, including sports, music, and intellectual pursuits. However, when discussing intellectual giftedness, general intelligence or cognitive ability likely plays a central role in any definition. Experts often differ on verbal definitions, and beyond general ability, there are specific domains of cognitive ability such as verbal, math, and spatial. With full acknowledgment that definitions of giftedness vary widely, this entry focuses on a quantitative definition of intellectual giftedness as a variable on a continuum, focusing on what is measurable—specifically discussing general cognitive ability (often indexed by IQ) and its connection to a variety of outcomes.

When intellectually gifted students are identified when young and followed up later in life, research shows they attain doctorates, publications, patents, higher income, and even university tenure at rates well above the general population. Examining countries as the unit of analysis, research has also demonstrated that the intellectually gifted in each country disproportionately impact innovation and even gross domestic product of that country. Many times, when individuals who have achieved high standing in their professions are retrospectively assessed to see whether they were intellectually gifted when they were young, they turn out in fact to have been highly intellectually gifted. Combining these sources of data shows that intellectual giftedness is important in the development of expertise and even impacts broader societal creativity and innovation.

The idea that beyond a certain cut point—for example, an IQ of 120—that more ability no longer appears to make a measurable difference has also been shown to be false. Even within a group with an IQ of 137 and higher and within highly select groups of millionaires, billionaires, and Fortune 500 CEOs, more ability predicts higher rates of earning doctorates, publications, patents, tenure in higher education, income, network power, and net worth. More ability continues to have a payoff even at the highest levels. This does not, however, rule out the importance of many other factors, including hard work, education, interests, personality, and luck. It does show, however, that the intellectually gifted tend to be largely overrepresented among the people who lead and create modern society.

Intellectually gifted students, as a whole, not only turn out to be very high achieving but also are no different from the general population in terms of psychological well-being or adjustment. Research on very young gifted children and on adults who were identified as gifted when they were children shows that in terms of family, friendships, romantic relationships, and broad indicators of well-being, adjustment, and satisfaction, these individuals fare quite well. Of course, this does not mean that intellectually gifted people don’t face challenges. Many do, and without appropriate academic challenge and acknowledgement of the psychological issues they might face, this can potentially do harm.

However, at least within the United States, little funding is directed toward supporting the development of gifted children in K–12 public education, most likely because of the U.S. focus on equity rather than excellence. Because financially advantaged students can access talent development opportunities outside of school but financially disadvantaged students cannot, a deep divide has developed between resource-rich and resource-poor students, impacting the representation of talented but disadvantaged students in elite levels of society. This has consequences for diversity and inequality.

Research shows that, perhaps counterintuitively, when students are selected on the basis of parent and teacher nomination initially, many disadvantaged and minority students are not properly identified as being intellectually gifted. Further, research has indicated that universal testing works in identifying students systematically and matching them with appropriate educational programming. A 2015 study found that universal testing led to large increases in disadvantaged and minority students being placed in gifted programs.

Jonathan Wai and Harrison J. Kell

See also Ability Tests; Aptitude Tests; Cattell–Horn–Carroll Theory of Intelligence; *g* Theory of Intelligence; Intelligence Quotient; Intelligence Tests

Further Readings

Assouline, S. G., Colangelo, N., VanTassel-Baska, J., & Lupkowski-Shoplik, A. E. (Eds.). (2015). A nation empowered: Evidence trumps the excuses that hold back America’s brightest students. Iowa City, IA: The Belin-Blank Center for Gifted and Talented Education.

Benbow, C. P., & Stanley, J. C. (1996). Inequity in equity: How “equity” can lead to inequity for high-potential students. Psychology, Public Policy, and Law, 2, 249–292.

Card, D., & Giuliano, L. (2015). Can universal screening increase the representation of low income and minority students in gifted education? (Working Paper 21519). National Bureau of Economic Research. doi:10.3386/w21519

Jensen, A. R. (1998). The g factor: The science of mental ability. Westport, CT: Praeger.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. Perspectives on Psychological Science, 1, 316–345.

Makel, M. C., Kell, H. J., Lubinski, D., Putallaz, M., & Benbow, C. P. (2016). When lightning strikes twice: Profoundly gifted, profoundly accomplished. Psychological Science. doi:10.1177/0956797616644735

Stanley, J. C. (1995). Varieties of giftedness. Retrieved from http://files.eric.ed.gov/fulltext/ED392825.pdf

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. Psychological Science in the Public Interest, 12, 3–54.

Thompson, L. A., & Oehlert, J. (2010). The etiology of giftedness. Learning and Individual Differences, 20, 298–307.

Wai, J., & Worrell, F. C. (2016). Helping disadvantaged and spatially talented students fulfill their potential: Related and neglected national resources. Policy Insights from the Behavioral and Brain Sciences, 3, 122–128.

Brandon W. Youker Brandon W. Youker Youker, Brandon W.

Goal-Free Evaluation Goal-free evaluation

735

739

Goal-Free Evaluation

Goal-free evaluation (GFE) is an evaluation approach in which the evaluator conducts the evaluation without reference to predetermined goals and objectives. The rationale behind GFE is that the evaluator should examine all relevant outcomes; additionally, stated goals and objectives only represent a limited number of potential outcomes. Furthermore, if a program is achieving its goals, then the goals and objectives should be apparent, otherwise the goals are irrelevant or trivial. The goal-free evaluator is nearly always external to the program and its stakeholders to ensure ignorance of the stated goals; thus, the goal-free evaluator either intentionally avoids knowing the program’s goals or simply disregards them. Even so, frequently the evaluator has a general idea of the program goals such as with a substance abuse treatment program, whose goal is reduction of use or abstinence, but the specific objectives are not as evident. Therefore, some have argued that “goal free” is a misnomer and suggest the title “objectives-free evaluation.” This entry describes GFE’s development, principles, practice, benefits, and limitations.

GFE Development

Michael Scriven introduced GFE in 1972 somewhat in opposition to the goal-based evaluation (GBE) models that dominated—and still dominate—program evaluation practice. With GBE (also known as objectives-oriented evaluation), the evaluator judges the program according to the attainment of the program’s goals and objectives. However, Scriven noticed that product evaluators rarely ask the product designers or manufacturers what they intended to do; rather, the product evaluator examines the product, develops criteria and performance standards, and then tests the product all while disregarding the manufacturer’s specific goals or objectives. Consumer Reports’ evaluations epitomize this process. Scriven analogizes GFE to the double-blind pharmaceutical study in which the evaluator is not privy to outcomes that the drug is supposed to produce; rather, the evaluator searches for all relevant positive, negative, or neutral effects.

GFE Principles

According to Brandon Youker, there are four principles underlying GFE. The first principle is that the goal-free evaluator identifies relevant effects or outcomes without referencing the program goals or objectives. Second, the evaluator investigates what actually happened or is happening without the cuing of the goals and objectives. The third principle of GFE is one of attribution, which determines what occurred and whether there is a logical connection between the outcomes and the intervention. The fourth principle is determining the degree to which the outcomes are positive, negative, or neutral.

GFE in Practice

GFE is amenable and adaptable to use with many evaluation models and data collection methods. The only caveat is that the model does not require goal orientation; therefore, historically, goal-free evaluators combined GFE with other evaluation models such as CIPP, utilization-focused evaluation, success case method, constructivist evaluation, and connoisseurship, among others. Furthermore, goal-free evaluators have preferred qualitative data collection methods with their GFEs. Most GFEs have included methods such as interviews with impactees, direct observation of program activities, review of program documents, utilization of preexisting checklists, and administration of open-ended survey questionnaires. Yet, there is nothing that obligates the goal-free evaluator to use qualitative methods. In fact, a few GFEs have included the counting of program outputs, used quantitatively based surveys that ask stakeholders to rate the program on general program characteristics, and administered standardized psychometric and educational assessments, for example.

Two Types of GFEs

There are two main types of GFEs practiced. The first type of GFE is one in which the evaluator intentionally avoids the stated goals and objectives. The second type of GFE is one in which the specific evaluation model disregards the goals—and is thus goal free by default—but the evaluator does not use special or deliberate precautions to avoid goals.

A critical operation with the intentional GFE is screening the goals and objectives from the evaluator. With this type of GFE, the evaluator intentionally avoids knowing the program's goals and objectives by having a third party serve as an intermediary between the evaluation client/program and the evaluator. Typically, the goal screener examines all materials and communiqués as well as facilitates the dialogue and meetings between the program representatives and the evaluator to eliminate all goal and objective-based references. This goal screener can be an administrative assistant, an evaluator who is not part of the evaluation design or data collection, or even the evaluation client, for example.

There are two versions of the intentional GFE: full and partial. A full GFE is one in which the entire evaluation is goal free, from initial meetings with the evaluation client to the final reporting. The majority of the scholarly literature on GFE concerns the full GFE. There are a couple of factors that may provide the impetus for evaluation clients and program administrators to consider the full GFE. For example, programs that are willing to relinquish some control over the evaluation tend to be somewhat more confident in their programs' results and they have existing goal-based monitoring and evaluation endeavors. These programs use GFE as a supplemental evaluation tool and as an independent check on their program outcomes. The second version of the intentional GFE is the partial GFE. The partial GFE begins goal free and then, at some point during the course of the evaluation, the evaluation client reveals the goals and objectives to the evaluators and the evaluators proceed with the knowledge of the stated goals. Partial GFEs allow the evaluation client to have many of the benefits of the full GFE while still ensuring some assessment of goal-specific outcomes. The partial GFE is typically used due to some skepticism of GFE and/or because program funders or administrators require reporting on goal attainment.

The second category of GFE consists of evaluation models that are goal free by default. Usually, these evaluators do not take steps to blind themselves from the goals; instead, they simply seek program outcomes. For example, models like most significant change, participatory assessment of development, and qualitative impact protocol ask program consumers about any changes or outcomes that they have experienced, typically within a given time frame; the evaluators collect these data without referencing the intervention’s goals. Evaluators of international development programs have employed this type of GFE as program scope and coverage are often geographically vast, and/or the potential program impact can sometimes be subtle. So rather than asking a program beneficiary about a specific program goal—for example, whether there has been a decrease in livestock mortality due to a vaccination program—the evaluator asks about any change that they have experienced or witnessed and then the evaluator identifies whether these reported changes are attributable to the intervention.

Benefits of GFE

Both Scriven and Youker identify several benefits of employing a GFE. The primary benefits include reducing goal-oriented biases, avoiding goal rhetoric, adapting to environmental changes, aligning program goals with actual activities and outcomes, and supplementing GBE initiatives.

The first benefit is that GFE can reduce bias. By maintaining goal ignorance, the evaluator avoids tunnel vision toward the goals, which can often lead to groupthink. Instead, the evaluator is increasingly able to search for any and all relevant outcomes, thus identifying potential unintended positive and negative side effects caused by the intervention. Furthermore, GFE also minimizes conflicts of interest as the evaluator is unable to sycophantically please the evaluation clients by telling them what they want to hear because the evaluator does not know the program’s intentions.

Another benefit of GFE is that it allows both the program people and the evaluator to avoid the rhetoric of goals. Program designers and administrators create program goals often couched in fleeting fads or idealistic aspirations. These administrators occasionally and/or sanctimoniously consult program consumers and other impactees, if at all. This begs the question: Whose goals matter? By disregarding the goals and objectives, the goal-free evaluator eliminates the dilemma of determining the true goals as well as deciding whose goals count. Moreover, poorly written goals and objectives frequently lead to goals and objectives that are irrelevant or are either too easily achieved or are unattainable. Goal hyperbole is not a distraction for the goal-free evaluator.

Just as a program is dynamic, GFE is adaptable to environmental changes within the program. Most GBEs are static while GFE can accommodate changes in consumer needs, program resources, and program practices. The goal-free evaluator can proceed despite changes in the program as long as the program's activities and the outcomes reflect adapting to these changes and the changes are observable.

GFE can align the program's goals with its actual activities and performance. The evaluation stakeholders can examine the GFE report and assess the degree to which the program actions, as recorded by the goal-free evaluator, match its established goals. Furthermore, there are instances in which goal-free evaluators, based on their data collection, report what they believe to be the program's goals. In both cases, GFE serves as a way of calibrating the program's goals, providing information for adding, eliminating, or editing goals and objectives that may result in improving future program monitoring and evaluation endeavors.

Lastly, whether full or partial, GFE supplements GBE, thus reducing the methodological limitations inherent with each individual approach. Employing two evaluation approaches acts as a method of triangulating by evaluation approach as one approach assesses the results of the other; this is especially the case when a GBE and GFE simultaneously yet independently evaluate the same program. Furthermore, when GBE and GFE employ different data collection methods, they triangulate data collection methods and data sources as well.

Limitations of GFE

GFE has six noteworthy limitations. First, GFE is inappropriate under certain circumstances. The second limitation is that GFE is not advocated as a standalone evaluation approach. Third, GFE may not be the most efficient way to evaluate program outcomes. Fourth, GFE disregards the opinions and goals of selected stakeholders. Fifth, some data collection methods are less apropos for use with GFE. Sixth, there is limited information for instructing the evaluator how to actually conduct a GFE.

There are three situations in which GFE is ill-advised, as each of these situations jeopardizes the goal-free nature of the evaluation. First, in a full GFE, when the evaluator and evaluation client fail to identify an independent goal screener, GFE is inappropriate. Second, GFE is imprudent when an evaluation client or program stakeholders are unwilling to adhere to the goal-free nature of the evaluation. Lastly, GFE is improper when the evaluator has extensive prior knowledge of or experience with the program, especially if the goal-free evaluator already knows the program's goals and objectives.

The recommendation is to use GFE as a supplement to GBE, not a replacement for it. GFE is not appropriate as the sole evaluation approach for evaluating a program; rather the role for GFE is as a component of a larger evaluation strategy. For evaluation clients and program administrators who request or mandate a report exclusively on goal achievement, GFE is not a suitable approach. This is often the case for evaluations of federally funded initiatives. For instance, programs funded by the National Science Foundation require that the evaluators specify the program's intentions as well as what should be examined during the evaluation. Furthermore, many requests for evaluation proposals dictate that the evaluator discusses logic models and program theories, which also compel the evaluator to refer to the program's goals and objectives.

The goal-free evaluator considers a wider or broader context of goals than does the goal-based evaluator, and therefore, a GFE may potentially deplete valuable evaluation resources. The goal-free evaluator dedicates substantial evaluation resources to searching for outcomes, some of which may not prove relevant, when the evaluator could be developing and refining evaluation instruments that target intended, program-specific outcomes. In other words, without the cuing of goals and by reducing interaction with program personnel, the goal-free evaluator may be at more risk than the goal-based evaluator for making incorrect assumptions about pertinent outcomes, sources for outcome data, and methods of data collection.

A common criticism of GFE, and hence a limitation, is based on the fact that the goal-free evaluator emphasizes the values and needs of the program consumer, and in doing so, the evaluators marginalize particular evaluation stakeholders, namely the evaluation clients as well as the program funders, administrators, managers, and staff. The argument is that sometimes a wide group of stakeholders carefully constructed and agreed upon the goals and objectives yet the goal-free evaluator ignores them and their goal-setting processes. With some evaluations, elected politicians are a stakeholder group whose goals the goal-fee evaluator also dismisses; this is noteworthy because elected officials should be representing the interests and goals of their constituents.

The fifth limitation is that certain research designs and data collection methods are less suitable with GFE. For example, experimental designs, random control are less suitable with GFE. For example, experimental designs, random control trials, and other theory and hypothesis testing designs are ill-suited for use with GFE as there are very few circumstances in which they would permit the maintenance of the goal-free nature of the evaluation. Thus, GFE data analysis has yet to involve inferential statistics. In addition, pretests–posttests and other quantitatively based questionnaires are rarely applicable with GFE.

The final limitation in using GFE is the limited information to guide the evaluator in conducting a GFE. To date, there is no guidebook or manual that describes GFE's field implementation or protocol. The prescriptions that do exist concern the aforementioned four principles of GFE, the goal-screening process, and a dos and don'ts checklist.

Brandon W. Youker

See also Accountability; Conflict of Interest; Consumer-Oriented Evaluation Approach; Goals and Objectives; Objectivity; Outcomes; Program Evaluation; Triangulation; Values

Further Readings

Coperstake, J. (2014). Credible impact evaluation in complex contexts: Confirmatory and exploratory approaches. Evaluation, 20(4), 412–427.

Davies, R., & Dart, J. (2003). A dialogical, story-based evaluation tool: The most significant change technique. American Journal of Evaluation, 24(2), 127–155.

Dietz, T., van der Geest, K., & Obeng, F. (2013). Local perceptions of development and change in Northern Ghana. In J. Yaro (Ed.), Rural development in Northern Ghana (pp. 17–3). New York, NY: Nova Science.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), School evaluation: The politics and process (pp. 319–328). Berkeley, CA: McCutchan.

Scriven, M. (1974). Prose and cons about goal-free evaluation. In W. J. Popham

(Ed.), Evaluation in education: Current applications (pp. 34–67). Berkeley, CA: McCutchan.

Scriven, M. (1991). Evaluation thesaurus (4th ed.). Newbury Park, CA: Sage.

Youker, B. W., & Ingraham, A. (2013). Goal-free evaluation: An orientation for foundations' evaluations. The Foundation Review, 5(4), 53–63.

Youker, B. W., Ingraham, A., & Bayer, N. (2014). An assessment of goal-free evaluation: Case studies of four goal-free evaluations. Evaluation and Program Planning, 46, 10–16.

Miles Allen McNall Miles Allen McNall McNall, Miles Allen

Goals and Objectives

Goals and objectives

739

741

Goals and Objectives

The terms *goals* and *objectives* are closely related, yet distinct. Although goals refer to the general aims or intended results of a program, objectives are the specific, measureable steps one takes to achieve a goal. Goals tend to be broad, intangible, and abstract; objectives are more precise, tangible, and concrete. This entry further defines goals and objectives and discusses how they relate to program development and evaluation.

Because accomplishing goals requires completing several intermediate steps, each goal is typically associated with several objectives. Well-specified goals and objectives are specific, measureable, achievable, relevant, and time bound (SMART). SMART goals and objectives are specific with regard to the desired outcome and population of interest, stipulate how the achievement of the objective will be measured, are achievable with the resources and capabilities of those attempting to accomplish them, are relevant to the larger vision and mission of the group, and specify the time frame within which they will be accomplished. For example, while the general goal of an educational program might be to improve the reading skills of third graders, a “SMART” version of the same goal might read: By 2020, 75% of third grade children will score proficient in reading as measured by the statewide standardized third-grade reading test.

Although the SMART version of the goal is clearer with regard to the intended outcome, method of measurement, and timing, it does not specify the concrete steps that would need to be taken to accomplish it, that is, the realm of objectives. Objectives related to this goal might include reviewing school district data to determine whether reading proficiency scores were lower among certain groups of students than among others, selecting or developing reading programs designed to improve reading skills among lower performing students and implementing reading programs. Each of these objectives could be further specified in SMART format.

Goals and objectives are essential elements of program theory and planning. Insofar as a program consists of an ongoing set of activities designed to accomplish some specific end, it is difficult to conceive of a program without considering its goals and objectives. Goals and objectives are also core elements of a program's theory, which entails assumptions about the relationship between a program's resources, activities, outputs, outcomes, and impacts. These elements are often represented visually in the form of a program logic model.

Together, goals and objectives serve as essential building blocks for developing program plans, garnering and deploying resources, establishing accountability for results, and assessing outcomes. For research and evaluation purposes, program goals and objectives often serve as the starting points for identifying program outcomes and developing evaluation designs to measure program outcomes and impacts. The "SMARTer" the program goal or objective, the easier it is to make the translation from goal or objective to measureable outcome or impact.

Even when a program's goals and objectives are written and clearly specified, it is often the case that program stakeholders hold different and sometimes competing understandings of a program's goals and objectives. The frequency with which the evaluator Joseph S. Wholey encountered widespread disagreement among stakeholders on the goals of federal programs led him to develop an approach known as *evaluability assessment*, the primary purpose of which is to determine whether a program is ready for a full-scale evaluation. One essential precondition for the evaluability of a program is a reasonable level of agreement among a program's stakeholders on the program's goals.

As useful as goals and objectives are for program planning, research, and evaluation purposes, there is some controversy about their proper role in program evaluation. Based on the seminal work of Ralph Tyler, regarded by many as the father of educational evaluation and assessment, objectives-based evaluation focuses on the clear specification of objectives and the precise measurement of outcomes. Because of the influential work of Tyler and others (e.g., W. James Popham), the objectives-based evaluation approach has (e.g., W. James Popham), the objectives-based evaluation approach has dominated educational evaluation for several decades and still wields considerable influence. However, as this approach gained ascendancy, it also encountered criticism on the grounds that its narrow focus on the measurement of objectives means that it fails to make true evaluative assessments of the merit, worth, or significance of programs.

Evaluation theorist Michael Scriven argues that goal-based evaluation is not properly considered evaluation at all; rather, it is a form of program monitoring whose deficiencies include a failure to consider the relevance of a program’s goals to the needs of the impacted population, unintended effects (both positive and negative), program costs, or comparisons to relevant alternatives. The shortcomings of goals-or objectives-based evaluation led Scriven to propose goal-free evaluation, the aim of which is to determine the *actual* effects of a program instead of what it is *trying* to accomplish. If the intended effects of a program are accomplished, then they will be discovered by a well-designed and well-implemented evaluation. If they are not, they are irrelevant.

The true merit of any program, Scriven argues, is the extent to which it meets the needs of its intended beneficiaries, not whether it meets programmatic goals. Although Scriven’s position might be considered somewhat extreme, it does serve as a useful caution to evaluators who might otherwise conduct evaluations that are too closely aligned with the interests of program managers or sponsors and insufficiently attuned to the needs of intended beneficiaries or to the unintended effects of the program in question. It also suggests that program goals and objectives should be closely linked to the carefully assessed needs of the population of interest, not to what others, including program managers, think they are. Consequently, any statement of programmatic goals and objectives should be grounded in a thorough assessment of the needs of the population of interest.

Miles Allen McNall

See also Collaborative Evaluation; Developmental Evaluation; Logic Models; Participatory Evaluation; Program Evaluation; Program Theory of Change; Utilization-Focused Evaluation

Further Readings

Popham, W. J. (1988). Educational evaluation (2nd ed.). Englewood Cliffs, NJ:

Prentice Hall.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). Evaluation: A systemic approach (7th ed.). Thousand Oaks, CA: Sage.

Scriven, M. (1991). Evaluation thesaurus (4th ed.). Thousand Oaks, CA: Sage.

Tyler, R. (1942). General statement on evaluation. Journal of Educational Research, 35, 492–501.

Wholey, J. S. (1987). Evaluability assessment: Developing agreement on goals, objectives and strategies for improving performance. In J. Wholey (Ed.), Organizational excellence: Stimulating quality and communicating value. Washington, DC: Heath.

Yang Lydia Yang Yang Lydia Yang Yang, Yang Lydia

Goodness-of-Fit Tests Goodness-of-fit tests

Goodness-of-Fit Tests

Goodness-of-fit tests include various tests that measure how well a statistical model (which is built from theory) fits the observed data. Depending on the types of distributions and the nature of the variables being examined, different goodness-of-fit tests are used. Commonly used tests include Pearson's chi-square (χ^2) test and R^2 measure of goodness of fit. This entry describes the rationale of each test and the steps taken to conduct each test.

Pearson's χ^2 Test

Pearson's χ^2 test is a goodness-of-fit test used in the context of discrete distributions (i.e., the data are categorical in nature). Introduced by Karl Pearson in 1900, Pearson's χ^2 test evaluates whether the frequency of observations in each category statistically differs from the theoretical prediction. A "good fit" model indicates we can reasonably suggest that the observed data have come from the theoretically predicted distribution. It uses the null hypothesis statistical significance testing procedure based on the χ^2 distribution.

Pearson's χ^2 test follows these steps:

1. State the null hypothesis: The frequency of observed values in the sample is statistically similar to the frequency predicted theoretically.
2. Select statistical significance level.
3. Calculate the χ^2 test statistic using the following formula:

$$\chi^2 = \sum\left[(O_i - E_i)^2 / E_i\right],$$

where O_i is the observed frequency count for the *i*th level of the categorical variable and E_i is the expected (theoretical) frequency count for the *i*th level of the categorical variable.

In the extreme case, if the observed data and the expected values are identical, χ^2 would be zero. This indicates that the theoretical prediction and observed data match perfectly. As the predicted values deviate farther from the observed data, the χ^2 statistic will increase until it reaches a critical point where it is deemed that the theoretical model and the observed data come from two different distributions. To find out this critical point, we move on to the next steps.

4. Determine the degrees of freedom, which is the number of categories minus 1.
5. Compare the calculated χ^2 statistic to the critical value from the χ^2 distribution.
6. Make a decision on the null hypothesis. If the calculated χ^2 statistic is greater than the critical value of χ^2, the null hypothesis is rejected. At this point, the difference between the frequency of the observed data and the predicted model is so large that they are most likely from different distributions. If the calculated χ^2 statistic is smaller than the critical value of χ^2, we fail to reject the null hypothesis. This means there is not sufficient evidence to suggest that the observed frequency distribution and the expected frequency distribution belong to different distributions. Therefore, it is reasonable to conclude the observed data have come from the predicted distribution.

To use Pearson’s χ^2 test, three assumptions must be met. First, expected frequency numbers in each cell (cell counts) should be adequate. The rule of thumb is at least five data points in each cell. Second, the overall sample size should be large enough. Although there isn’t any agreed upon rule of thumb for sample size in χ^2 tests, small samples can lead to Type II error. Lastly, all observations should be independent of each other.

R^2 Measure of Goodness of Fit

R^2 can be used as a measure of goodness of fit with continuous distributions (i.e., the data are interval or ratio). R^2 indicates to what extent the observed data fit a statistical model. In linear regression model, if the differences between the observed values and the values predicted from a regression line are small and unbiased, the data and the model fit well. R^2 is often interpreted as the fraction of the total variability in the observed outcome that is explained by the model. The R^2 value ranges from 0.0 to 1.0. An R^2 of 1 (i.e., 100%) means the model perfectly explains the total variance in the observed outcome. In this case, all observed values in the data set fall exactly on the regression line, thus the regression line and the data fit perfectly. An R^2 of 0 means the model does not explain any variance of the outcome, indicating the regression line does not fit the data at all. In general, a high R^2 value indicates a “good fit” model, meaning the model explains a high proportion of the variability in the observed data.

R^2 is calculated directly using the following formulas:

$$R^2 = SS_{\text{model}} / SS_{\text{total}},$$

$$SS_{\text{model}} = \sum\left[(y_i - y)^2\right],$$

$$SS_{\text{total}} = \sum\left[(y_i - y)^2\right],$$

where y is the mean of all the observations of the outcome variable, is the ith predicted value of the outcome variable, and y_i is the ith observation of the outcome variable.

The R^2 measure of goodness of fit has one limitation. As more predictor variables are added into a model, R^2 always increases due to chance alone. If the model has a large number of predictors, R^2 is more likely to be biased. That is, although R^2 may increase in absolute values, the model may not better explain the variance in the outcome. In this case, adjusted is often used to adjust for the number of predictor variables in a model. can be calculated by modification of R^2:

$$R^2_{\text{adjusted}} = 1 - \left[\left(1 - R^2\right)(N-1)/(N-p-1)\right],$$

where p is the number of predictor variables and N is the total sample size.

If adding an additional predictor does not improve prediction as much as chance alone, then decreases. In extreme cases, can be negative, but it is always less than the value of R^2 in practice.

Yang Lydia Yang

See also Alpha Level; Chi-Square Test; Hypothesis Testing; R^2; Type II Error

Further Readings

Cameron, A. C., & Windmeijer, F. A. G. (1997). An *R*-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics, 77, 329–342.

Greenwood, P. E., & Nikulin, M. S. (1996). A guide to chi-squared testing. New York, NY: Wiley.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. International Statistical Review, 51, 59–72.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). Understanding regression analysis: An introductory guide (Vol. 57). Thousand Oaks, CA: Sage.

Frederick Burrack Frederick Burrack Burrack, Frederick

Grade Equivalent Scores Grade equivalent scores

742

743

Grade Equivalent Scores

The grade equivalent score compares a child’s performance on a grade-level examination to the median (central most) score of other students on the same material in the test’s norming group. This score is expressed as a decimal number with the left digit representing the grade level of a student’s performance and the right digit representing the approximate month in a 10-month academic school year. For example, if a sixth-grade student earned a grade equivalent score of 6.5, this would indicate that the raw score of this sixth-grade student is at the level of the sixth-grade student in the 5th month of the academic year or February (September is month 0, October is month 1, and so on through June, which is month 9). If a sixth-grade student receives a grade equivalent score of 7.8, this means that on this particular exam, this student scored at the level at which a seventh-grade student near the end of the school year would score on the sixth-grade exam.

Common Misconceptions

Grade equivalent scores represent a student’s ability in comparison to students who were in the specific test’s norming group. It is important to understand that grade equivalent scores above or below a student’s grade are common and should only be interpreted as if the student on this particular examination scored above or below the median score on the test. A score of 8.4 by a sixth-grade student does not indicate that student is capable of doing eighth-grade level work.

Grade equivalent scores are often misinterpreted as being a grade-level standard. A score of 6.5 on a sixth-grade examination may not represent the desired level of achievement for all sixth-grade students. It simply represents the median (central most) score of sixth-grade students during the 5th month of the academic year on this particular examination. Achieving this score may not be appropriate for either an individual or an entire group of students.

Appropriate Use of Grade Equivalent Scores

Grade equivalent scores should not be used in mathematical calculations such as comparing the average (mean) because the scale is not an equal interval scale. It is also not useful in comparing progress. For example, going from 3.4 to 3.9 is not a similar amount of growth as going from 8.4 to 8.9. In actuality, going from 2.5 to 2.9 is much greater growth than going from 8.4 to 8.9. But grade equivalent scores can be calculated for a group of students. If the average (mean) score is drawn from the group’s raw scores, then this mean score can be converted to a grade equivalent score for the group as a whole.

Grade equivalent scores cannot be used as a justification for grade-level placement or to determine the appropriate level of material for students in a course of study. These scores are simply useful to identify whether a child is functioning at, above, or below the central level of achievement on a particular examination.

Frederick Burrack

See also Age Equivalent Scores; Lexiles; Median Test; Norm-Referenced Interpretation; Ordinal-Level Measurement

Further Readings

American Educational Research Association/American Psychological Association/National Council on Measurement in Education Joint Committee. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Airasian, P. W. (1994). Classroom assessment (2nd ed.). New York, NY: McGraw-Hill.

Miller, M., Linn, R., & Gronlund, N. (2009). Measurement and assessment in teaching (10th ed.). Hoboken, NJ: Pearson Education.

Stiggins, R. J. (2009). Student-centered classroom assessment (2nd ed.). Upper Saddle River, NJ: Merrill, an imprint of Prentice Hall.

Lotta C. Larson Lotta C. Larson Larson, Lotta C.

Grading

Grading

743

746

Grading

The term *grading* generally refers to the process of rating student progress or performance in areas of academic achievement, activities, or behaviors using a coding system or established scales of values (such as letters, symbols, numbers, or percentages). This entry explores the history of grading, common methods used for reporting grades, the purposes of grading, and alternative grading practices.

On a daily basis, teachers use both formative and summative evaluations to gather information about students' progress in achieving learning goals. Formative assessment procedures are generally less formal and are used to guide student learning while it is still in progress. Often more formal, summative assessments help evaluate student achievement at the end of an instructional unit. Summative assessment results are commonly used to inform the grading process at the end of a marking period to create a report card, which may be shared with students, parents, and school officials. Teachers are generally expected to generate report cards several times during the school year. While report cards differ, they commonly address students' academic achievement; students' participation in extracurricular activities; and/or students' behaviors, personal skills, or dispositions (e.g., ability to collaborate, ability to solve problems, ability to be a self-directed learner).

History of Grading

Formal testing systems and related grading practices are relatively recent educational phenomena. In fact, grading and reporting were virtually unknown in schools in the United States until the mid-19th century. With students of mixed ages and backgrounds grouped together with one teacher in one-room schoolhouses, most did not pursue an education beyond elementary studies. Teachers shared students’ learning progress with parents, usually during visits to students’ homes. With the implementation of compulsory attendance and increased student enrollment in the early 20th century, formal evaluation methods were established to determine whether students could progress to the next level. At the elementary level, narrative report cards became common practice, while high school teachers used percentages and other markings to document students’ achievements. Although there have been many variations in grading practices over subsequent years, this was the beginning of the grading and reporting system that exists today.

Methods for Reporting Grades

It is important that any classroom practices associated with grading are meaningful and communicated to students, parents, or other stakeholders with accuracy and precision. Teachers must ensure that grading and reporting meet established criteria for validity and reliability. The methods for reporting grades vary at different grade levels. For example, letter grades are commonly used in the upper elementary through high school levels, whereas checklists and parent–teacher conferences are more frequent at the elementary level. No single reporting method adequately serves all purposes, and schools often combine multiple methods. For example, during parent–teacher conferences, teachers may share letter grades. Similarly, a report card may include a combination of letter grades, checklists, and narrative comments. What follows are common methods of reporting student achievement.

Checklist

A checklist typically contains a list of specific skills or behaviors that teachers mark as student progress, or gaining mastery, throughout the school year. Some checklists use verbal descriptors such as excellent, good, and needs improvement, whereas others use simple pass/fail indicators. Checklists containing behaviors often include descriptors such as always collaborates with others, often collaborates with others, or seldom collaborates with others.

A common criticism of checklists is that they do not provide detailed information about an individual's performance. However, they are generally quick and easy for teachers to complete and simple for parents or other stakeholders to interpret.

Letter Grades

When using letter grades, teachers generalize a great deal of information about a student's achievement into a single letter. Traditionally, letter grades used in the United States are A, B, C, D, or F; A being the highest and F (symbolizing failure) the lowest.

Letter grades are often based on a percentage system in which test items are scored either right or wrong. In this system, getting 90%–100% of answers correct on a test generally converts to the grade of A; 80%–89% of answers correct converts to the grade of B. Commonly, if below 60% of questions are answered correctly, a grade of F (failure) is assigned.

Critics of letter grades argue that the meaning of, or academic performance associated with, each letter grade is arbitrary, and cut offs between grades are difficult to justify. Letter grades often assess more than just academic performance and include factors such as class participation, late assignments, and student attendance. Consequently, letter grades do not always reflect students' academic abilities. Furthermore, letter grades may lack the richness of other more detailed reporting methods such as narrative accounts.

Narrative Accounts

Narrative reports are written descriptions of a student's achievement and educational development in relation to instructional goals. Narrative accounts often include unique information about individual students' progress and skills. Critics of narrative accounts point out that they are labor-intensive for the teacher and they can be biased or insensitively written. Parents may interpret narrative comments differently from what a teacher intended to communicate. However, when done well, narrative accounts can provide a rich description of a student's progress.

Parent–Teacher Conferences

A parent–teacher conference is a meeting or conference between parents or guardians and a teacher to discuss a student’s academic achievements, learning strengths and needs, and future social or academic goals. Parent–teacher conferences can build strong relationships among parents and teachers. However, preparing for and conducting conferences can be time-consuming for teachers, and not all parents have the time or desire to attend parent–teacher conferences.

Purposes of Grading

It is important for schools and districts to identify clear purposes for grading, as these tend to inform the particular functions and reporting methods of grading. Although there are many different purposes for grading, primary reasons include providing feedback, serving administrative purposes, and providing motivation.

Feedback

Communicating how well a student has achieved instructional goals is a fundamental purpose of grading. Grades provide important feedback to students as they help students keep track of their own progress and help identify individual strengths and areas for improvement. For students, grades may even impact important decisions regarding college and future careers. A student who regularly receives strong grades in a content field may be inclined to pursue further work or study in that area.

Communicating a student’s academic progress to parents is often done via report cards, during parent–teacher conferences, or by sharing results from standardized tests. When communicated formally, grades ensure that parents receive an overview of a student’s general performance, while indicating particular strengths or weaknesses. In addition to grades in subject areas, report cards may include feedback on a student’s behaviors, dispositions, or participation in activities.

Feedback from grades can help teachers make instructional decisions. Therefore, teachers who assign grades should pay close attention to changes in the distribution of grades—especially when the assessments bringing about the grades are consistent over time. Changes in grades may reflect changes in instructional approaches or the effectiveness of new curricula or instructional materials. Teachers can also use feedback from students to materials. Teachers may also use feedback from grades to group students according to interests or abilities, differentiate instruction, or provide remedial or enriching services.

Administrative Purposes

Grades serve a variety of administrative functions, often involving district-level decisions concerning student enrollment and placement of transfer students. Schools also use grades to determine a student’s class ranking, graduation requirements, and a student’s suitability for promotion or retention. In many middle schools or high schools, grades impact students’ eligibility for participation in athletics and other extracurricular activities. Students often make honor roll or earn academic awards based on grades. Furthermore, decisions regarding college admission and scholarships are often based primarily on grades.

Motivation

Grades also serve the purpose of motivating students. Supporters of using grades as motivation assume that receiving a low grade will encourage students to try harder, whereas receiving a high grade will motivate students to continue or renew their efforts. Teachers objecting to use grades for the purpose of motivation argue that student achievement should be intrinsically motivated and driven by the joy of learning. However, many students strive to get good grades and stay eligible for athletics and other extracurricular activities. Hence, grades may serve as strong motivators for students.

Alternative Grading Practices

Traditional report cards and standardized test scores provide students and parents with straightforward methods of monitoring academic progress. Such common grading practices are deeply rooted in today’s education systems. Although these practices serve multiple purposes, some educators, researchers, and parents question the effectiveness of grades. Alfie Kohn, an author and lecturer on education, is an outspoken critic of grades and test scores and argues that modern grading practices, particularly letter grades, can have negative effects on students’ creative thinking, engagement, and motivation to take on challenges.

In response to criticism against conventional grading practices, some school districts are replacing traditional report cards with standards-based grades that measure students' proficiency on clearly defined course objectives. Supporters of standards-based grades argue that while letter grades focus on points and percentages, standards-based practices evaluate how well a student meets measurable benchmarks and objectives. On standards-based report cards, each subject area is divided into a list of skills or standards. Students receive a separate mark for each standard.

Supporters of alternative grading practices point to research indicating that self-motivated learners do not need grades as motivation to learn and that students often base their self-worth on academic performance, which may result in stress, anger, and academic deficiencies. There is some support for the idea that teacher feedback and constructive criticism from peers and teachers encourage greater growth in academic performance than letter grades. Furthermore, engaging students in self-assessment practices can promote responsibility and self-motivation, resulting in student-driven accountability for learning.

Lotta C. Larson

See also Accountability; Evaluation; Motivation; Reliability; Standardized Tests; Standards-Based Assessment; Validity

Further Readings

Guskey, T. R. (2015). On your mark: Challenging the conventions of grading and reporting. Bloomington, IN: Solution Tree Press.

Kohn, A. (1994). The issue is not how but why. Educational Leadership, 52(2), 38–41.

Munoz, M. A., & Guskey, T. R. (2015). Standards-based grading and reporting will improve education. Phi Delta Kappan, 96(7), 64–68.

Reeves, D. (2011). Elements of grading: A guide to effective practice. Bloomington, IN: Solution Tree Press.

Townsley, M. (2013/2014). Redesigning grading—Districtwide. Educational Leadership, 71(4), 68–71.

Irene Kaimi Irene Kaimi Kaimi, Irene

Christoforos Mamas Christoforos Mamas Mamas, Christoforos

Graphical Modeling

Graphical modeling

746

750

Graphical Modeling

Graphical modeling uses graphs, which present the different ways the variables in a model depend on each other, to represent and visualize the model. The model's variables can be simply associated or be connected through causal relationships. The resulting displays rely on probability and graph theory, graph algorithms and machine learning; as such, they connect concepts from statistics and computer science.

A wide range of different types of graphical models and methods have been developed in a variety of areas including, but not limited to, medical diagnosis, image understanding, speech recognition, and natural language processing. The use of graphical models can also enable understanding of social and technical features of organizations and structures. In education, such systems may extend from the classroom unit to the school and from the educational system of a country to the educational systems of several countries. Visualization and interpretation of the underlying structures between members of these systems can help in identifying isolated members, which potentially share common characteristics. This in turn can lead to the introduction of improved policies and practices, so that the educational and social needs of all (or groups of the) corresponding members (e.g., students, schools, and educational systems) are better met. This entry presents some of the basic ideas of graphical modeling and then illustrates the concepts in the context of social network analysis.

Some Probability Concepts

Probabilities are used in everyday life and determine our decision-making processes. For example, the chance of rain informs one's plans for the weekend. Each time we model the real-world uncertainty, there will be some underlying *random experiment* (such as flipping a coin to decide whether to cycle or drive to work). If the experiment (e.g., the toss of a coin) is repeated a very large (in theory infinite) number of times, the event happens roughly a fraction p of the time (e.g., half of the time we will get heads); the larger the number of repetitions, the closer we will get to the true probability of the event.

For a random experiment (e.g., the toss of a fair coin), the *sample space* is the set of all possible outcomes (e.g., heads and tails), an *event* (e.g., heads) is a subset of the sample space, and the *probability of the event* will be a number between 0 and 1 (in the fair coin example, p (heads) = p (tails) = ½).

A *random variable*, usually denoted by a capital letter, say X, is a variable of which the possible values are the numerical outcomes of a random experiment (in the fair coin example, if X is the number of heads in one toss of the coin, X can take the values 0 and 1).

Statistical inference, in its simplest form, uses an observation to draw conclusions about some unknown quantity. Both the unknown quantity and the observation are represented here by random variables and the modeling objective is to decide whether and in what way the two random variables relate. For the events of getting heads in the toss of a coin, and rain the following day, X represents the number of heads in the toss of a coin ($X = 0$ or $X = 1$), and Y is an indicator random variable that it will rain tomorrow ($Y = 0$ or $Y = 1$). Then, the observation that the random variable X takes on a specific value (e.g., $X = 0$) is not expected to affect the random variable Y. In this case, X and Y are *independent*; conditional on the observed value of X, the probabilities of the values of Y remain unchanged.

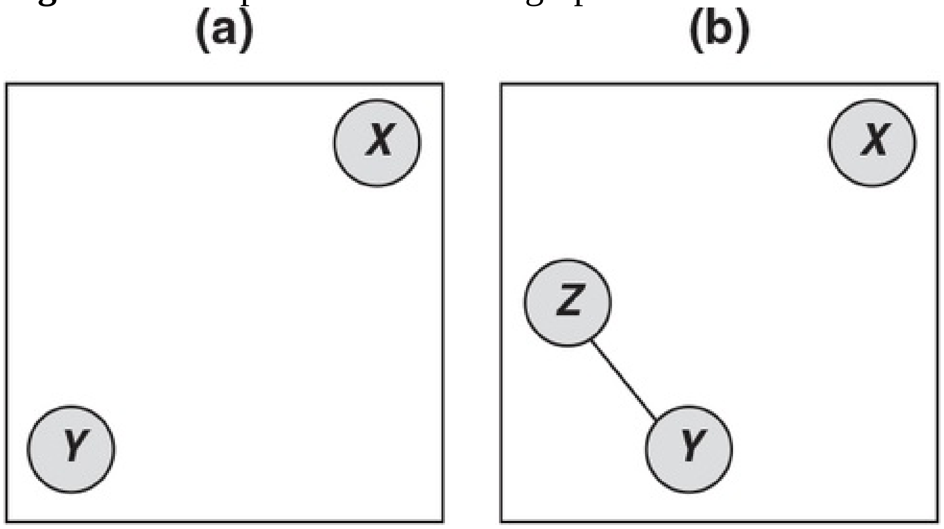
If Z is considered to be an indicator random variable that one will cycle rather than drive to work tomorrow ($Z = 0$ or $Z = 1$), then observation for Y will actually affect the perception of which are the likely or unlikely values for Z. Y and Z are *dependent* random variables. Conditional on the observation for Y, the probability for the outcome for Z changes.

Basic Examples of Graphical Models

A graph allows visualization of the relationships between variables based on the details of their forms. Graphical models can be used to define straightforward algorithms that implement probabilistic inference, that is, algorithms able to derive the probability of one or more random variables taking a specific value or set of values; this implementation is efficient and does not require enumeration of all settings of all variables in the model.

Let *X* and *Y* denote two independent random variables. Then, the joint probability of the two random variables is the product of the two individual probabilities, whereas conditioning on *X* does not affect the distribution of *Y* and vice versa. Graphically, the joint distribution of *X* and *Y* is represented as two circles. Their independence suggests that one should not connect these two circles, as shown in Figure 1a.

Figure 1 Examples of undirected graphs



In a new scenario, *X* and *Y* are independent, but a third random variable, *Z*, possibly depends on *Y*. Then, statistical theory methods require calculation of joint, marginal, and conditional probabilities that will determine which pairs of the three random variables *X*, *Y*, and *Z* are independent. Using these probabilities and taking into account that there is a joint probability table for *Y* and *Z*, but such a table is not required for the pairs *X* and *Y* as well as *X* and *Z*; this new graphical representation is shown in Figure 1b.

Undirected Graphs

The examples in the previous section suggest that each model involves a graph and tables of probabilities that decide the form of this graph. For a graph, the circles are formally called *nodes* or *vertices*, and the lines are its *edges*. The graph is *undirected* when the edges do not have directionality associated with them. In addition, each node corresponds to a random variable and each edge suggests whether there is a possible association between the two connected nodes. The larger the number of edges, the larger the number of probability distributions the graph implies that the model can interpret.

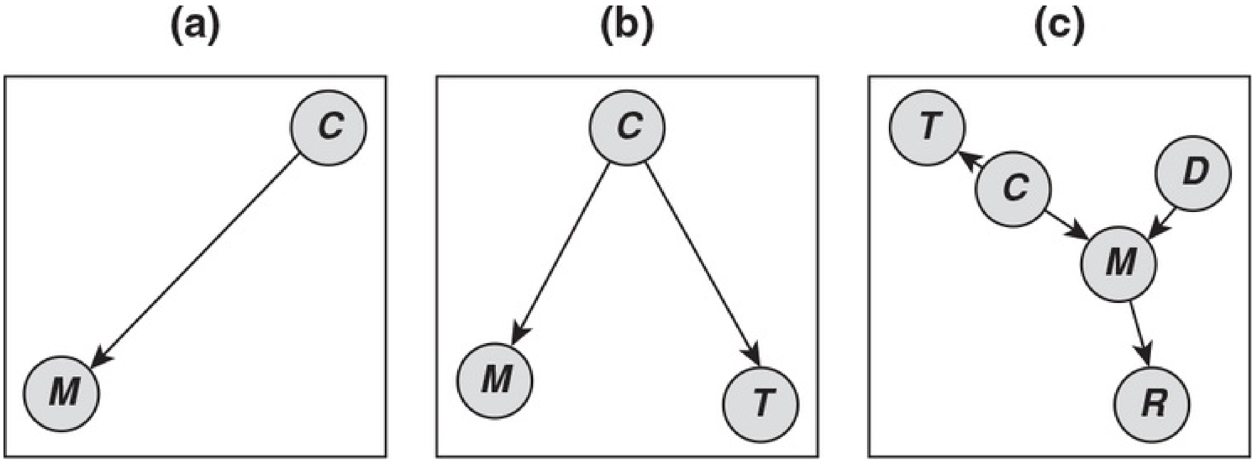
In order to specify a Graph G, we need to specify the set of nodes N and the set of edges E. Each edge consists of a pair of vertices s, t from the set of edges E. For undirected graphs, there is no distinction between edge (s, t) and edge (t, s). In Figure 1a, where *X* are *Y* are independent, the graph only involves the two nodes, but no edges. In Figure 1b, *Y* and *Z* (but not *X* and *Y* nor *X* and *Z*) are possibly dependent, hence the graph contains both the set of nodes V = {1, 2, 3} and the set of edge(s) E = {(2, 3)}. Undirected graphical models are also called *Markov random fields* or *Markov networks*.

Directed Graphs

The use of undirected graphical models may understate some independencies. Often, two variables are connected because some other variable depends on them. *Directed graphs* or *Bayesian networks* are not limited to representing distributions satisfying the strong independence assumptions inherent in Markov networks. The realistic independence properties of a given setting are taken into account when portraying the distribution.

The next example is about a finance company, wishing to hire a recent university graduate. Although the company aims for clever new employees, their cleverness cannot be tested directly. Instead, the company uses the graduates' average mathematics degree mark to decide. For illustration purposes, the assumption that the variables *C* (cleverness) and *M* (average degree mark) are binary (clever/not clever; high/low average mark) is made. In this instance, the resulting network has one node for each of the two random variables *C* and *M*, with an edge from *C* to *M* representing the direction of the dependence (or *influence*) in this model, as shown in Figure 2a.

Figure 2 Examples of Bayesian networks



It should be noted that a graduate can obtain a high average degree mark by working hard even if the graduate is not clever. A third binary variable *T* (high/low score) refers to the score of a candidate in a general mathematics skills knowledge test the candidates have to undertake before the final hiring decision is made by the company. For any realistic representation of this information, the graduate's cleverness is correlated both with his average degree mark and his score in the test.

The degree mark and the test score are also not independent. Yet, a conditional independence property may hold; given that a graduate is clever, a high average degree mark does not provide additional information about the graduate's test performance. This is only true under the rather strong assumption that random variables *M* and *T* are only correlated because of the graduate's cleverness and not, for example, for his ability to perform well in written exams; such approximations to reality are often made in graphical models. Figure 2b gives a graphical representation of this model; *M* and *T* are only *conditionally independent* (there is no direct edge between them, but both are correlated with C, which is the influence for both).

In a more lifelike situation, the difficulty of the degree, *D*, is also assumed to influence the average degree mark of a graduate. In addition, in this situation the graduate needs a reference letter when applying for this position, so the graduate asks one of her final year professors to write a reference letter for her. The professor will provide a reference letter (random variable *R*) and this can be either good or bad (binary), depending on the graduate's degree mark (that the nature of the reference letter would depend only on the graduate's degree mark is a strong assumption that may not hold in practice, as the lecturer may for instance also remember the in-class participation of the graduate). Hence, the model in Figure 2c gives a more accurate representation of the truth. The degree difficulty and the student's cleverness are not related and are thus independent, whereas the graduate's average mark depends on both of these variables. On the other hand, the graduate's score on the company's test only depends on her cleverness, and the quality of the reference letter only depends on the graduate's average degree mark.

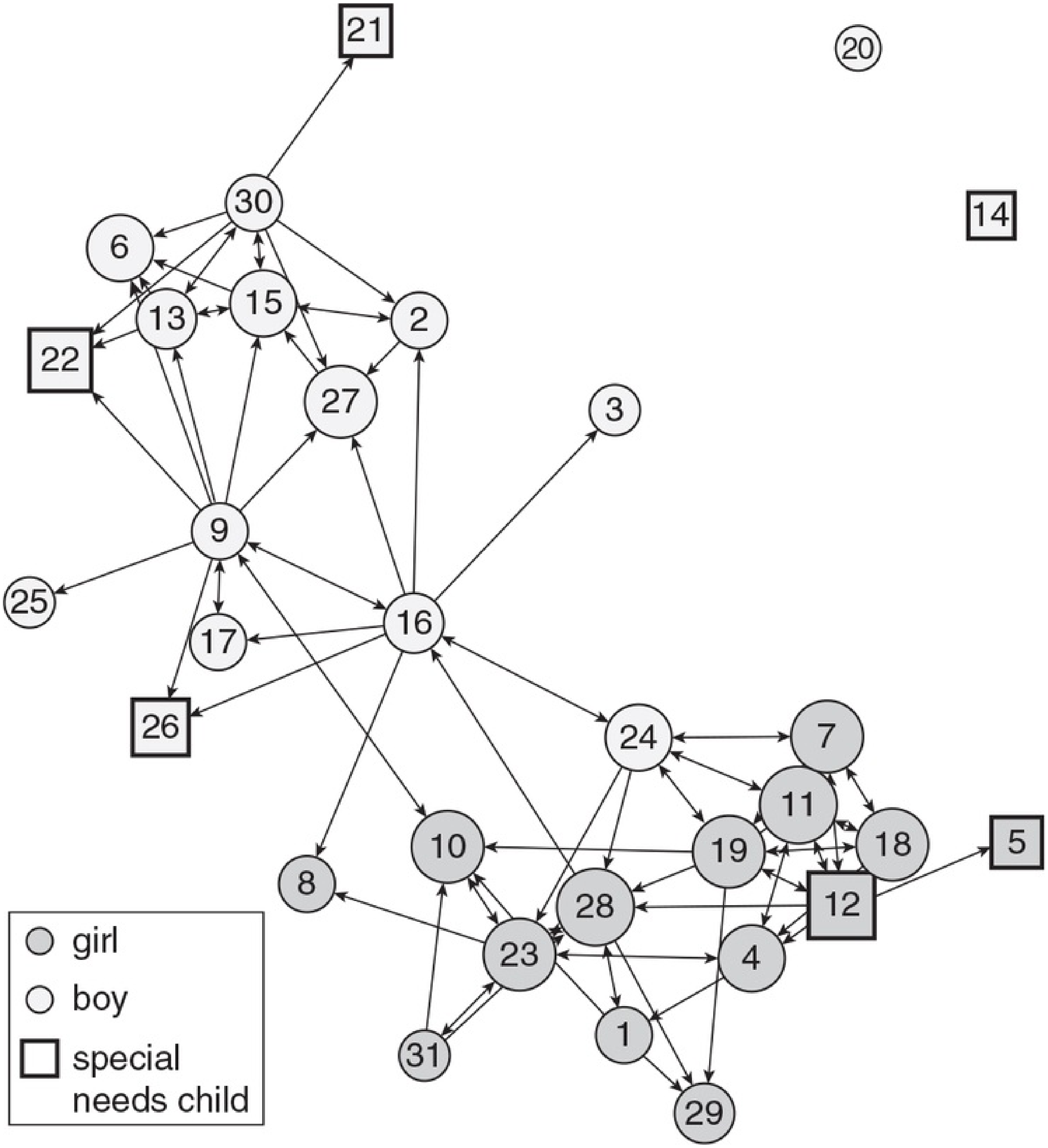
In general, a Bayesian network is represented using a directed graph, of which the nodes are the random variables of interest and the edges essentially relate to direct influence of one node on another. A directed graph G = (V, E) is formed by a collection of vertices V = {1, 2, …, m} and a collection of edges E. Each edge of a directed graph consists of a pair of vertices s, t from the set of edges E and (s→t) indicates the direction (in this case, from s to t).

More Complicated Graphical Models

All examples presented thus far involve a small number of random variables. In the typical real-world data situations, interest lies in understanding how several—often hundreds—of random variables are associated. Graphical models make use of the conditional (in)dependencies in a network of random variables to provide a condensed picture of a high-dimensional joint probability distribution of random variables.

One such instance is when the objective is to describe the friendships of elementary school children. For simplicity, a classroom of 31 children is considered. Each child in the classroom is treated as a random variable, who will potentially name any of his or her classmates as his or her friend. This results in 31 random variables which may or may not depend to each other. Figure 3 shows a graphical model that corresponds to this setup, in which it is easy to identify isolated students (students 14 and 20), as well as groups of students who tend to cluster together (e.g. girls—pink and boys—blue).

Figure 3 Example of classroom social network graph



The graphical model in Figure 3 is a practical way to present the information contained in 31 questionnaires and probability tables of size 31 by 31. This is an example of *social network analysis*, which uses graph theory to examine a social structure. Social network analysis brings together social sciences/humanities and computer/mathematical sciences in developing visualizations of social networks. In the education context, this approach is underused, but promising for interpreting networks and their dynamics, in order for the relationships between network members (e.g., students) to be better understood by people less familiar with the technical details (e.g., teachers), so that actions can be taken where needed.

Yet more complicated graphical models can be constructed for understanding the relationships between, for example, friends on Facebook, followers on Twitter, or trade and monetary networks.

Irene Kaimi and Christoforos Mamas

See also Bayesian Statistics; Conditional Independence; Matrices (in Social Network Analysis); Social Network Analysis; Social Network Analysis Using R

Further Readings

Blitzstein, J., & Hwang, J. (2014). Introduction to probability. Boca Raton, FL: CRC Press.

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). Graphical models with R. New York, NY: Springer-Verlag.

Jordan, M. I. (Ed.) (1999). Learning in graphical models. Cambridge, MA: MIT Press.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: Principles and techniques. Cambridge, MA: MIT Press.

Scott, J. (2012). Social network analysis (3rd ed.). Thousand Oaks, CA: Sage.

Wasserman, S., & Faust, K. (1994). Social networks analysis: Methods and applications. Cambridge, UK: Cambridge University Press.

Jennifer C. Greene Jennifer C. Greene Greene, Jennifer C.

Great Society Programs Great society programs

750

753

Great Society Programs

This entry describes the Great Society programs of the 1960s and looks at their effects on education and on other aspects of society and their implications for educational measurement and evaluation. The Great Society programs constitute an unprecedented set of federally funded educational, social, and environmental programs that were enacted during the presidency of Lyndon B. Johnson (1964–1968). These programs were intended to eliminate poverty and racial injustice and to ensure that the country’s systems of political, educational, employment, and health opportunities were open to all Americans. Many of the Great Society programs have endured until the present time, although they have gone through numerous revisions over the decades.

The policies and initiatives of Great Society programs have shaped education in the United States for over half a century. Among the most ambitious of the Great Society programs were those focused on improving educational access and quality for all children and youth, particularly for those from high-poverty communities. The signature piece of Great Society legislation in the education domain was the Elementary and Secondary Education Act (ESEA) of 1965.

The centerpiece of the original ESEA was its Title I, which provided extra funding to schools nationwide that served high concentrations of children and youth from low-income families and neighborhoods. In the same year that ESEA was passed, Head Start was launched and the Higher Education Act was passed, providing scholarships and low-interest loans to any young person who wanted to pursue higher education. The 2015 incarnation of the ESEA is the Every Student Succeeds Act, which replaced the No Child Left Behind Act of 2001.

The Great Society programs also reached into many other aspects of citizens’ lives. The multiple strands of this grand policy-making endeavor include the lives. The multiple strands of this grand policy-making endeavor include the following, each presented with an illustration of relevant legislation that was passed during Johnson’s presidency.

Civil rights

In 1964, the Civil Rights Act outlawed discrimination based on race, color, national origin, religion, or sex; and in 1965, the Voting Rights Act banned literacy tests as a requirement for voting.

Poverty

In 1965, the Economic Opportunity Act established educational, employment, and training programs targeted for the poor, as cornerstones of the War on Poverty.

Health

In 1965, legislation was passed establishing Medicare and Medicaid, health insurance programs for, respectively, the elderly and for low-income individuals and families.

Arts and media

In 1967, the Public Broadcasting Act provided financial assistance for noncommercial television and radio broadcasting, including the Public Broadcasting Service and National Public Radio.

Environment

During Johnson’s presidency, Congress passed an Air Quality Act, a Water Quality Act, and three acts aimed at preserving the nation’s wilderness, rivers, and scenic and recreational trails.

Housing and urban development

In 1965, the Omnibus Housing Bill provided substantial grants for low-income people to move into new housing projects and for both low-income homeowners and small businesses in blighted communities to rehabilitate their properties. The Cabinet-level Department of Housing and Urban Development was created that

Cabinet-level Department of Housing and Urban Development was created that same year.

Immigration

In 1965, the Immigration Act abolished immigration privileges previously afforded to immigrants from Europe.

Implications for Educational Measurement and Evaluation

The massive undertaking of the Great Society programs called for a significant response from education researchers, especially measurement experts and evaluators. Further, as the Great Society programs were modified, redirected, or reinforced over the decades, educational inquiry experts were expected to keep pace. Developments in educational measurement and evaluation became an integral part of the political, governance, and accountability structures of the times.

The 50-plus years since Lyndon Johnson’s presidency have also been a time of transformational change in the technologies of educational inquiry—change that remains dynamic and ongoing. For example, an early yet highly influential technological contribution to testing was the development of the high-speed scanner in the mid-1950s. The scanner enabled affordable multiple-choice testing of all students. Beyond acknowledging the powerful contributions of ever-evolving technologies to educational inquiry, the remainder of this entry concentrates on the policy and practice dimensions of educational measurement and evaluation that were influenced by the massive political experiment of the Great Society.

Recasting Large-Scale Testing in Service to Educational Accountability

In tandem with the Great Society educational programs, and with advances in technology, large-scale standardized state testing programs proliferated in the 1960s and 1970s and continue to the present time. These testing programs variously assessed “minimum competencies,” state-established content standards, and state-established accountability standards. Because these state tests served political needs for educational accountability, they were vulnerable to criticism from all sides. In response and in anticipation of future critiques, the measurement community developed increasingly sophisticated techniques for test item construction, for sampling (of both students and items), and for statistical analyses of results.

Hierarchical linear modeling is one example of the sophisticated statistical developments in the field because the increased emphasis on testing and educational accountability began in the 1960s. Hierarchical linear modeling is designed to account for shared variance in multilevel data (e.g., for students in the same classroom) and can yield accurate estimates of performance results at the student, classroom, grade, school, and district levels. Still, an inherent, and likely inevitable, tension remains between the political accountability agenda for large-scale testing (often supplied by private companies) and the scholarly agenda for developing high-quality, defensible tests that yield trustworthy information on how well schools serve children and youth.

Catalyzing the Development of the Professional Practice of Educational Evaluation

The professional field of educational program evaluation was just emerging at the time that the Great Society programs were planned, developed, and implemented. There were a few earlier educational evaluation studies of note. Well known to historians of educational inquiry is an early evaluation study known as the Eight-Year Study, conducted in the 1930s in the United States.

The Eight-Year Study was led by assessment expert Ralph Tyler, and its primary purpose was to assess the success of a traditional subject-driven high school curriculum in preparing students for college, compared to a more problem-based, cooperative curriculum and learning environment. The latter was intended to better serve youth who did not opt to attend college. Thirty high schools were chosen for the experimental curriculum, and students in these schools were matched with students attending schools with conventional curricula. The results clearly indicated that students in the 30 experimental schools performed just as well academically as students in matched schools and were more involved in cultural and artistic activities.

The Eight-Year Study of the 1930s well demonstrated the value and contributions of educational program evaluation, especially for innovative ideas. However, there was little further demand for evaluation until the avalanche of Great Society programs arrived in the mid-1960s. Accompanying this avalanche was a growing demand from elected officials and federal and state agencies to find out whether these programs “worked,” that is, whether they reached their intended objectives.

Among those who responded to these demands were educational and social science research experts, many from universities, who had the methodological expertise requisite for this work. University professors from multiple disciplines endeavored to study, document, and assess the success of educational programs implemented under the ESEA and other major Great Society initiatives. Most engaged these challenges with the dominant objectivist, quantitative methodology of that era, namely, randomized experimental studies.

However, rigorous experimental designs were often a poor fit to studying the educational programs developed as part of the Great Society. Although some of these programs were anchored in well-established principles of teaching and learning, others were trying out relatively new and as-yet-untested ideas. In studying the latter type of programs, an exclusive focus on intended outcomes was not a sensible evaluation strategy. Further, these programs, and thus also their evaluations, were conducted in real-life schools, classrooms, and playgrounds, where it was not possible to exert full experimental controls. Notably, randomization of individual students to experimental and control groups was rarely possible, as students were intrinsically nested within classrooms and schools.

Over the next decade, the response of the fledging evaluation community to these challenges of evaluating Great Society educational programs was to rethink and reimagine how the practice of educational evaluation could better address these challenges. Experimental designs remained, then as still today, favored by policy makers and some evaluators, as they directly assess how well a program reaches its intended outcomes. These designs gained enhanced practical relevance as quasi-experimentalism, designed as a more practical fit to real-life contexts, gained credibility through widespread use.

At the same time, in the late 1960s and into the 1970s, other evaluative questions and relevant audiences arose, and other kinds of evaluation approaches and designs were crafted for these other purposes and people. For example, educators in schools were interested in the quality of the educational learning experiences being provided for their students, in addition to the outcomes attained. Program developers wondered about the quality and contextual fit of the educational programs they designed. Minority communities and parents wondered how well their children were faring in programs designed specifically to enhance educational opportunity and equity for all children. A 1989 review of early evaluations of the Head Start program by Sadie Grimmett and Aline M. Garrett well illustrates these developments.

Legacy of the Great Society Programs in Educational Evaluation Today

The visible and public importance of the Great Society programs, including its educational programs, served to catalyze the development of the contemporary era of evaluation. This era of rapid and multipronged development of the evaluation enterprise encompassed an increasing diversity of evaluation purposes and audiences, evaluation approaches and methodologies, and intended evaluation uses. It also stimulated an exponential growth in evaluation’s footprint and the now fully global evaluation community.

In contemporary times, policy-oriented evaluation remains focused on intended outcomes and on establishing strong evidence bases for programs successful in attaining these outcomes. A wide range of program evaluation approaches and methodologies is available to meet the needs of all those with an interest in the evaluation of programs, including program leaders and staff, community leaders and activists, and intended program participants. Evaluation approaches today range from experimentalism to democratic and critical evaluation, methodologies from surveys to participatory narratives, and the intended uses of evaluation range from decision making to learning to social critique.

Jennifer C. Greene

See also Evaluation, History of; Every Student Succeeds Act; Head Start; Hierarchical Linear Modeling; No Child Left Behind Act; Policy Evaluation; Program Evaluation

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. Journal of Curriculum Studies, 32(2), 159–181.

Grimmett, S., & Garrett, A. M. (1989). A review of evaluations of Project Head Start. Journal of Negro Education, 58(1), 30–38.

Ritchie, C. C. (1971, February). The eight-year study. Educational Leadership, 484–486

Tumulty, K. (2014, May 17). The Great Society at 50. The Washington Post. Retrieved from http://www.washingtonpost.com/sf/national/2014/05/17/the-great-society-at-50/

The Washington Post. (2014, May 17). Evaluating the success of the Great Society. Retrieved from http://www.washingtonpost.com/wp-srv/special/national/great-society-at-50/

Julius Sim Julius Sim Sim, Julius

Grounded Theory Grounded theory

753

756

Grounded Theory

Grounded theory is a well-established and highly influential approach to qualitative research. Developed in the 1960s by two sociologists, Barney Glaser and Anselm Strauss, in the United States, it is characterized by a theory-building approach, based on an iterative, inductive process of data analysis. This entry reviews the history and development of the theory, its key characteristics, the process of analyzing data while using grounded theory, and other considerations regarding the theory.

History and Development

Grounded theory was an approach first presented in detail in Glaser and Strauss’s 1967 work, *The Discovery of Grounded Theory*. Fundamental to this approach was a concern to develop theory—both formal and substantive—from data in a systematic, inductive process. This was in reaction to a more traditional and dominant approach whereby theory was developed a priori and then tested against data through an essentially deductive process. Grounded theory subsequently became a widely used method in qualitative research, particularly in the context of health and illness. However, it has been the subject of considerable debate and some degree of controversy.

The most fundamental development in the subsequent history of grounded theory took the form of a schism between Glaser and Strauss. The publication in 1990 of the first edition of *Basics of Qualitative Research*, authored by Strauss and Juliet Corbin, provoked a strong reaction from Glaser, who felt that in this book, the emphasis on allowing theory to emerge directly from the data, free from the influence of theoretical preconceptions, had been lost and that the process of analysis had become excessively structured and prescriptive. In a series of texts, Glaser sought to reassert what he regarded as the authentic nature of grounded theory. Although Strauss did not respond openly to Glaser's concerns, this episode led to two distinct models of grounded theory, one Glaserian and one Straussian, and each with different emphases in terms of method and terminology.

Another important development is the use of grounded theory methods in research that is broadly inductive in its approach but does not necessarily adopt the principle of theory building central to the broader grounded theory approach. Grounded theory can therefore refer either to an overall *approach* to qualitative research or to a set of *methods* to be used in the analysis of qualitative data in the context of a different methodological approach.

Key Characteristics

Despite the emergence of differing perspectives in grounded theory, and taking due account of the varying emphases placed within these perspectives, it is possible to identify some fundamental features of grounded theory research. The foremost of these features is a reliance on an *inductive* approach to data analysis. Instead of identifying a number of broad theoretical concepts or categories in advance and then applying these to the data, as occurs in some forms of thematic content analysis, grounded theory insists that concepts and categories should be identified from the data. The resulting theory is thereby "grounded" in the data.

This model of analysis requires the researcher to engage in a particularly close and detailed reading of the data, applying analytical codes to particular pieces of data, prior to subsuming these codes under broader theoretical categories. It further implies that the analyst should return to the data repeatedly so as to check these codes and categories and the emerging theory—data analysis and theorizing thereby constitute an iterative process involving *constant comparison*. Specifically, constant comparison entails repeatedly comparing instances of data within a category with other instances in that category. The meaning of the category is thereby tested, developed, and refined as appropriate. Moreover, categories may be renamed and may be restructured in the process; it may be decided that what was one category should become two or that two categories should be merged to form a single category.

Two other key characteristics of grounded theory are *theoretical sampling* and *saturation*. Theoretical sampling means that, as a study proceeds, new sources of data—participants, situations, or social contexts—are specifically selected in relation to the theory that is being developed through the analysis of the data, so as to develop or refine emerging categories, or to elaborate the relationship between them. Imagine a researcher conducting a grounded theory study on the development of moral sensitivity among social workers. After interviews with a number of participants, certain analytical categories have been identified, including one labeled “empathy.” In order to develop a fuller understanding of the theoretical concept represented by this category, the researcher might specifically seek out practitioners whose work has placed them in particularly stressful situations, in which the notion of empathy might be expected to be powerfully illustrated. Similarly, the researcher might have discerned an emerging theoretical relationship between two categories—“vulnerability” and “cruelty”—and might look to interview practitioners who work either with children or with older people in order to gather data that might be expected to elucidate this relationship. An important aspect of theoretical sampling is deliberately seeking out *deviant cases* (also referred to as *negative cases*), which are sources of data that may question or disconfirm aspects of the theory emerging from the data.

Saturation also has to do with sampling but is a means of determining when additional data are no longer necessary and the process of data collection can be terminated. Accordingly, saturation may be considered to have occurred when a particular category has sufficient examples of data within it; there are sufficient examples of the category within the data to establish its meaning, and no new properties of the category are likely to be revealed through further examples. Equally, saturation may indicate the point in data collection at which no new categories, or relationships between categories, emerge (e.g., while further examples of existing categories may still be found, the data are no longer suggesting fresh categories or additional theoretical links between existing categories).

The Process of Analysis

The data analyzed in grounded theory studies most often come from semistructured or in-depth interviews, though they may also derive from notes made during participant observation or from textual sources such as diaries. The data may also come from focus group transcripts, though these often do not lend data may also come from focus group transcripts, though these often do not lend themselves to a fully inductive analysis.

There are varying accounts of how such data should be analyzed within grounded theory and specific contrasts between Glaserian and Straussian approaches; what follows will not draw exclusively from either the Glaserian or the Straussian model. The first stage, however, is the attaching of codes to pieces of data, whether paragraphs, sentences, phrases, or individual words. This is normally referred to as *open coding* and begins almost as soon as the data are collected. These codes are conceptual labels, and these labels may initially be quite descriptive and are often expressed in terms of the language that the research participant has used (referred to as *in vivo* codes). During the process of constant comparison, the meaning of codes, and thus the labels attached to them, often change, and codes are characteristically named in more theoretical terms.

Following open coding, a process of *axial coding* (Strauss) or *selective coding* (Glaser) takes place. Here, open codes are further developed in terms of key conceptual categories, including those that play a central role in the development of theory—*core categories*. A core category is an overarching category that represents a central concept or phenomenon shared by a number of lower order categories. To return to the earlier example, a core category called “moral action” might be considered to encapsulate, conceptually, the essence of a number of other categories, such as “doing one’s duty,” “acting on another’s behalf,” “removing or protecting from harm,” and “providing care.”

As theoretical insights develop, and the relationship of individual codes to broader conceptual categories is clarified, codes may be merged or split as part of the constant comparative process. As an extension of this, analytical relationships between codes or categories are proposed, as part of what Glaser calls *theoretical coding*. This process is assisted by what Glaser calls *coding families* and Strauss calls *coding paradigms*. Broadly, these are sets of abstract concepts that guide the conceptual development of open codes into conceptual categories in axial or selective coding, and the creation of theoretical relationships between these categories (and the core categories in particular), which will in turn be integrated within the theory that results from the analysis. Examples of a Straussian coding paradigm are causal conditions; action/interaction strategies; context; intervening conditions; and consequences. Despite the similarity of function between coding and families and coding paradigms, the different ways in which they were operationalized by Glaser and Strauss were the basis of much of the divergence that developed between their approaches to grounded theory, and in particular of Glaser’s claim that the Straussian model was too prescriptive.

Throughout the process of analysis, the researcher can utilize memos and diagrams. Memos are a reflective written record of the ongoing process of analysis and theorizing, allowing evolving insights to be clarified, reflections on field notes to be made, questions to be posed by the analyst to him-or herself, and a record to be kept of the sequence of decisions made in the process of analysis. As the emergence of new theoretical insights may cause insights developed earlier in the analysis to be revisited and revised, memos greatly assist this process. Diagrams are conceptual visualizations of the data. They assist the analyst in developing relationships between codes and generating theoretical propositions and may incorporate coding paradigms to facilitate this process.

As the grounded theory approach has to do with discovering or building theory, the role of prior theoretical understanding is important. It is sometimes claimed that grounded theory precludes one from conducting a literature review at the outset of a project, as to do so would be incompatible with the goal of discovering theory from, rather than imposing it upon, the data. Although this injunction does appear on page 37 of *The Discovery of Grounded Theory*, it should not be taken too literally. It does, however, highlight the need to avoid entering a study with clear theoretical preconceptions or expectations, and in particular, the importance of not determining analytical codes or categories in advance of collecting the data. Importantly, engaging with the literature once a theory has begun to emerge from the data is generally seen as an important part of the analytical process in grounded theory.

Other Considerations

Grounded theory originated within the symbolic interactionist tradition within sociology, but it does not presuppose a particular theoretical perspective. Recently, an approach to grounded theory centered in social constructionism has been developed by Cathy Charmaz. Rather than seeing theory as being “discovered” in data by the analyst, in a manner that Charmaz terms “objectivist,” her model of grounded theory regards theory as the “created” product of a more reflexive and relativist engagement with the data, and one in which theory is cocreated by the analyst and the participant. Adele Clarke, meanwhile, has developed a form of grounded theory based on situational analysis that adopts a postmodernist ecological perspective and, through the use of “maps,” incorporates more contextual and structural notions of social life alongside those related to agency.

Reflecting the fact that grounded theory can refer both to an overall approach to qualitative research and to a set of methods, many of the techniques utilized in grounded theory may be employed in other approaches to qualitative analysis. For example, memos and diagrams, or their equivalent, can find a place in a variety of styles of qualitative analysis, particularly those that are interpretive and reflexive in nature. Similarly, a form of theoretical sampling may be used in other types of research where the research design is emergent in response to ongoing data analysis, such as in forms of ethnography, and the way in which concepts are generated from the data may be reflected in other approaches that are essentially inductive, such as interpretative phenomenological analysis.

Finally, the process of analysis in grounded theory can be facilitated by the judicious use of computer software. NVivo and ATLAS.ti are two examples of programs that are commonly used for this purpose. Although these programs can assist data analysis by helping to organize the data and represent the relationships between codes and categories, they are not a substitute for the researcher’s own analytical judgment and theoretical sensitivity.

Julius Sim

See also Constructivist Approach; Interviews; NVivo; Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Clarke, A. E. (2005). Situational analysis: Grounded theory after the postmodern turn. Thousand Oaks, CA: Sage.

Corbin, J., & Strauss, A. (2015). Basics of qualitative research: Techniques and procedures for developing grounded theory (4th ed.). Thousand Oaks, CA: Sage.

Glaser, B. G. (1992). Basics of grounded theory analysis: Emergence vs forcing. Mill Valley, CA: Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago, IL: Aldine.

Ronli Diakow Ronli Diakow Diakow, Ronli

Growth Curve Modeling Growth curve modeling

756

759

Growth Curve Modeling

Growth curve modeling is a statistical method for analyzing change over time using longitudinal data. Data collected from individuals at multiple time points is used to analyze trends over time and variation in changes over time among individuals. Growth curve models focus both on similarities among individuals, captured by the mean structure, and on differences among individuals, captured by the covariance structure. The model can also be extended to explain change over time and variations in that change in terms of other factors. This entry further describes growth curve modeling, then discusses unconditional growth curve models and extensions to these models.

Growth curve modeling has also been called latent growth curve modeling, latent growth modeling, and latent curve analysis. The word *growth* is used because positive change over time, for example, increases in scores on an achievement test. The word *curve* is used to reflect the focus on the shape of the change over time, even when linear change is assumed. The word *latent* is sometimes used to emphasize the fact that the parameters for the between-person model for change are modeled using latent variables in some of the statistical models.

Models for longitudinal data when the outcome of interest is observed have been developed in both hierarchical linear modeling (HLM) and structural equation modeling (SEM) frameworks. In HLM, observations at different time points are considered to be nested within individuals, just as, for example, students are considered to be nested within schools. This leads to a two-level growth curve model, with time points at Level 1 and individuals at Level 2. In SEM, the covariance among the repeated observations is modeled and the latent variables that account for the relationships among observed variables represent the parameters of the growth curve.

Unconditional Growth Curve Models

HLM

The application of HLMs to longitudinal data considers repeated measurements as being clustered within individuals. When examining change over time for individuals, the focus is on a model for change within an individual, such as a model where the outcome at each time is regressed on a function of time. In order to compare changes over time among individuals, the focus is on a model for average change across the population. This leads to a two-part representation for change containing both a model for individual growth and a model for individual differences in growth. These two parts fit naturally within an HLM framework, with a within-subject Level 1 model for individual growth and a between-subject Level 2 model for variation in growth between individuals.

An HLM for linear growth is given as: Level 1:

$$ytp = \beta 0_p + \beta 1_p \,\text{time}_{tp} + \epsilon_{tp}.$$

Level 2:

$$\beta 0_p = \gamma 00 + u0_p . \beta 1_p = \gamma 10 + u1_p.$$

The Level 1 model (Equation 1) specifies individual growth curves, and the Level 2 models (Equations 2 and 3) specify the population growth curve. y_{tp} is the observed measurement for person *p* at time *t*, and time$_{tp}$ gives the timing of the measurement occasions. β0$_p$ is a person-specific intercept, often called the initial status; γ00 gives the mean at the initial time, and u0$_p$, called the random intercept, represents person-specific deviations from that mean. β1$_p$ is a person-specific rate of change, often called the rate of change or growth rate; γ10 is the mean rate of change, and u1$_p$, called the random slope, are person-specific deviations from that mean. $\in_{tp}$ are occasion-specific deviations from the person's growth curve.

Typically, model parameters are estimated using maximum likelihood, which requires the additional assumptions of normally distributed random effects (including error terms). $u0_p$ and $u1_p$ are assumed to follow a multivariate normal distribution with means of 0, variances of ψ00 and ψ11, and covariance of ψ10. $\in_{tp}$ is assumed to follow a normal distribution with mean 0 and variance τ. $\in_{tp}$ is assumed to be independent of $u0_p$ and $u1_p$. This model can be estimated using any of the standard software for HLMs.

Interpretation of the model focuses on the parameters for between-person growth. Note that all individuals are assumed to have growth curves of the same shape, as in their average, but the magnitude of growth can vary. γ00 and γ10 describe the shape of the average growth curve. ψ00 and ψ11 describe the extent to which individual growth curves vary around the mean growth curve. Smaller variance indicates more similarity among individuals while larger variance indicates more differences.

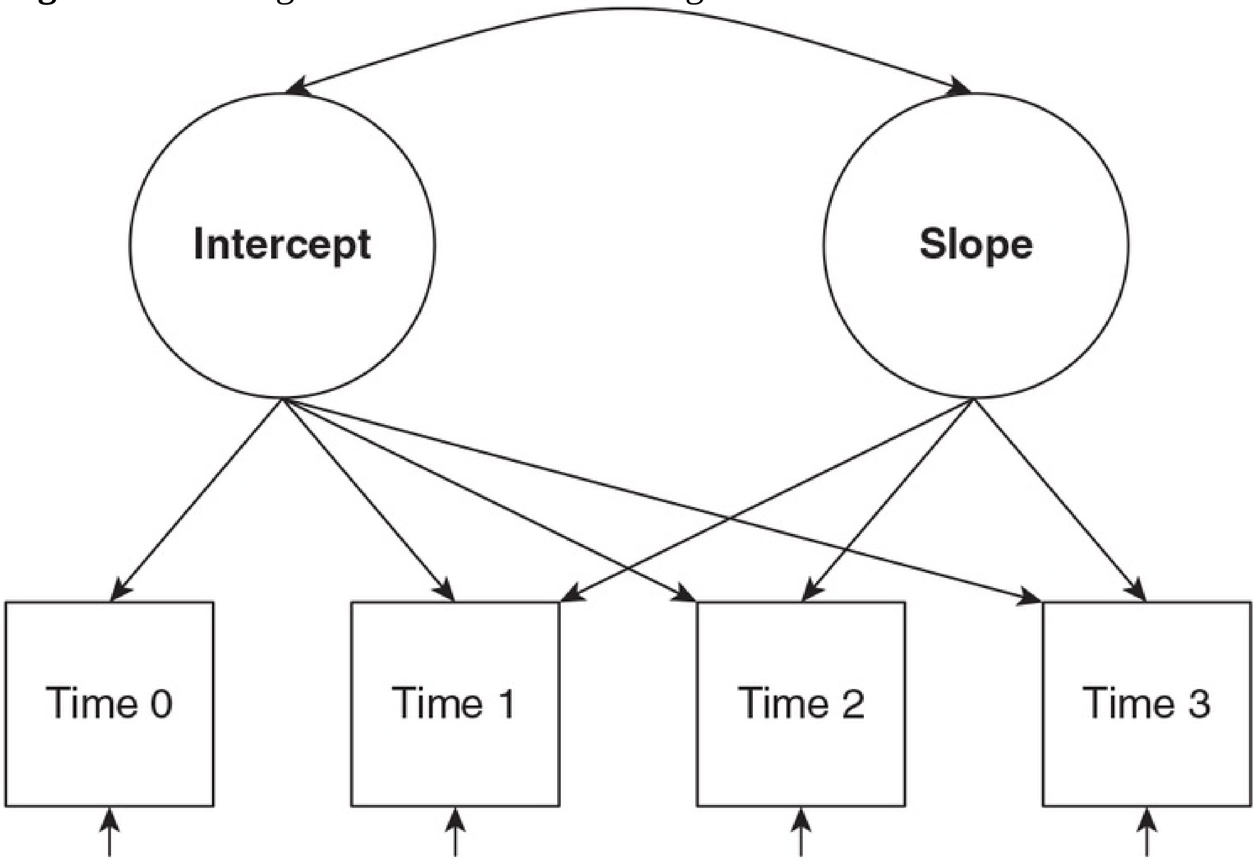
The growth curve model in Equations 1 to 3 models linear growth. Nonlinear growth can be accommodated in a number of ways. Two of the most common are including polynomial transformations of the time variable as additional terms in the model and including indicator variables for separate, nonoverlapping sections of the timescale as variables in the model to model piecewise linear growth. Because the timing of the measurements is included as a covariate, it is not necessary to have equally spaced measurement occasions or to have all individuals observed at the same time points (i.e., it is unnecessary to have balanced data).

SEM

The application of structural equation models to longitudinal data considers the covariance of repeated measurements over time. The same considerations that prompted the development of HLM models for longitudinal data, modeling both individual growth and a structure for combining individual growth functions, motivated SEM models for longitudinal data. John McArdle and colleagues gave the name *latent (growth) curve analysis* to the procedure.

A structural equation model for growth is represented by the path diagram in Figure 1. The model contains two latent variables (in circles), labeled *intercept* and *slope*, to account for the correlations between the observed measurements (in squares) at each time. These latent variables are called the growth factors; the intercept is often also called the initial status. The loadings (i.e., arrows) from the intercept to each observed variable are fixed at 1. The loadings from the slope to each observed variable are fixed equal to the observation time (i.e., at 0, 1, 2, 3, resulting in a linear growth pattern); this requires the same timing of the measurement for each person. The mean structure and covariance structure for the intercept and slope latent variables are estimated. The residual variances of the observed variables are also estimated and are typically not constrained to be equal across occasions. This model can be estimated using any of the standard software for SEMs.

Figure 1 Path diagram for an unconditional growth curve model



The shape of the curve is defined by the loadings (represented by the arrows in the figure) from the latent variable representing the slope (labeled *slope*) to the observed variables (represented by the squares). As in Figure 1, these are often restricted to linear. However, these loadings can be freely estimated, resulting in an estimated shape for the growth curve. Alternatively, growth curves with specific functional forms can be modeled by adding additional latent variables and by fixing the loadings to other values, for example, by adding a third latent variable to represent a quadratic term and fixing the loadings at 0, 1, 4, 9, and so on. Because the measurement timing is built into the structural part of the model, standard growth curve models in the SEM framework assume that all individuals are measured on the same occasions or with the same spacing between measurement occasions.

Connection Between HLM and SEM

In general, models for longitudinal growth under the HLM and SEM frameworks are mathematically equivalent because each is subsumed under the broader statistical framework of generalized latent variable modeling. For example, the model in Equations 1–3 is identical to the model in Figure 1; in the latter model, the slope loadings fixed for linear growth and residual variances constrained to be equal across time. Applying either of these models to a data set will yield the same parameter estimates. In an SEM framework, the Level 1 equations (e.g., Equation 1) are called the measurement model and the Level 2 equations (e.g., Equations 2 and 3) are called the structural model. The random intercept ($\beta0_p$) and slope ($\beta1_p$) in Equation 1 correspond to the two latent variables in Figure 1 (labeled intercept and slope, respectively). The article by John Willett listed in the further readings at the end of this entry provides more detail on the correspondence between longitudinal data analysis using HLM and SEM.

Extensions

There are a multitude of extensions to the unconditional growth curve models presented earlier. Most of these extensions can be applied regardless of which modeling framework is being used. However, some extensions are more straightforward to accomplish under one of the paradigms and in software designed for one type of model. Adding additional levels to account for other forms of clustering, such as individuals within classrooms or schools, is very straightforward if using an HLM. Estimating the effects of other latent variables on the growth process is straightforward if using SEM by embedding the growth curve within a larger structural model.

Additional (observed) covariates can be added to the growth curve model under either framework. Both time-invariant covariates (e.g., characteristics of the individuals such as gender) and time-varying covariates (e.g., characteristics that could change across occasions such as number of hours spent studying) can be included. These covariates explain variation in growth between individuals and across time. Time-invariant covariates, which explain variation in growth between individuals, are entered in the Level 2 or structural equations for the growth curve model and explain differences in the growth parameters such as the intercept and slope. Time-varying covariates, which explain variation in growth within and between individuals, are entered in the Level 1 or measurement equations for the growth model.

The growth curve model presented earlier assumed continuous observed outcomes (y_{tp}). The model can also be extended to account for dichotomous or categorical observed outcomes. Under the HLM framework, this is usually done by changing from a linear regression framework to one with a different link function, such as logistic or probit regression. Under the SEM framework, this is usually done by estimating the model based on polychoric (rather than Pearson) correlations. The model could also be extended to account for a latent outcome that is observed using multiple indicators. Under the HLM framework, this is usually done by incorporating a measurement model as the lowest level. The growth curve model is then a three-level model for items, occasions, and individuals. Under SEM, this is also done by expanding the measurement model to incorporate multiple indicators. This model has been given a variety of names including a curve of factors model, a longitudinal model with multiple indicators, and a second-order latent growth model. Additional possible extensions include modeling more complex residual structures, relaxing the assumptions of normality, and growth mixture modeling.

Ronli Diakow

See also Generalized Linear Mixed Models; Hierarchical Linear Modeling; Repeated Measures Designs; Structural Equation Modeling

Further Readings

Bollen, K. A., & Curran, P. J. (2006). Latent curve models: A structural equation perspective. Hoboken, NJ: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. Psychological Bulletin, 101(1), 147–158.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. Child Development, 58(1), 110–133.

Muthen, B. O., & Khoo, S. T. (1998). Longitudinal studies of achievement using latent variable modeling. Learning and Individual Differences, 10(2), 73–101.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). Latent growth curve modeling. Los Angeles, CA: Sage.

Raudenbush, S. W. (1989). The analysis of longitudinal, multilevel data. International Journal of Educational Research, 13(7), 721–740.

Willett, J. B. (1988). Questions and answers in the measurement of change. Review of Research in Education, 15(1), 345–422.

Robert L. Johnson Robert L. Johnson Johnson, Robert L.

Bradley D. Rogers Bradley D. Rogers Rogers, Bradley D.

Guiding Principles for Evaluators Guiding principles for evaluators

759

762

Guiding Principles for Evaluators

Evaluation in its simplest form is the act of ascertaining the amount, value, or effectiveness of an object or action. In this informal sense, it is an activity that is exceedingly common and nearly universally practiced. Most everyday evaluative actions involve relatively low cost items or activities and can, therefore, afford to lack the scrupulousness of a systemic methodology that is desired when evaluating something of more consequence. The evaluative rigor a person might use when acquiring a house, for example, will likely be substantially different than that used when purchasing a piece of fruit from a market.

Professional educational evaluation needs to be even more scrupulous. A disciplined system informed by guiding principles helps to ensure the quality and consistency of the evaluative process, a quality and consistency that is much desired with higher stakes evaluations such as those conducted in an educational setting. This entry briefly describes the development of educational evaluation and the process that led to the development of a formal set of guiding principles for evaluators. The entry concludes with an overview of each of the guiding principles.

Background and Development

Educational evaluation traces its origin to the domestic policies of the administration of U.S. president Lyndon Johnson. Under his leadership, the United States enacted a set of laws with the express purpose of eliminating poverty and ameliorating a host of other social ills; these initiatives are often collectively known as the Great Society. The federal government invested millions of dollars into programs in education, health care, urban renewal, housing, and other similar areas.

Unlike market-based enterprises that can rely on natural external markers of success, as well as internal and external systems that provide constant feedback, public sector programs often lack intrinsic mechanisms with which to ensure effective allocation of funding and means by which to judge their success. Members of Congress expressed concerns regarding these issues during debate on a key part of the Great Society legislation, the Elementary and Secondary Education Act (ESEA) of 1965. To address these concerns, Congress included in the ESEA a requirement that each grant recipient file an evaluation report that detailed the specific results of the program. This requirement of the ESEA is generally recognized as the event most responsible for the development of modern program evaluation.

Since the passage of the ESEA, program evaluation has developed into a full-fledged professional discipline. In the wake of the ESEA's evaluation requirement, universities established programs specializing in training professional evaluators. In 1976, 12 professional associations concerned with ensuring the quality and consistency of program evaluation created the Joint Committee of Standards for Educational Evaluation and tasked it with developing a set of standards to be used by professional evaluators. The fruits of this endeavor, *The Standards for Evaluations of Educational Programs, Projects, and Materials*, was published in 1981. The third edition of this work was published in 2010. In addition to the Joint Committee of Standards for Educational Evaluation standards, a set of ethical guidelines was developed by the Evaluation Research Society and was published in 1982. These guidelines were updated and revised in 2004 and are made available by the American Evaluation Association as the *Guiding Principles for Evaluators*.

Evaluator Principles

This section chiefly focuses on summarizing the American Evaluation Association guiding principles, but its description of these principles is informed by the Joint Committee of Standards for Educational Evaluation standards and the experience of evaluators. It should be noted that these are guiding principles and are not intended to be dictates but are the result of dialogue and consensus, and as such, they are subject to revision and continued refinement. Further, it must be recognized that the principles are themselves products of their specific historical moment and location, namely, the United States of the early 21st century. That is not to suggest that guiding principles are without value in other cultural settings, it is merely to acknowledge that circumstances are different from country to country and epoch to epoch.

Competence

It has been said that it is a poor craftsman who blames his tools. The proverb is meant to convey that there is an essential personal responsibility and expertise that is necessary for quality work. This is generally true of most crafts and is no less true for professional evaluation. This notion is reflected in the guiding principle of competency. Principal evaluators as well as each member of the evaluation team must possess the appropriate education, experience, skill, and training needed to effectively execute the tasks required to properly conduct an evaluation. An effective evaluation can only be produced by effective evaluators working within their circumscribed areas of expertise. Although perhaps counterintuitive, it is good practice for evaluators to decline evaluations or tasks that are substantially outside the range of their education, ability, or skill.

It is incumbent upon evaluators to not only recognize the bounds of their competencies but also to seek out continuing education and training opportunities that will serve to maintain and expand those competencies. Formal and informal means of improving one’s skills can readily be found through participation in a community of practicing evaluators; this can typically be facilitated by joining a professional organization. In addition, competent evaluators are better able to project and maintain the credibility necessary to engender the trust of the various stakeholders as well as that of the entity that is being evaluated (i.e., the evaluand).

Accuracy and Credibility

In addition to general competency, evaluators must make every effort to ensure the accuracy and credibility of their evaluations by conducting data-based, systematic inquiries. The methodologies of inquiry employed should be appropriate to the questions posed and, importantly, should conform to the highest technical standards of a given methodology. It is left to evaluators to use their own expert judgment to decide the appropriateness of a methodology; the guidelines avoid wading into any methodological controversy and thus do not make recommendations regarding this issue.

It is good practice for evaluators to discuss with the evaluand the relative weaknesses and strengths of the evaluation questions and the methods of inquiry that will be used to answer them. It is imperative that this is done throughout the evaluation process and in a clear, understandable, and contextually appropriate manner. Evaluators likely will find it useful to be involved with the evaluand from the planning stages of a program. This will allow the evaluator to assist program staff with formulating answerable evaluation questions and establishing realistic evaluation objectives. Further, evaluators can provide feedback on the design of the project that can help facilitate the execution of systematic inquiry methodologies.

Honesty and Integrity

The next guiding principle suggests that evaluators should act with honesty and integrity both in their own personal behavior and in such a manner that ensures the integrity of the entire evaluation process. The responsibility to deal honestly begins with the initial negotiations with the evaluand concerning the terms of the evaluation. The cost of the evaluation, tasks to be performed, limitations of methodologies, and scope and usage of the results must all be clearly communicated to the evaluand and relevant stakeholders during this initial process. The responsibility to initiate a discussion of these issues lies not with the evaluand or the stakeholders but rests solely upon the evaluator.

Before accepting the evaluation project, the evaluator should disclose any potential or actual conflicts of interest. If the evaluation project is accepted despite a conflict or the appearance of a conflict, these should be clearly noted in any and all reports produced for the evaluation; to fail to report potential conflicts risks impugning the integrity of any findings of the evaluation, no matter how well established. It is possible that in some cases, a conflict could be implicit to the evaluation itself, for example, if the evaluation is funded by an entity that has a vested interest in specific outcomes. A situation such as this should be handled similarly to any other conflict, that is, it should be disclosed.

Evaluators should make every attempt to adhere to the agreed upon contract; however, situations may arise that require changes be made to the originally negotiated project plan. Should this occur, a careful record of all changes must be kept. Further, if certain changes are found to affect the scope or results of the evaluation, then prompt disclosure to the evaluand and relevant stakeholders must be made of those changes.

Acting with honesty and integrity must not be equated with fidelity to the evaluand. It is possible that at times the wishes of the evaluand may run counter to the demands of acting honestly or ensuring the integrity of the evaluation. Evaluators must make every effort to ward against any misrepresentation of their procedures, data, or findings. If it is determined that a procedure or activity has a likelihood of producing misleading data or might result in specious conclusions, then it is the responsibility of the evaluators to communicate their concerns to the evaluand. If the concerns of the evaluator are not allayed, then the evaluator should resolve to decline to conduct the evaluation.

Respect

Throughout the entire process of conducting an evaluation, from contract negotiation to final report, the guiding principles recommend that evaluators should be careful to hold in highest regard the corporate and individual dignity and worth of all those participating in or affected by the evaluation. The kind of respect here intended extends beyond sentiment or politesse; it is an ordered process with real implications and goals.

Evaluators should begin by seeking a thorough apprehension of the context within which the evaluation is to be conducted. Every evaluation will consist of unique contextual factors that can affect the evaluation process, the way it is perceived by stakeholders, and even the results. These factors can include, among others, timing, location, political climate, and economic conditions. More narrowly, evaluators should also seek to understand the differences among the various individual participants in the evaluation, such as differences in culture, religion, gender, and age. Any potential implications of either the contextual factors or individual differences must be accounted for and should inform all stages of the evaluation.

Another element of respect relates to the management of risk and the potential for harm. Evaluators should adhere to all professional standards related to risk of harm and informed consent. In brief, risk of harm to the evaluand or stakeholders should be minimized and benefits maximized without compromising the integrity of the evaluation or its findings. In addition, the evaluand and all participants should be informed of any risk of harm and their consent granted. This includes disclosure of the level and limits of confidentiality that can be expected by the participants. Those participants who bear the risk of harm must do so willingly and must be made aware of any opportunities to receive the benefits of the evaluation. Also, to avoid the appearance of coercion, it should be clearly communicated that eligibility to receive benefits or services related to the evaluation does not depend upon participation in the evaluation.

General Welfare

Building on the principle of respect is the larger responsibility of evaluators to the public welfare. Although evaluators should have a special relationship with the evaluand, a balance must be maintained between its interests and needs and those of other stakeholders as well as the general public. The wide diversity of interests and values of the full range of stakeholders should be taken into account from the planning to reporting stages of the evaluation. Essential to ensuring that these various interests and values are respected is the full freedom of information as far as the restrictions of confidentiality will allow. Information should be disseminated to stakeholders and to the public as frequently as is reasonable and should be uniquely communicated in such a manner so that it is clearly intelligible by the intended group. It is possible that at times the interests of any of the various groups (funder, evaluand, and other stakeholders) will be in conflict. In these instances, it is good practice for evaluators to look beyond the narrow interests of any particular stakeholder and take in to consideration the general societal welfare. This obligation is especially pertinent when an evaluation is publicly funded as is frequently the case in educational evaluation.

Robert L. Johnson and Bradley D. Rogers

See also Conflict of Interest; Ethical Issues in Evaluation; Evaluation; Evaluation, History of; Evaluation Versus Research; External Evaluation; Great Society Programs; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Alkin, M. C. (2011). Evaluation essentials: From A to Z. New York, NY: Guilford.

American Evaluation Association. (2004). Guiding principles for evaluators. Retrieved from http://www.eval.org/p/cm/ld/fid=51

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). Program evaluation: Alternative approaches and practical guidelines (3rd ed.). Boston, MA: Pearson.

House, E. R. (1995). Principled evaluation: A critique of the AEA guiding principles. New Directions for Program Evaluation, 1995(66), 27–34.

Kellaghan, T., & Stufflebeam, D. L. (2003). International handbook of educational evaluation. Dordrecht, the Netherlands: Kluwer Academic.

Mertens, D. M., & Wilson, A. T. (2012). Program evaluation theory and practice: A comprehensive guide. New York, NY: Guilford Press.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). The program evaluation standards: A guide for evaluators and evaluation users. Thousand Oaks, CA: Sage.

Richard R Sudweeks Richard R Sudweeks Sudweeks, Richard R

Guttman Scaling

Guttman scaling

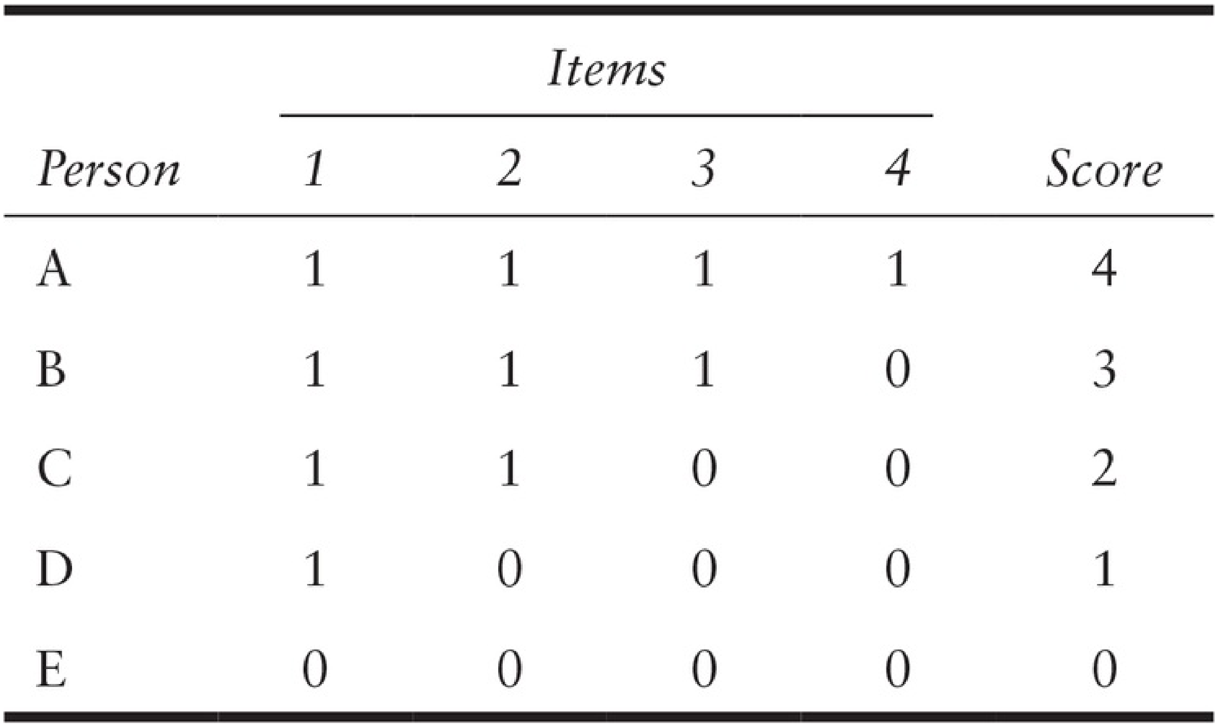
763

766

Guttman Scaling

Guttman scaling or scalogram analysis was developed by Louis Guttman in an attempt to find a way to measure attitudes that would improve on what he perceived to be the limitations of Thurstone scaling and Likert-type scaling. Guttman believed that an individual’s attitude toward some psychological object could be measured by presenting the person with statements that had been ordered in terms of their favorableness or unfavorableness toward the target object. He theorized that a perfect scale would consist of a set of statements that were hierarchically cumulative in the sense that an individual who endorsed a particular statement would also endorse all less extreme statements in the set and that an individual who failed to endorse a given statement would not endorse any statements representing more extreme feelings about the target object. This entry describes the purpose of Guttman scaling, the computation and meaning of the coefficients of reproducibility and scalability, and the differences between Guttman scaling and Likert-type scaling.

Table 1 shows an example of a perfect Guttman Scale consisting of responses from five individuals to four different statements. In this example, the responses to the various items are coded dichotomously: 1 = *agree*, and 0 = *disagree*. An individual’s score is determined by how many statements were endorsed by that person.



	Items				
Person	*1*	*2*	*3*	*4*	*Score*
A	1	1	1	1	4
B	1	1	1	0	3
C	1	1	0	0	2
D	1	0	0	0	1
E	0	0	0	0	0

In a perfect Guttman Scale, each possible score is associated with one and only one pattern of responses. Consequently, once a person’s score is known, it is possible to exactly reproduce that individual’s response pattern. For example, if Person B received a score of 3, we know that Person B responded favorably to Items 1, 2, and 3 and did not endorse Item 4. Similarly, if Person D received a score of 1, we know that Person D endorsed Item 1 and none of the other items. This characteristic that each possible response pattern is associated with a unique score distinguishes Guttman Scales from both Likert-type scales and Thurstone Scales.

Measurement Error and Coefficient of Reproducibility (CR)

The claim that each different possible score is associated with a unique pattern of responses assumes that there is no measurement error in the data. In reality, perfect Guttman Scales do not exist because measurement error is always present to some degree. Each of the five response patterns shown in the rows of Table 1 (1111, 1110, … 0000) are perfectly consistent with Guttman’s theory, but in a real application, there will likely be at least some inconsistent or mixed response patterns such as 1101 or 1010 that do not fit Guttman expectations. Guttman recognized this problem, but he believed that a perfect scale could be approximated as long as the percentage of errors was relatively low. He defined an *error* as any deviation of an observed response from the ideal response that would have been expected by the cumulative model. In addition, he proposed a statistic that could be used to describe the degree to which a set of data was free from deviant responses. He called this statistic the *CR* and defined it as follows:

$$\text{CR} = 1 - (\text{number of errors})/(\text{total number of errors}).$$

CR can be interpreted as the percentage of the total number of responses that can be accurately predicted from knowledge of the total scores. If the value of this coefficient exceeds 90% for a set of items, then Guttman would claim that the set is scalable.

Application of Guttman's coefficient necessitates an operational definition of what counts as an error. According to Guttman's definition, the number of errors in a response pattern is the least number of positive responses that would need to be changed to negative or the least number of negative responses that would need to be changed to positive in order to convert the observed pattern to the ideal pattern. However, his rule for counting errors was later shown to be inadequate and has been replaced by a refined version originally advocated by Ward Goodenough and later by Allen Edwards.

According to the Goodenough–Edwards perspective, the ideal response pattern for a respondent is best predicted by the number of items to which the individual responded positively. The Goodenough–Edwards rule for counting errors can be illustrated by examining an inconsistent response pattern such as 1010. Because a person manifesting this pattern endorsed two statements, the person ideal response pattern would have been 1100 indicating that the person would have responded positively to the first two items and negatively to the remaining two. In this example, two changes would be needed to transform the observed pattern to the ideal pattern, namely, the 0 in the second position would need to be changed to 1 and the 1 in the third position would have to be changed to 0. This refined definition of what counts as an error is more consistent with Guttman's cumulative interpretation of scaling theory than the definition he used.

Edwards subsequently showed that satisfying Guttman's 90% criterion was not a sufficient condition for assessing scalability. He demonstrated that the CR is influenced by the proportion of responses in the modal category (the category which has the most responses) of each item and that an artificially high coefficient with dubious meaning can result if the distribution of responses to the items is highly skewed. Edwards argued that in order to properly interpret CR, a researcher needs to know how small it can be given the observed distribution of responses to each item. He developed the minimal marginal reproducibility (MMR) statistic to estimate the lower bound of CR given the observed data. He defined MMR as the mean proportion of responses in the modal category averaged across the *K* items. The value of MMR can be interpreted as the smallest value of CR that is possible given the observed proportion of respondents who agreed and disagreed with each statement.

Table 2 displays the proportions involved in computing MMR for a scale consisting of 4 items responded to by 10 persons. In this example, 10 persons have responded to 4 items. The data are coded: 1 = *agree* and 0 = *disagree*. The cell entries in the row labeled *p* report the proportion of respondents to each item who responded “agree.” Similarly, the entries in the row labeled *q* indicate the proportion who chose “disagree” in response to each item.

Example 2:

	Items				*Errors*	
Person	*1*	*2*	*3*	*4*	*Guttman*	*G-E*
A	1	1	1	1	0	0
B	1	1	1	0	0	0
C	1	1	0	0	0	0
D	1	1	1	0	0	0
E	1	1	1	0	0	0
F	1	1	0	1	1	2
G	1	1	1	1	0	0
H	1	1	0	0	0	0
I	0	1	1	0	1	2
J	0	0	0	0	0	0
p	.8	.9	.6	.3		
q	.2	.1	.4	.7		
Number of errors					4	
Total number of responses					40	
Sum of nonmodal frequencies					10	
Percentage of errors					10.0%	
Coefficient of reproducibility					90.0%	
Minimum marginal reproducibility					75.0%	
CR-MMR difference					15.0%	
Coefficient of scalability					60.0%	

Note: CR-MMR = coefficient of reproducibility–minimal marginal reproducibility.

Three of the cells in the *p* row and one of the cells in the *q* row are shaded. These shaded cells indicate the proportion of respondents in the modal category for each item. Hence, MMR is (.8 + .9 + .6 + .7)/4 = .75 or 75%. Comparing the observed value of CR (90%) with MMR (75%) reveals that the reproducibility of the data in this example is 15% larger than the minimum possible value of CR given the proportion of persons who agreed and disagreed with each item.

Coefficient of Scalability (CS)

Herbert Menzel formalized a procedure for comparing CR and MMR by proposing a statistic called the *CS*. This new coefficient describes the degree of observed improvement in reproducibility (CR − MMR) divided by the maximum amount of improvement that would be possible (1 − MMR) given the proportion of respondents who agreed or disagreed with the various items.

$$\mathrm{CS} = (\mathrm{CR} - \mathrm{MMR})/(1 - \mathrm{MMR}).$$

Menzel agreed with the previous criterion that a set of data should have a CR value of at least 90%, but he prescribed the additional criterion that a set of data have a CS value of 60% or more in order to be scalable. This additional criterion means that the observed improvement should be at least 60% of the possible improvement.

The CR for the data in Table 2 was computed based on the Goodenough–Edwards definition of error. For the sake of contrast, the number of errors using Guttman’s original definition of error are also reported. Only 5% of the total responses are errors when Guttman’s definition is used. This means that the value of CR would have been 95% instead of 90% if Guttman’s definition had been used, but his definition underrepresents the actual number of errors in the data.

The data in Table 3 provide a contrast to the data in Table 2. Note the similarities and the difference in the shaded cells in the marginal frequencies (the *p* and *q* rows) in each data set.

Although the modal category values are the same in the two examples, the *p* and *q* values for Item 3 in the two examples are in reverse order. Because both data sets have the same proportions in the modal category for each item, they have the same MMR value. However, CR = 90% in Table 2, but in Table 3, CR = 75%. Furthermore, the CR for Table 3 is no larger than MMR. Hence, the value of CS for Table 3 is 0. The data in Table 3 fail to comply with both the minimum accepted CR value of 90% and the minimum accepted CS value of 60%.

Example 3:

	Items				
Person	*1*	*2*	*3*	*4*	*G-E Errors*
A	1	1	1	1	0
B	1	1	1	0	0
C	1	1	0	0	0
D	1	0	1	0	2
E	1	1	0	0	0
F	0	1	1	0	2
G	1	1	0	1	2
H	1	1	0	0	0
I	0	1	0	0	2
J	1	1	0	1	2
p	.8	.9	.4	.3	
q	.2	.1	.6	.7	
Number of errors				10	
Total number of responses				40	
Sum of nonmodal frequencies				10	
Percentage of errors				25.0%	
Coefficient of reproducibility				75.0%	
Minimum marginal reproducibility				75.0%	
CR-MMR difference				0.0%	
Coefficient of scalability				0.0%	

Note: G–E errors = Goodenough–Edwards definition of error; CR-MMR = coefficient of reproducibility–minimal marginal reproducibility.

The third criterion a set of data should satisfy in order to be scalable focuses on whether there is any evidence of more than one dimension in the data. Guttman anticipated that some response patterns would include deviant responses, but he assumed that the errors would be essentially random and therefore unsystematic. Therefore, the presence of any error pattern that occurs disproportionately is cause for concern. There is no single statistic for determining whether this criterion is satisfied. A workable procedure is to identify the various error patterns that occur in a set of data and then to compare their relative frequency and make a subjective judgment as to whether any patterns occur more frequently than would be expected due to random variation.

Guttman Scaling and Likert-Type Scaling

The items in a Guttman Scale need not be statements. They can also represent other definable characteristics of the target object. Polytomous items can also be used rather than dichotomously scored items. One advantage of the Guttman approach is that each scale score can be obtained from one and only one pattern of responses. In contrast, when Likert-type scaling is used, the same score may be obtained by persons representing several different patterns of responses.

Gutttman Scales have not been as widely used as Likert-type scales. The principal reason is that Likert-type scales are easier to construct. However, Guttman Scales have been successfully used (a) in anthropology to scale cultural characteristics such as household wealth, (b) in education to analyze the role of conceptual knowledge in procedural learning, (c) in political science to scale voting patterns of U.S. senators and the decisions of Supreme Court justices, (d) in psychology to scale depression, and (e) in sociology to scale stages of drug use.

Richard R Sudweeks

See also Likert Scaling; Scales; Thurstone Scaling

Further Readings

Edwards, A. L. (1957). Techniques of attitude scale construction. New York, NY: Appleton-Century-Crofts.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139–150.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. Educational and Psychological Measurement, 7, 247–249.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), Measurement and prediction (pp. 60–90). Princeton, NJ: Princeton University Press.

McIver, J. P., & Carmines, E. G. (1981). Unidimensional scaling. Newbury Park, CA: Sage.

Torgerson, W. S. (1958). Theory and methods of scaling. New York, NY: Wiley.

H

Timothy Franz Timothy Franz Franz, Timothy

Hawthorne Effect

Hawthorne effect

767

769

Hawthorne Effect

The Hawthorne effect is when research participants act in a way that is consistent with their perception of the researcher’s expectations during a study, which then biases the outcomes of that research study. For example, imagine that a researcher was conducting a study about one type of helping behavior—door-opening behavior. If participants for some reason know that they were being studied and also know that the purpose of the study is about helping, they may be more likely to help by opening doors for others because that is what they think the researcher expects (regardless of the actual purpose, hypotheses, or methods of the study) rather than walking through the door and paying little attention to whether they should hold it for the next person. This entry begins by providing a brief historical account of the original Hawthorne studies and the importance of their findings and then offers a summary of six ways to minimize the Hawthorne effect in research. The name for the Hawthorne effect comes from a series of classic research studies conducted in the 1920s and 1930s, which are collectively known as the Hawthorne studies. These studies began as an examination of the impact of the quality and amount of lighting on worker efficiency in the Hawthorne plant of Western Electric Company. The first of the studies controlled the level of the lighting and measured factory worker productivity. At points, they increased the level of the lighting, decreased it, or kept it the same. They expected that increased lighting would improve productivity on the factory line, but instead found that factory line productivity increased regardless of the level of illumination in the plant.

Because of these unexpected findings, they continued to study the employees in the plant by manipulating other human factors, such as temperature, humidity, work hours, incentives, rest, and fatigue. The researchers were surprised to find work hours, incentives, rest, and fatigue. The researchers were surprised to find the results showed an increase in productivity regardless of the particular variable being manipulated. As a result, the research team set up a controlled experiment in a relay test room and then interviewed over 21,000 employees.

These original Hawthorne studies provided two key findings that are still relevant today. The first, though not the most pertinent for this summary, is part of the historical foundation of the field of industrial and organizational psychology. The Hawthorne studies demonstrated the importance of the informal influence networks and social factors have on employee behavior in organizations. Prior to this, many considered the formal leadership structure to be the most important influence on employee behavior. The second key finding, and the more important one for this summary, is that the Hawthorne studies demonstrated the importance of participant expectations, or demand characteristics, on participant behavior. Specifically, the studies revealed that a participant will react to things other than just the variable manipulated by a researcher. It is clear from this series of studies that workers in the Hawthorne studies acted in a way that was inconsistent with the predictions and instead likely consistent with what they thought the researchers wanted.

Although there is considerable historical evidence that the Hawthorne Effect occurs, there is also evidence that Hawthorne Effects are potentially misinterpreted or overexaggerated. For example, in 1974, H. M. Parsons suggested that the Hawthorne effect can actually be interpreted through operant conditioning principles. In 1981, Dana Bramel and Ronald Friend reinterpreted the findings in terms of power and resistance. Finally, in 1989, a meta-analysis by John Adair, Donald Sharpe, and Cam-Loi Huynh showed that there is little to no evidence for the Hawthorne effect when statistically summarizing the results of 39 fairly recent research studies. Regardless, these authors still recommended avoiding research designs that may potentially induce participant bias and Hawthorne effects.

Minimizing the Hawthorne Effect in Research

There are many ways in which a researcher can reduce the likelihood of the Hawthorne effect occurring. Six of the more common research tools for avoiding Hawthorne effects are briefly summarized here.

Withholding the True Purpose

One of the easiest research tools to prevent Hawthorne effects is withholding from participants any mention of researcher expectations. As one part of giving consent to participate in a study, participants have an ethical right to know about the procedures (as well as the risks and benefits). However, this does not mean that a researcher must inform participants about the specific predictions, expectations, or research questions that are being examined. In fact, in many studies, it is best to leave the expectations out of any descriptions when describing the study background. When a researcher withholds the true nature of a study, participants have a lower probability of behaving in an expectation-consistent manner.

Deception

Withholding the true purpose of a study is, in part, deception by omission—a researcher is leaving out key information. This is different from research that involves true deception, which is actually purposefully deceiving participants about the true purpose of the study. Many research ethics guidelines allow for some level of deception in research as long as there is (a) no other possible method of gaining the information, (b) no additional harm generated, and (c) the possibility to debrief participants. Deception, when absolutely necessary, can obscure the nature of the study from participants and allow researchers to study a phenomenon in a way in which participants are unlikely to know the research expectations.

Placebo Control

In psychological and educational research, using a placebo control group is also an effective tool for understanding Hawthorne effects. Placebo control group designs are often confused with Hawthorne effects. However, a placebo is a control that allows researchers to examine a phenomenon in comparison to a treatment. Specifically, participants in a placebo control group have similar expectations as those in a treatment group. Thus, if both groups change similarly, it is possible that a Hawthorne effect has occurred. On the other hand, if the treatment group changes at a different rate than the placebo control group, a Hawthorne Effect is unlikely to have occurred.

Blind/Double Blind Study

Blind/Double-Blind Study

One of the most effective research tools to control for a Hawthorne Effect is a blind, or even better, a double-blind study. In a blind study (which may be used as part of a placebo-control design), participants are unaware of the treatment condition that is being studied. Even more powerful is a double-blind study, when neither participants nor the researchers collecting the data know about the treatment group(s) or placebo. In these situations, it is very unlikely that participant expectations may affect the findings because neither participants nor researchers know what those expectations are.

Naturalistic, Unobtrusive Observation

Naturalistic observation designs are also particularly powerful at avoiding participant bias and Hawthorne effects. Naturalistic observation is a design in which researchers observe, unobtrusively, people's behavior in the real-world settings. Further, those being observed are unaware that they are even in a research study. Although these designs are unable to provide any conclusions about cause and effect, they are powerful tools for studying people in their natural setting when they are behaving in a natural way rather than a way that is impacted by the study setting.

Multi-Method, Multi-Measurement Research Designs

As in all research, the best choice is triangulation, which means examining a phenomenon from multiple perspectives. These perspectives may include multiple ways to measure a phenomenon and (more importantly) multiple research studies/designs that test what we assume we know about people. For example, a researcher may study participant behavior using surveys, experiments, and naturalistic observation. If the findings are, for the most part, replicated across measures and methods, it is more likely that any findings about human behavior are an indication about how people think, feel, or act.

Future Directions

When people are studied in systematic ways, there is always the potential for bias. The Hawthorne effect is a part of a larger set of participant reaction biases, including demand characteristics and participant reactance, and other artifacts of the measurement process, such as mere measurement effects and question-behavior effects. These are all threats to the internal validity and, hence, meaningfulness of a research study.

Timothy Franz

See also Placebo Effect; Pygmalion Effect

Further Readings

Adair, J. G., Sharpe, D., & Huynh, C. L. (1989a). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. Review of Educational Research, 59(2), 215–228.

Adair, J. G., Sharpe, D., & Huynh, C. L. (1989b). Placebo, Hawthorne, and other artifact controls: An overview of a research program. Mathematical and Theoretical Systems, 1(1), 99–110.

Baritz, L. (1960). Hawthorne. The servants of power: A history of the use of social science in American industry (pp. 77–95). Middletown, CT: Wesleyan University Press.

Bramel, D., & Friend, R. (1981). Hawthorne, the myth of the docile work, and class bias in psychology. American Psychologist, 36(8), 867–878.

McCambridge, J. (2015). From question-behavior effects in trials to the social psychology of research participation. Psychology … Health, 30(1), 72–84.

Parsons, H. M. (1974). What happened at Hawthorne? New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies. Science, 183(4128), 922–932.

Roethlisberger, F. J., Dickson, W. J., & Wright, H. A (1939/1961). Management and the worker: An account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago (12th ed.). Cambridge, MA:

Harvard University Press.

Sommer, R. (1968). Hawthorne dogma. Psychological Bulletin, 70(6), 592–595.

Vera Lynne Stroup-Rentier Vera Lynne Stroup-Rentier Stroup-Rentier, Vera Lynne

Head Start

Head start

769

771

Head Start

Head Start is a program administered by the U.S. Department of Health and Human Services that provides comprehensive services including health, nutrition, and early childhood education to children and families who live below the poverty line. The program’s goals include (a) building and supporting stable family relationships, (b) enhancing children’s physical and social–emotional development, and (c) improving literacy, language, and problem-solving skills to strengthen cognitive development.

Head Start provides several advantages to the children and families it serves, including giving children the opportunity to attend preschool and helping them become better prepared for kindergarten. The program is designed to address the effects of poverty by providing substantial intervention to participating children and their families. There have been questions about the efficacy of Head Start and whether the modest gains achieved in Head Start are worth the investment in the program. This entry first discusses the history of Head Start and then describes its programs and policies and research on its effectiveness.

History of Head Start

Head Start started as a result of President Lyndon B. Johnson’s War on Poverty. The Office of Economic Opportunity launched an 8-week summer program called Project Head Start in 1965. Led by a pediatrician and psychologist, this comprehensive child development program helped communities across the nation meet the needs of disadvantaged preschool children, aged 3–5 years.

In 1966, Congress authorized Head Start as a year-round program. Head Start began in the federal Office of Economic Opportunity, which was later discontinued. Grant funding and oversight of Head Start programs are now conducted by the federal Administration for Children and Families in the Department of Health and Human Services.

In 1994, Early Head Start began in an effort to serve children from birth to age 3 in response to research evidence; this was an optimal time to intervene to impact children's long-term development. Since its creation, Head Start has served more than 34 million children, birth to age 5, and their families. Head Start programs were funded to serve nearly 1 million children and pregnant women during the fiscal year ending September 30, 2016. Federal spending for Head Start that year totaled approximately US$9.2 billion.

Programs and Policies

Head Start provides services each year to children and families across all 50 states and the District of Columbia, Puerto Rico, and other U.S. territories. Services in Head Start have expanded to include health screenings, health checkups, dental checkups, and developmental screenings. Educational curriculum for young children is decided by individual programs but must follow federal Head Start performance standards. Family advocates help families to access community resources such as education and employment. Services are designed to respect the family's culture and experience. In addition to Early Head Start and the main Head Start preschool program, there is specialized Head Start programming that targets (a) migrant and seasonal farm workers, (b) indigenous Americans in centers on or near reservations, and (c) homeless children and families.

Families are a centerpiece of the Head Start program. This emphasis on family partnership provides the rationale for the Head Start governing body called the policy council. Over half of the members of this group must be parents of currently enrolled children. The policy council is required to meet once a month at a time that is mutually convenient to all persons attending the meeting. The policy council approves budget, spending, and new hires.

Federal law required that by 2013 at least 50% of Head Start teachers needed to have a bachelor's or advanced degree in early childhood education or have at least a bachelor's degree in another subject along with coursework equivalent to a major relating to early childhood education with experience teaching preschool-age children. While education requirements for Head Start are now similar to those requirements for education professionals in school districts, Head Start programs are typically administered through social services agencies and not local school districts.

Effectiveness of Head Start

Numerous studies of Head Start have been conducted beginning soon after the program started, with inconclusive evidence of the program's effectiveness. California's Head Start programs have been studied several times with large data sets including 12,000 families and almost 50,000 children. Results of these studies indicate that Head Start has a positive impact on families and on the ability of preschoolers to meet age expectations once they get to kindergarten if they are enrolled in Head Start for 2 full years at ages 3 and 4. Other studies have found Head Start graduates are more likely to graduate from high school and attend college and are less likely to commit crimes.

In 1998, Congress mandated an effectiveness study of Head Start that involved 5,000 children of age 3 and 4 years. The study measured Head Start's effectiveness as compared to other forms of community support and educational intervention. Benefits to children improved with early participation and varied across ethnic and racial groups. A comprehensive review completed in 2005 stated that Head Start's long-term benefits are mixed but positive. Some studies have discussed a Head Start "fade," which means the initial positive impact of programming is no longer seen by second or third grade. However, other studies find decreases in grade retention and special education placement for preschoolers who participated in Head Start. In terms of the effectiveness of Head Start, it appears to produce some benefit including a few long-term benefits for those who participate as children.

Vera Lynne Stroup-Rentier

See also Applied Research; Childhood; Great Society Programs

Further Readings

Bierman, K. L., Nix, R. L., Domitrovich, C. E., Welsh, J. A., & Gest, S. D. (2015). The head start REDI project and school readiness. Health and education in early childhood: Predictors, interventions, and policies, 208–233.

DeLoatche, K. J., Bradley-Klug, K. L., Ogg, J., Kromrey, J. D., & Sundman-Wheat, A. N. (2015). Increasing parent involvement among Head Start families: A randomized control group study. Early Childhood Education Journal, 43(4), 271–279.

Haines, S. J., Summers, J. A., Turnbull, A. P., Turnbull, H. R., & Palmer, S. (2015). Fostering Habib's engagement and self-regulation: A case study of a child from a refugee family at home and preschool. Topics in Early Childhood Special Education, 35(1), 28–39.

Kline, P., & Walters, C. (2015). Evaluating public programs with close substitutes: The case of Head Start (No. w21658). National Bureau of Economic Research.

Stroup-Rentier, V. L., Summers, J. A., Palmer, S., & Turnbull, A. P. (2015a). An exploration of how the foundations intervention influences family-professional partnerships in Head Start: A case study. NHSA Dialog, 17(4).

Stroup-Rentier, V. L., Summers, J. A., Palmer, S., & Turnbull, A. P. (2015b). Family-professional partnerships in Head Start: Practical strategies using a partnership intervention. NHSA Dialog, 17(4).

U.S. Department of Health and Human Services (HHS), Administration for Children and Families (2017). Head Start program facts fiscal year 2016. Retrieved from https://eclkc.ohs.acf.hhs.gov/data-ongoing-monitoring/article/head-start-program-facts-fiscal-year-2016

Matthew B. Fuller Matthew B. Fuller Fuller, Matthew B.

Health Insurance Portability and Accountability Act Health insurance portability and accountability act

771

773

Health Insurance Portability and Accountability Act

The Health Insurance Portability and Accountability Act (HIPAA; Pub.L. 104–191, 110 Stat. 1936) was adopted by the U.S. Congress and signed into law by President Bill Clinton in 1996. It is also known as the Kennedy–Kassebaum Act or Kassebaum–Kennedy Act after two of its leading sponsors, senators Nancy Kassebaum (R-KS) and Edward Kennedy (D-MA).

The law has two main components, known as Title I and Title II. Title I protects health insurance coverage for workers and their families when they change or lose jobs. Title II requires the establishment of standards for health-care transactions and for protection of health-care information and electronic records. Although the law is designed to focus on health-care agencies, it defines *protected health information* (PHI) and *covered entities* in such a way that in addition to traditional medical providers, psychological services and postsecondary institutions of higher learning are considered covered entities.

Elementary and secondary schools are generally exempt from HIPAA protections. However, health information may be collected through educational records and, as such, would be covered by the Family Educational Rights and Privacy Act (FERPA). Although Title I offers protection of health care in times of employment transition for employees, Title II involves health information transaction and records of importance to most education agencies and is the focus of this entry. The entry provides an overview of Title II, then discusses the law's applicability to educational settings and educational research and its relationship to FERPA.

HIPAA establishes a Privacy Rule (45 CFR Part 160 and Part 164, Subparts A and E) that legislates national standards in protecting individuals’ personal health information. These privacy standards apply to health insurance providers, health-care billing agencies (called clearinghouses), and health-care providers conducting specific health-care transactions electronically. The Privacy Rule requires appropriate safeguards in protecting the privacy of individuals’ PHI and sets limits on the disclosure of PHI without patient consent. The Privacy Rule also provides patients with rights to examine and obtain copies of their records and to request corrections.

The Privacy Rule establishes definitions for types of organizations covered by HIPAA in section 1861(s), 42 U.S.C. 1395x(s). Covered entities are defined as (a) health plans, (b) health-care clearinghouses, and (c) health-care providers who electronically transmit any health information in the course of normal business. Generally, these transactions concern billing and payment for services or insurance coverage but they may also include sharing of diagnosis, results, test data, or orders for medical treatment. Any entity—a hospital, an academic medical center, a research facility, and a physician—who electronically transmits PHI directly or through another organization to a health plan would be considered a covered entity. Academic medical centers, research centers, psychological services, and other organizations are covered under this definition.

The Privacy Rule and other sections of Title II provide for a number of protections of what is considered PHI. PHI is defined in 45 CFR 160.103 as any individually identifiable health-care information that is collected or maintained electronically or in any other medium or format. PHI excludes any information collected as a matter of educational records covered under FERPA or employment records held by a covered entity in its role as employer. Therefore, the two definitions of *covered entity* and PHI govern whether HIPAA protections are in place for organizations offering health care, collecting or maintaining health information, and transmitting individually identifiable health information.

Basic patient rights that covered entities must adhere to when dealing with PHI include a number of measures related to disclosing or transmitting information between parties, de-identifying information, notifying patients of their rights, and responding to patient requests for corrections or to view records. For example, if a patient asks to view PHI or receive a copy of PHI, it must be provided to them within 30 days of the request.

The Privacy Rule also establishes the conditions in which patients must provide written consent for the disclosure of PHI between entities. A covered entity is permitted to disclose PHI without patient consent if the transmission is used to support the treatment or billing. However, all other forms of disclosure require written consent from the patient, and the covered entity must make a reasonable effort to disclose only the minimum necessary information required. The Privacy Rule gives individuals the right to request that inaccurate information in their health records be corrected. Patients must also be notified of disclosures of their PHI and have choices in how they may be notified such as via a specific phone number or address.

Applicability to Educational Settings and Educational Research

As previously mentioned, elementary and secondary schools have generally been viewed as exempt from HIPAA regulations on the basis that they often do not meet the definition of a covered entity or usually do not transmit health records electronically. Moreover, schools seldom bill for health services rendered, further reducing the likelihood that HIPAA regulations apply to their contexts. However, postsecondary institutions of higher education have been found to meet HIPAA’s definition of a covered entity and often do transmit health records electronically. Moreover, psychological service providers, counselors, and social service agencies may also be providing medical care and thus meet the definition of a covered entity under HIPAA.

Relationship to FERPA

When higher education institutions began implementing HIPAA regulations for medical centers on their campuses, a number of questions arose. Are psychological or counseling centers considered covered entities? If a professor learns of a student’s illness or other medical condition and shares it with others, has the professor violated the student’s HIPAA rights? In 2008, the U.S. Department of Education and U.S. Department of Health and Services issued a set of guidelines on the status of educational records in light of FERPA and HIPAA regulations that addressed these and other questions.

The guidelines established that for most situations arising in educational settings, FERPA regulations on educational records will guide the treatment of student health information. However, if health information is collected through hospitals, psychosocial counseling centers, or other more traditional health-care settings housed within an educational institution, HIPAA provisions should also be considered. The guidelines note that FERPA and HIPAA regulations offer a complementary, not contradictory, suite of regulations in the care and education of youth.

The HIPAA Privacy Rule establishes conditions under which PHI can be disclosed by covered entities for the purpose of research. The Privacy Rule (45 CFR 164.501) defines *research* as “a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge.” A covered entity is authorized to disclose health information that has been de-identified (in accordance with 45 CFR 164.502(d) and 164.514(a)-(c) of the rule) for the sake of research. Care should be taken to disclose only information necessary to complete the research.

Some have also considered whether HIPAA adversely affects participants’ engagement in educational or medical research. A 2005 study found that patient participation in survey research following a heart attack dropped by 62.4% and that patients who did participate in posttreatment outcomes research tended to be older, were more likely to be married, and had lower mortality rates 6 months after treatment following the implementation of HIPAA as compared to pre-HIPAA studies. This led to the conclusion that the HIPAA Privacy Rule adversely affected the number of patients in medical research and introduced severe selection bias in data collection for patient registries. J. Michael Oakes has argued, however, that HIPAA was unlikely to have a large effect on evaluation research.

Matthew B. Fuller

See also Confidentiality; Family Educational Rights and Privacy Act; Institutional Review Boards

Further Readings

Armstrong, D., Kline-Rogers, E., Jani, S. M., Goldman, E. B., Fang, J., Mukherjee, D., & Eagle, K. A. (2005). Potential impact of the HIPAA Privacy Rule on data collection in a registry of patients with acute coronary syndrome.

Archives of Internal Medicine, 165(10), 1125–1129.

Boruch, R. F., Michael D., & Joe, S. C. (1996). Fifty years of empirical research on privacy and confidentiality. In B. H. Stanley, J. E. Sieber, & G. B. Melton (Eds.), Research ethics: A psychological approach (pp. 129–73). Lincoln: University Press of Nebraska.

Koocher, G. P., Norcross, J. C., & Greene, B. A. (2013). Psychologists' desk reference (3rd ed.). Oxford, UK: Oxford University Press.

Nass, S. J., Levit, L. A., & Gostin, L. O. (2009). Effect of the HIPAA Privacy Rule on health research. Washington, DC: National Academies Press.

Oakes, J. M. (2002). Risks and wrongs in social science research: An evaluator's guide to the IRB. Evaluation Review, 26(5), 443–479.

U.S. Department of Education, … U.S. Department of Health and Human Services. (2008). Joint guidance on the application of the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) to student health records. Retrieved from http://www2.ed.gov/policy/gen/guid/fpco/doc/ferpa-hipaa-guidance.pdf

U.S. Department of Health and Human Services. (n.d.). Health information privacy. Retrieved from http://www.hhs.gov/hipaa/index.html

U.S. Department of Health and Human Services. (2013, July 26). Does the HIPAA Privacy Rule apply to an elementary or secondary school? Retrieved from https://www.hhs.gov/hipaa/for-professionals/faq/513/does-hipaa-apply-to-an-elementary-school/index.html

U.S. Department of Health and Human Services. (2013, July 26). Research. Retrieved from http://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

Nichola Shackleton Nichola Shackleton Shackleton, Nichola

Hierarchical Linear Modeling Hierarchical linear modeling

773

778

Hierarchical Linear Modeling

Hierarchical linear modeling is also known as using *multilevel models, variance component models,* or *random effect models*. These models are used when data have a hierarchical or clustered structure. Hierarchical structures are the norm in the social sciences; for example, patients are treated within hospitals, people live in households, employees work within companies, and children learn within the same classrooms. This structure introduces dependence into the data, as units observed within clusters are more similar than units chosen at random from the population.

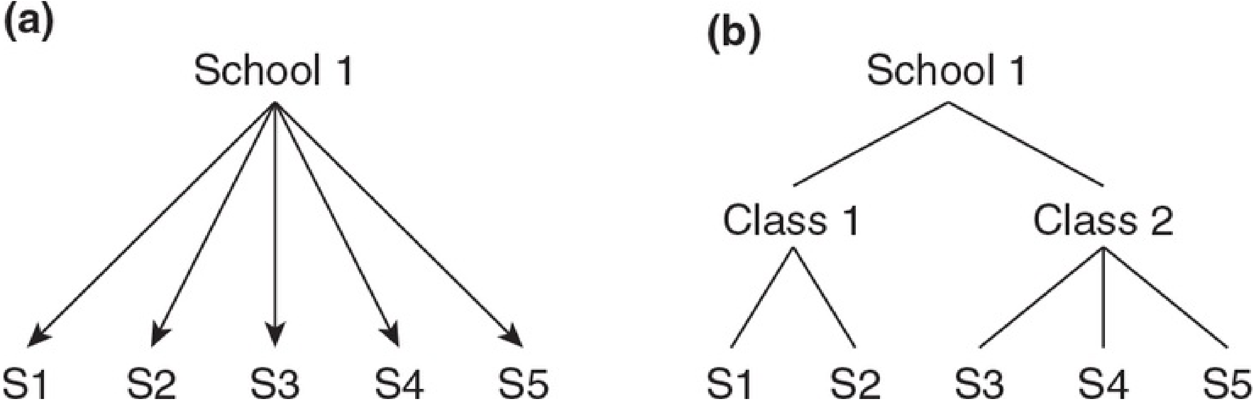
Traditional multiple regression techniques assume that observations are independent. Ignoring the clustered structure of the data leads to an underestimation of the standard errors of regression coefficients, leading to an overstatement of statistical significance. However, there are other methods for adjusting standard errors without fitting hierarchical linear models. Hierarchical linear models are most useful when the researcher is interested in group effects specifically. This entry discusses the basic principles and estimation procedures of hierarchical linear modeling, more advanced applications of these models, and the models' limitations.

Basic Principles and Estimation Procedures

Hierarchies

Hierarchical structures involve lower level units nesting within higher level units. Throughout this entry, the lowest level of observation in the hierarchy is referred to as Level 1, where units are nested within groups and these groups are referred to as Level 2; and when these groups are nested within higher order groups, the higher order groups are referred to as Level 3. Figure 1 demonstrates two hierarchies: a two-level hierarchy (a) with students (Level 1) nested within schools (Level 2) and a three-level hierarchy (b) with students (Level 1) nested within classrooms (Level 2) and within schools (Level 3).

Figure 1 Pictorial representation of (a) two-level hierarchy and (b) three-level hierarchy



Many kinds of data in the social sciences have a hierarchical structure, and it is worth noting that individuals are not always the Level 1 units. If schools were the unit of analysis, then schools (Level 1) could be nested within local authorities (Level 2). Equally, measurements taken at multiple time points (Level 1) may be nested within the individuals who were measured (Level 2). Research design can create data hierarchies through sampling. Clustered sampling techniques, or cluster randomized control trials, specifically recruit groups of people within hierarchies. Higher level units, such as schools, are selected for participation rather than randomly selecting individual students (Level 1).

Once hierarchies are established in the social world, they result in nonindependent (correlated) data. This can be the result of selection (where the characteristics of the individuals determine their groupings) and social processes (the interaction between individuals within groups, and exposure to the same context). For example, in some cases, schools attract students with similar characteristics (e.g., socioeconomic position, exam performance, and ethnicity) partially because of the schools' location and performance. Therefore, selection into schools results in students having similar characteristics and behavioral patterns than would be expected if selection into schools was random. In addition, socialization processes and interactions with other students and staff in the same school cause students to become more similar through the formation of the same school cause students to become more similar through the formation of perceived or actual social norms about expectations and behaviors. Therefore, even if allocation to higher level groups (in this example schools) was random at the outset, social processes and the exposure to the same environment create similarities in the group members. This creates dependence in the data when measuring the group members, which can be accounted for by using hierarchical linear models.

Statistical Validity

Hierarchical linear modeling is an extension to ordinary regression. Imagine a researcher has data from a survey of 600 pupils drawn from 50 schools. The researcher wanted to investigate the relationship between performance on a school leaving test at age 16 (y_i for pupil *i*) and a measure of ability on entry to high school (x_i). Single-level ordinary least squares regression would estimate a single equation by pooling all 600 cases expressing the achievement on the school-leaving test as a linear function of ability.

$$y_i = b_0 + b_1 x_i + e_i,$$

where b_0 is the intercept and *b* is the slope coefficient, and both are parameters to be estimated. The term e_i is the residual.

Equation 1 states that for a 1-unit increase in the measure of ability on entry to high school (x_i), the score on the school-leaving test (y_i) increases by the value of b_1. One assumption of single-level regression models is that the observations are independent, that is, the residuals (e_i) are uncorrelated. If school-leaving test scores (y_i) are clustered by school, and this is not taken into account in the analysis, the standard errors of the regression coefficients will generally be underestimated. This will result in confidence intervals being too narrow and an overstatement of statistical significance. If the clustering is not taken into account, then the information from all 600 students is treated as unique information. Correct standard errors will be estimated only if variation among groups is allowed for in the analysis.

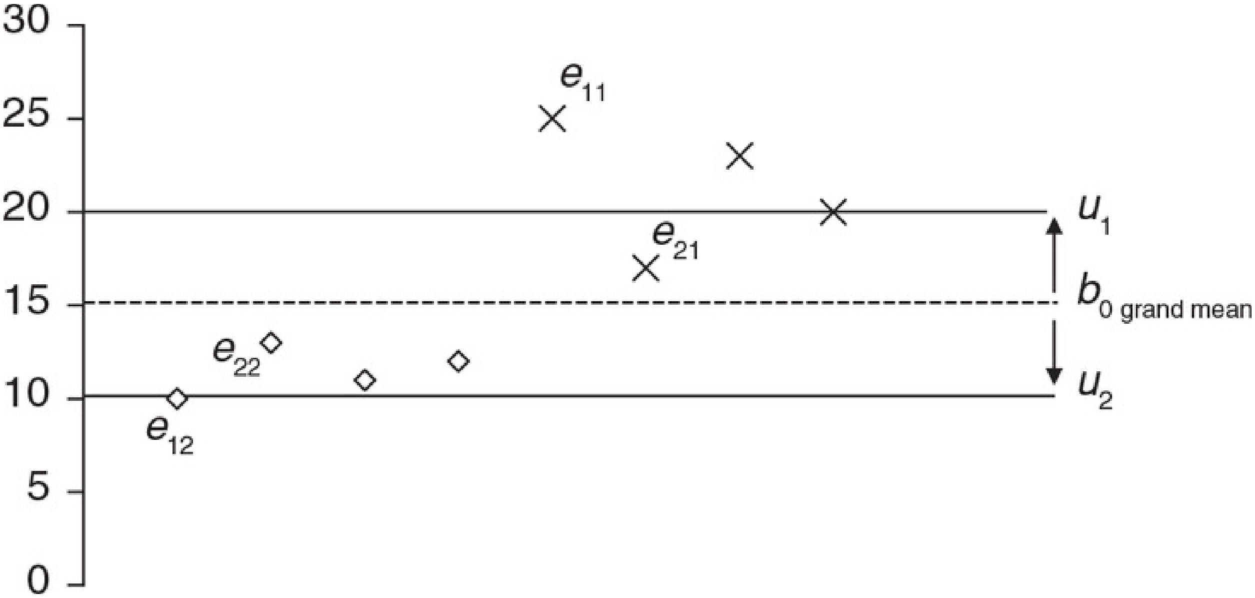
Hierarchical linear models form an appropriate generalization of Equation 1, to allow for group-level variation, by changing the suffixes, so that y_{ij} is the score on the school-leaving test for pupil *i* in school *j* and u_j is the “effect” for the *j*th school. The school effects u_j (Level-2 residuals) are random variables assumed to follow a normal distribution with a mean of 0. This model, the most common multilevel model, allows for different school-level intercepts. This model is also known as a *random intercepts model*.

$$y_{ij} = b_0 + b_1 x_{ij} + u_j + e_{ij}.$$

In Equation 2, b_0 is the overall mean of *y*; therefore, it represents the mean school-leaving test score across all students and schools (the grand mean). The mean of *y* for group *j* is $b_0 + u_j$, and so the group-level residual, u_j, is the difference between group *j*’s mean (cluster mean) and the overall mean (grand mean). This allows schools to have different mean scores on the school-leaving test. Some schools will have means that are higher than the grand mean, suggesting that their students perform better on the school-leaving test on average, and some schools will have lower cluster mean values. The group-level residual, u_j, is assumed to follow a normal distribution. The individual-level residual, e_{ij}, is the difference between the value of *y* for individual *i* and the individuals group mean $b_0 + u_{j.}$ This reflects differences in students’ individual school-leaving test scores from their schools cluster mean.

The random intercepts model is presented pictorially in Figure 2. In the example shown in Figure 2, the overall mean of *y* is 15. Student scores in Hypothetical School 1 are represented by the cross symbol and student scores in Hypothetical School 2 are represented by the diamond symbols. The school-level residual for School 1 is determined by the difference between the grand mean (15) and the mean for school 1(20) $u_1 = +5$.

Figure 2 Pictorial representation of a random intercept model



There are other ways to control for clustering and obtain correct estimates for the standard errors of regression coefficients. Survey methodologists adjust standard errors for *design effects*. They describe the adjustments made in terms of effective sample size; this reflects the reduction in unique pieces of information that each observation contributes within clusters. *Marginal models* (also known as population-averaged models) can also be used to analyze clustered data and obtain correct standard errors. The key advantage to hierarchical linear models is that they do not treat clustering as a nuisance; they specifically model the variability both within and between groups and the effects of group-level characteristic on individual outcomes.

Hierarchical linear models allow the use of covariates measured at any of the levels of a hierarchy. This enables the researcher to explore the extent to which differences in Level-1 outcomes (*y*) between groups are explained by both Level-1 and Level-2 covariates (and Level 3). Hierarchical linear models also allow for cross-level interactions, such that a covariate at Level 1 can be interacted with a Level-2 covariate. For example, we can examine differences in average student school-leaving test scores conditional upon characteristics of the students (e.g., sex) and the school (e.g., school practice), and whether different characteristics of the school are more or less beneficial for students with different characteristics (e.g., Sex × School Practice).

Harvey Goldstein has written extensively on the use of hierarchical linear models. Goldstein used hierarchical linear models to study the extent to which the variation between schools in graduation differed by prior attainment of the students. This is described as the value added by schools, as it provides an assessment of the distance traveled by a student within a school, accounting for the fact that students enter school performing at different levels and therefore showing how much schools are adding to an individual student’s performance. A student who achieves highly on entry to high school is likely to continue to achieve highly, but this metric also shows the gains made by students who were middle and low achieving on entry to high school. This changed the way schools were evaluated and how comparisons were made between schools in the United Kingdom.

Variance Partition Coefficient (VPC)

The VPC measures the proportion of total variance that is due to differences between groups. In simple hierarchical linear models, VPC is equivalent to the intraclass correlation coefficient, which is interpreted as the correlation between the outcomes of two randomly selected individuals from the same group. If all observations are independent of one another, ICC = 0. If all the responses from observations within all clusters are exactly the same, ICC = 1. If the ICC is 0.15, then the VPC interpretation states that 15% of the variation is between groups and 85% within. The ICC interpretation states that the correlation between randomly chosen pairs of individuals belonging to the same group is 0.15. Both interpretations are correct.

For continuous outcomes, the VPC/ICC is calculated as shown in Equation 3.

$$\mathrm{ICC} = \mathrm{var}\left(u_0\right) / [\mathrm{var}\left(u_0\right) + \mathrm{var}\left(e_0\right)],$$

where var(u_0) is the Level-2 residual variance, and var(e_0) is the variance of the Level-1 residuals.

There are different methods available for calculating the VPC/ICC for binary variables. A popular method that ensures the ICC estimates are not smaller than 0 and that within-cluster variance does not depend on cluster prevalence is shown in Equation 4.

$$\mathrm{ICC} = \mathrm{var}\left(u_0\right) / [\mathrm{var}\left(u_0\right) + \pi^2 / 3],$$

where var(u_0) is the Level-2 residual variance and $\pi^2/3$ (which is equal to 3.29) is by assumption the variance of the Level-1 residuals.

The discrepancy between the estimation of the VPC and the ICC occurs in three-level models when interpreting the variance at the second level. Imagine the three-level data structure shown in Figure 1b, with students nested within classrooms and classrooms nested within schools. The school-level ICC is calculated by:

$$\text{Level-3 ICC} \\ = \text{var}(v_0)/[\text{var}(v_0) + \text{var}(u_0) + \text{var}(e_0)],$$

where var(v_0) is the Level-3 (school) residual variance, var(u_0) is the Level-2 (classroom) residual variance, and var(e_0) is the variance of the Level-1 (student) residuals.

The classroom-level ICC is calculated as the correlation between two students within the same classroom, within the same school:

$$\text{Level-2 ICC} \\ = \text{var}(v_0) + \text{var}(u_0)/[\text{var}(v_0) + \text{var}(u_0) + \text{var}(e_0)],$$

where var(v_0) is the Level-3 (school) residual variance, var(u_0) is the Level-2 (classroom) residual variance, and var(e_0) is the variance of the Level-1 (student) residuals.

By contrast, the VPC for the classroom level does not include the school-level residual variance on the numerator:

$$\text{Level-2 VPC} \\ = \text{var}(u_0)/[\text{var}(v_0) + \text{var}(u_0) + \text{var}(e_0)]$$

Random Slopes Model

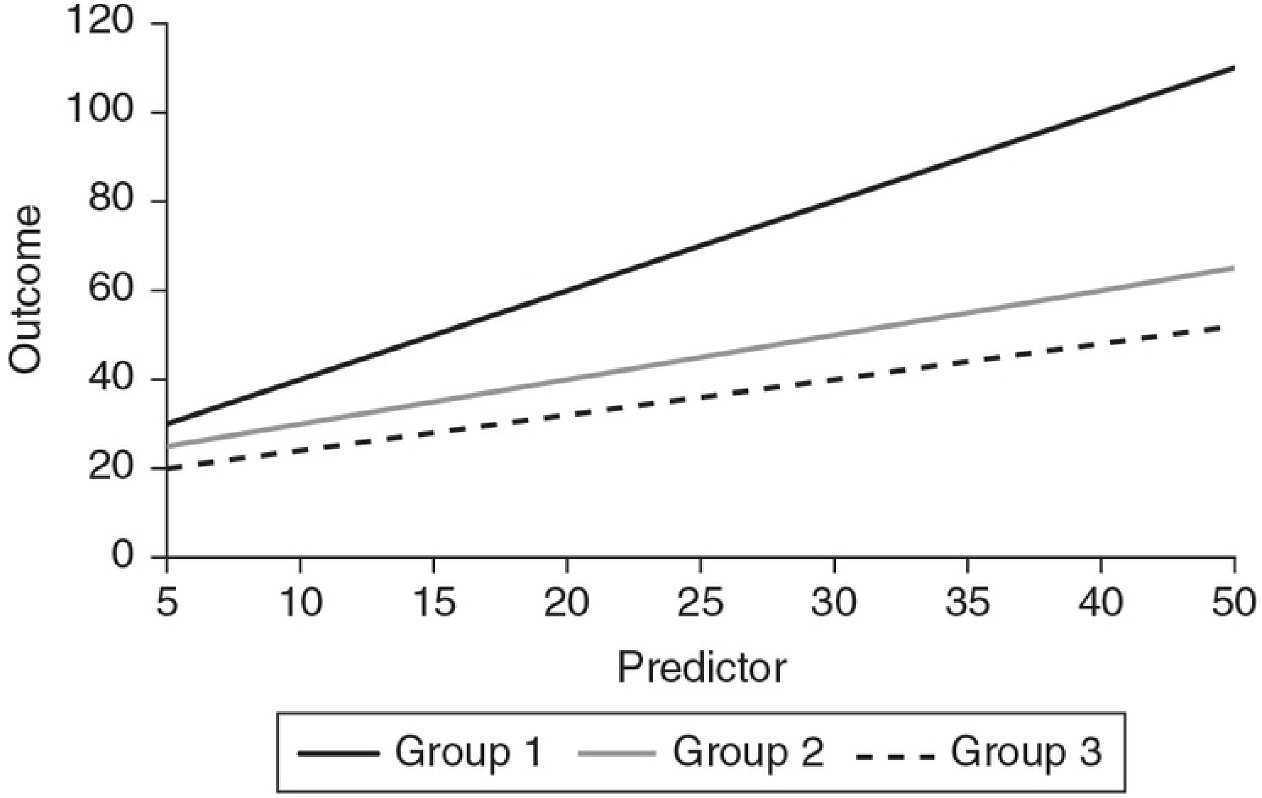
A random intercept model assumes that the relationship between the outcome and the predictors is the same for each group. This assumption can be relaxed by allowing for different slopes for each group. By interacting the group-level effect with the predictor variable, a differential slope is estimated for each group, as shown in Equation 8.

$$y_{ij} = b_0 + b_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij}.$$

Equation 8 is an extension of Equation 2, with the addition of the term $u_{1j}x_{ij}$. Subscript “0” differentiates the random effect for the intercept u_{0j} from the random effect for the slope u_{1j}. The intercept for group j is $b_0 + u_{0j}$. The slope of the line for group j is $b_1 + u_{ij}$. b_1 is the average slope across groups. Both random effects are now assumed to follow a normal distribution. The variance of the intercept and slope are assumed to be correlated; the covariance between the intercept and slope is estimated as part of the random slopes model. The random slopes model is also commonly known as the random coefficient model or a growth curve model when using repeated measures data.

Figure 3 presents a pictorial representation of a random slope model. Hypothetical values for the outcome (*y*-axis) and the predictor variables (*x*-axis) are shown. Groups 1, 2, and 3 have differential intercepts, the values for the intercepts are 30 for Group 1, 25 for Group 2, and 20 for Group 3. The slope for Group 1 is much steeper than the slopes for Groups 2 and 3. As the value of the predictor increases, the estimated value of the outcome increases at a greater rate for Group 1 than for either Group 2 or Group 3.

Figure 3 Pictorial representation of random intercept and random slopes



Growth curve models are random slope models fitted to repeated measures data. Repeated measures data are structured such that multiple measurement occasions (level 1) are nested within individuals (level 2). For example, a researcher might be interested in reading test scores of a set of children measured four times from age 5 to age 9. Growth curve models allow researchers to consider the average change in reading scores, to identify the variance in the intercept (how much do children differ in their initial reading scores), to identify what predicts differences in the intercepts (the initial reading score), to identify the variance in the slope (the change in scores over time) and to consider what predicts these differences. The growth curve model allows for change over time to be nonlinear through the use of higher order polynomials.

More Advanced Applications

The examples provided in this entry focus on hierarchical structures whereby lower level units (such as students) are clustered within a single higher level unit (such as schools). Extensions to hierarchical linear modeling can account for two additional types of hierarchies that exist: cross-classified structures and multiple memberships. Cross-classified structures appear when lower level units belong to combinations of higher level units that are not hierarchically ordered. For example, students may be nested within neighborhoods as well as schools, but students who attend the same school are not necessarily from the same neighborhoods. Multiple memberships occur when lower level units are nested within multiple higher level units. For example, older students do not tend to stay in one class formation with the same classmates but are in different class formations for different subjects. Extensions to hierarchical linear models are available for estimation with these data structures.

Hierarchical linear models are increasingly used as dynamic (autoregressive) models with repeated measures data. Dynamic models are used when previous responses are believed to exert a causal influence on subsequent responses, for example, how we expect prior achievement to exert a causal influence on current achievement. The most common application of these models involves including a lag of the dependent variable, where the dependent variable is regressed on a value of the dependent variable measured at a previous time point. Variables are included in the model to explain the variance in current achievement that is not accounted for by prior achievement, for example, changes in students' health status or families' economic resources.

Furthermore, hierarchical linear modeling has been incorporated into the structural equation modeling framework. This allows hierarchical linear models to be estimated using latent variables as outcomes and predictors.

Limitations

While this entry has highlighted the utility and flexibility of hierarchical linear models, the random intercept and random slopes models do not necessarily produce an estimate of causal effects; indeed, a strict set of assumptions is required for causality to be implied. These assumptions include a lack of correlation between the included covariates and the Level-1 residual, referred to as Level-1 exogeneity, and a lack of correlation between the included covariates and the random intercept, referred to as Level-2 exogeniety. The model also assumes that conditional upon the covariates in the model, the variance of the Level-1 residual is homoscedastic, the variance of the random intercept is homoscedastic, and that the random intercepts are uncorrelated between groups. The utility of the model is dependent upon the accuracy of the model specification.

Nichola Shackleton

See also Cluster Sampling; Multiple Linear Regression; Structural Equation Modeling

Further Readings

Fielding, A., & Goldstein, H. (2006). Cross-classified and multiple membership structures in multilevel models: An introduction and review (Research Report No. 791). Department of Education and Skills, University of Birmingham, England.

Goldstein, H. (1995). Hierarchical data modeling in the social sciences. Journal of Educational and Behavioral Statistics, 20, 201–204.

Goldstein, H., Huiqi, P., Rath, T., & Hill, N. (2000). The use of value-added information in judging school performance. London, UK: Institute of Education, University of London.

Leckie, G. (2013). Multiple membership multilevel models. LEMMA VLE Module 13, 1–61. Retrieved from http://www.bristol.ac.uk/cmm/learning/course.html

Leckie, G., & Goldstein, H. (2015). A multilevel modelling approach to measuring changing patterns of ethnic composition and segregation among London secondary schools, 2001–2010. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178, 405–424.

Steele, F. (2008). Introduction to multilevel modelling concepts. LEMMA VLE Module 5, 1–45. Retrieved from http://www.bristol.ac.uk/cmm/learning/course.html

Daniel B. Hajovsky Daniel B. Hajovsky Hajovsky, Daniel B.

Matthew R. Reynolds Matthew R. Reynolds Reynolds, Matthew R.

Hierarchical Regression Hierarchical regression

Hierarchical Regression

Hierarchical regression (HR) is one of several regression methods subsumed under multiple regression. HR is primarily focused on explaining how effects are manifested by examining variance accounted for in the dependent variable. The aim of HR is typically to determine whether an independent variable explains variance in a dependent variable beyond that already explained by some other independent variable(s).

It is typical that the additional amount of explained variance is evaluated for statistical significance based on change in R^2 (ΔR^2). R^2 represents the amount of variance in a dependent variable that is explained by an optimal linear combination of independent variables. Thus, ΔR^2 represents the change in variance explained in the dependent variable by including an additional independent variable.

For example, say a researcher is interested in studying influences on math achievement. Specifically, the researcher is interested in whether math self-concept explains variance in math achievement. But, the researcher knows, based on a literature review, that socioeconomic status (SES) and intelligence also explain variance in math achievement and math self-concept (and importantly, a possible relation between math self-concept and math achievement). Therefore, the researcher is interested in whether math self-concept explains variance in math achievement beyond that of SES and intelligence.

The researcher collects data on the three independent variables (SES, intelligence, and math self-concept), along with data on the dependent variable intelligence, and math self-concept), along with data on the dependent variable—math achievement scores. HR is used to analyze the data. The two so-called blocks of entry are used. In the first block, SES and intelligence are included, simultaneously, to explain variance in math achievement scores. In the second block, math self-concept is included to explain variance in math achievement, beyond that explained by the variables in the first block. Because SES and intelligence are already included in the regression, the math self-concept variable explains variance beyond that already explained by the combination of SES and intelligence.

In this example, if the ΔR^2 is statistically significant, then math self-concept explains unique variance in math achievement. To determine practical significance, the researcher may interpret the ΔR^2. A better estimate of the practical significance may be obtained by taking the square root of ΔR^2. This estimate is called the semipartial correlation. Using the aforementioned example, the semipartial correlation represents the relation between math self-concept and math achievement, after removing the influences of SES and intelligence from math self-concept.

The example we used seems to be the most common of HR. HR has several other uses including estimating total, direct and indirect effects, providing a standardized measure of effect size in the form of proportion of variance explained or semipartial correlation, and performing moderator analyses. Although ΔR^2 is often the focal point in HR, the regression coefficients can still be interpreted individually for their statistical significance, sign, and magnitude. One issue with interpreting the regression coefficients is how to interpret them in each block because they change when additional independent variables are included in subsequent blocks. Ultimately, the rationale for order of entry in HR should be based on theory and be logically defensible.

Daniel B. Hajovsky and Matthew R. Reynolds

See also Multicollinearity; Multiple Linear Regression; Partial Correlations; Residuals; Simple Linear Regression; Stepwise Regression

Further Readings

Jaccard, J., Guilamo-Ramos, V., Johansson, M., & Bouris, A. (2006). Multiple regression analyses in clinical child and adolescent psychology. Journal of Clinical Child and Adolescent Psychology, 35, 456–479.

Keith, T. Z. (2015). Multiple regression and beyond: An introduction to multiple regression and structural equation modeling (2nd ed.). New York, NY: Routledge.

Tonya Rutherford-Hemming Tonya Rutherford-Hemming Rutherford-Hemming, Tonya

High-Stakes Tests

High-stakes tests

779

780

High-Stakes Tests

High-stakes testing is an evaluation process whereby a major consequence is attached to a standardized test. “High stakes” refers to the outcome or consequence of the process, which for the student can be a grade or the potential to fail a course. While any test can be perceived by the test taker as high stakes if a grade is associated with it, high stakes here refers to standardized tests developed specifically to evaluate student achievement and school effectiveness.

Proponents of high-stakes testing believe that attaching significant rewards or major penalties to the evaluation method will motivate students and teachers to achieve better learning outcomes. The U.S. law known as the No Child Left Behind Act (NCLB) heralded a new age of increased high-stakes testing. When implemented in 2002, the NCLB enforced how and what educators would teach and how and what students would learn. It supported standards-based education reform built on the premise that setting high standards and establishing measurable goals could improve individual outcomes in education.

The NCLB heightened the stakes of standardized tests for schools and school districts because under the law, the test scores were publicly reported and schools that did not make adequate yearly progress for multiple years could face sanctions. The most severe sanctions under the law were replacement of the principal and school closure.

Many opponents of the NCLB contended that high-stakes testing had deleterious effects on students and teachers and that the consequences attached to poor performance failed to motivate students or improve teacher practices. In 2015, the NCLB was replaced by the Every Student Succeeds Act, which continues the annual testing requirement but gives states more discretion over their accountability systems. Still, high-stakes testing continues to be suggested as a means for educational reform.

Stressful effects of high-stakes testing are extensively reported in the literature. Stress is most often reported in terms of behavior (focused attention), cognition (outside worries/thoughts about the results), and physiology (increased heart rate). Most often reported is the notion that students with high test anxiety do not perform well when compared to students with low test anxiety. There has been relatively little research examining interventions for test anxiety.

Tonya Rutherford-Hemming

See also Adequate Yearly Progress; Every Student Succeeds Act; No Child Left Behind Act; Standardized Tests; Standards-Based Assessment

Further Readings

Banks, J. (2015). “Your whole life depends on it”: Academic stress and high-stakes testing in Ireland. Journal of Youth Studies, 18(5), 598–616.

Nichols, S. L., Glass, G. V, & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. Education Policy Analysis Archives, 20(20). Retrieved from http://epaa.asu.edu/ojs/article/view/1048

Segool, N. K., Carlson, J. S., Goforth, A. N., von Der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students’ anxious responses to high-stakes testing. Psychology in the Schools, 50(5), 489–499.

von Der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000–2010. doi:10.1002/pits.21660

HIPAA

See Health Insurance Portability and Accountability Act

Alon Friedman Alon Friedman Friedman, Alon

Histograms

Histograms

780

782

Histograms

A histogram is a bar chart in which data values are grouped together and put into different classes. These classes often present the frequency distributions found in the data set. Histogram coordinate systems are based on the horizontal axis and vertical axis. These two axes give the histogram the widths of the group that are equal to the class intervals and heights equal to the corresponding frequencies. Bars often represent the visual aspect of histograms; the height of each bar corresponds to its class frequency. As a result, the histogram makes the middle of the distribution visually apparent. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. This entry discusses the history of histograms and how they are used with statistics, data, and probability distributions.

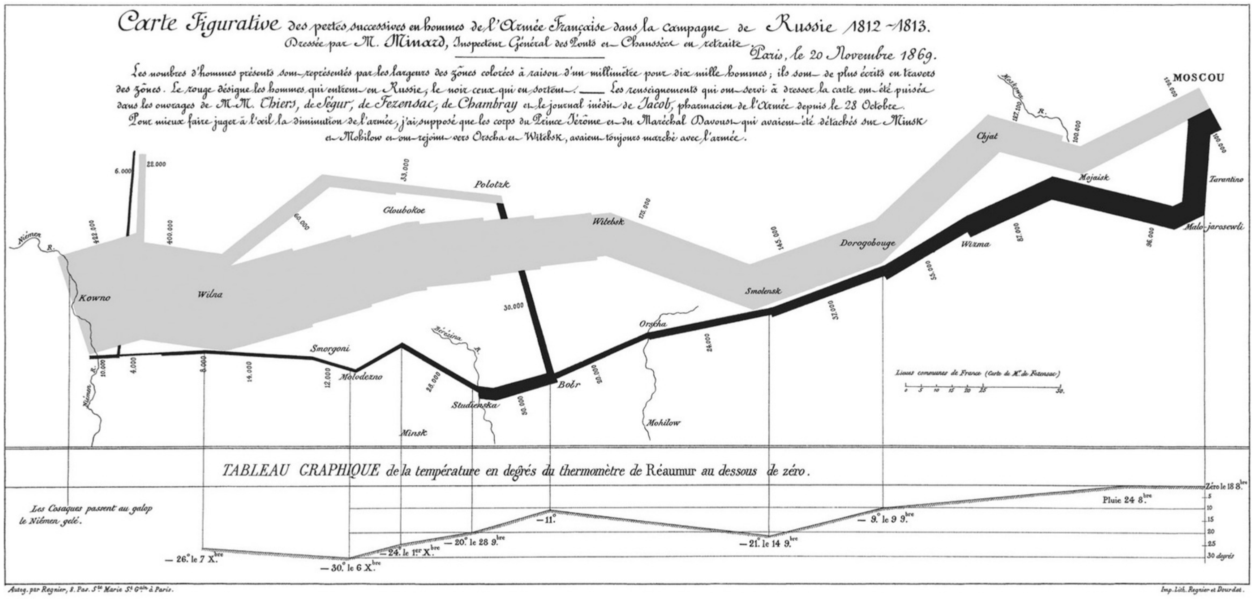
History

Histograms were introduced into the context of statistics as a columnar representation of frequency distributions arranged along the *x* axis. Karl Pearson defined histograms as an estimate of the probability distribution of a given variable by depicting the frequencies of observations occurring in certain ranges of values, also known as continuous variables.

The graphic display of a histogram is an important aspect of measuring the distribution. Although the graphic display of the histogram can show many visual patterns, many agree that a histogram should always display information succinctly. Charles Joseph Minard created an influential histogram showing the losses suffered by Napoleon’s army in the Russian campaign of 1812 (see Figure

1). This histogram is notable for the two-dimensional representation of six types of data: the number of Napoleon’s troops, distance, temperature, the latitude and longitude, direction of travel, and location relative to specific dates.

Figure 1 Charles Joseph Minard’s map of Napoleon’s Russian campaign of 1812


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Kowno
Wilna
Smorgoni
Molodezno
Gloubokoe
Polotzk
Witebsk
Studienska
Minsk
Bobr
Orscha
Mohilow
Smolensk
Dorogobouge
Wizma
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Niémen R.
Bérézina R.
422.000
400.000
6.000
22.000
60.000
33.000
175.000
145.000
100.000
30.000
10.000
4.000
8.000
14.000
12.000
28.000
20.000
24.000
37.000
50.000
55.000
96.000
Lieues communes de France (Carte de M.r de Fezensac)
0 5 10 15 20 25 50.
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Zéro le 18 8.bre
Pluie 24 8.bre
— 9.° le 9 9.bre
— 21.° le 14 9.bre
— 11.°
— 20.° le 28 9.bre
— 24.° le 1.er X.bre
— 30.° le 6 X.bre
— 26.° le 7 X.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.
Imp. Lith. Regnier et Dourdet.

Statistics

The benefit of histograms as a visual presentation of frequency distribution is summarized through five indicators that provide strong evidence for the proper distributional model:

1. The center, that is, the location of the data distribution
2. Spread, that is, the scale of the range of the data
3. Skew of data
4. Presence of outliers
5. Presence of multiple models extends scope of the data.

Under a frequency distribution, the data are arranged into numerically ordered class groupings. When developing a frequency distribution, each class grouping should have the same width. In order to determine the width of a class interval, the range of the data is divided by the number of desired class groupings.

Data

Data

In cases of large data sets, it is easier to present and handle the data by grouping the values into class intervals, which are sometimes known as bin widths. Sturges’s rule states that the data range should be split into k equally spaced classes where k = [1 + log10*n*] and where Log10(*N*) is the log base 10 of the number of observations. According to this rule, 1,000 observations would be graphed with 11 class intervals because 10 is the closest integer to Log2(1,000). The ceiling operator takes the closest integer above the calculated value. However, if the data are not normally distributed, additional classes may be required. The idea of skewness of the distribution is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative or even undefined. The formula is written as $k = [+ \log_2 n]$ where it is estimated the third moment of the skewness of the distribution, where it is derived from a binomial distribution and implicitly assumes an approximately normal distribution.

As a result, Sturges’s rule often leads to oversmoothed histograms, especially for large samples. If this is the case, the Rice rule, where one can choose the number of intervals by multiplying 2 to the cube root of the number of observations, can be used. The formula for the Rice rule is represented by *K* = [2*n* 1/3]. The call for the Rice rule as an alternative to Sturges’s rule appears when the moderate *n* is less than 200 observations. David P. Doane introduced his own modification to Sturges’s rule in 1976, aiming to improve its performance with nonnormal data. In Sturges’s original equation, the skewness of the distribution was the center of the distribution calculation.

Probability Distributions

Another aspect of the histogram is its ability to capture probability distributions. Under probability theory, the random variable is a function that describes the relative likelihood for selecting a random variable. The density of a continuous random variable, also known as the probability density function, focuses on the random variable that takes on a given value. The continuous random variable takes on an infinite number of possible values. The second type of variable is known as a discrete variable and can only take on a finite number of values. For a discrete random variable X that takes on a finite number of possible values, one can determine $p(X = x)$ for all of the possible values of X and call it the probability mass function. For continuous random variables, the probability that X takes on any particular value x is 0. That is, finding $p(X = x)$ for a continuous random variable X is not going to work. Instead, researchers will need to find the probability that X falls in some interval (a, b), that is, they will need to find $p\ (a < X < b)$. This will be completed by using the probability density function.

Alon Friedman

See also Bar Graphs; Data Visualization Methods; Quantitative Research Methods

Further Readings

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). Graphical methods for data analysis. Wadsworth, OH: Wadsworth International Group.

Doane, D. P. (1976). Aesthetic frequency classification. American Statistician, 30, 181–183.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 186, 343–414.

Sturges, H. (1926). The choice of a class-interval. Journal of American Statistical Association, 21, 65–66.

Patrick Radigan Patrick Radigan Radigan, Patrick

Sylvia L. Mendez Sylvia L. Mendez Mendez, Sylvia L.

Historical Research

Historical research

782

785

Historical Research

The word *history* can refer both to the sum of all human experience over time and to the study of these experiences. Therefore, history not only produces its own history, but it can provide a sense of one’s own identity and allow for the better understanding of the present human condition. It also corrects misleading lessons of the past, helps to develop acceptance and appreciation of other cultures or realities, and affords the opportunity to engage in deep critical thinking. A study of the past can provide modern researchers with many interesting lessons; undertaking a historical study requires researchers to think as historians, identify source materials, analyze them carefully, and employ refined methodologies that can lead to sound interpretations. This entry looks at how researchers approach history, the sources historians use, analyzing sources, methods of interpretation, and how interpretations change over time.

Thinking as a Historian

Even within the confines of a narrow topic of history, the amount of available evidence often is overwhelming. To manage this enormous task and maintain objective interpretations, researchers must learn to approach the past with increasingly mature stages of historical consciousness. The four progressive stages of consciousness are: 1. History as fact: New researchers of history tend to focus solely on concrete and particular events, situations, or outcomes. Studies conducted in such a fashion typically attempt to clarify or solidify discrepancies in the historical record.

2. History as a causal sequence: At this stage, individuals begin to explore cause and effect relationships between concrete events. Researchers typically search for a single, primary cause with which to explain the events of the past.

3. History as complexity: Researchers begin to appreciate the complexity of the human experience and the difficulty associated with understanding a given event due to the lack of complete evidence in the historical record. They also develop a willingness to accept multiple causality in favor of a single truth.

4. History as interpretation: At this final stage, researchers learn to appreciate the complexity, deficiency, and often contradictory evidence for any historical event. As mature students of history, they understand that they are responsible for selecting the evidence to be presented and that their own interpretations rarely provide an absolute truth to any given subject.

Identifying Sources

Sources are the artifacts for any historical study. Artifacts can be relics or testimonies. Relics are the physical artifacts of the past, such as drawings, paintings, architecture, or antiques. Testimonies are written or oral records that were purposively or inadvertently left for future generations and typically fall into the categories of primary or secondary sources.

Primary Sources

Primary sources, or direct sources, are original artifacts that were produced during the time in question and can be divided into two broad types: manuscripts and published sources. Manuscripts are handwritten, typed, or oral sources that have not been printed or intended for public consumption. They can include diaries, personal letters, transactional records, rosters, notes, memoranda, or any number of mundane scraps that have survived over time. Manuscript sources on any given topic can be difficult; however, many libraries, museums, and private collections contain a wealth of archived materials.

Published sources are intended for public consumption or professional study and typically fall into two broad groups: those intended for immediate publication and those originally not intended for publication. Materials for immediate publication include newspaper articles, public notices, legislative procedures, court rulings, autobiographies, and census reports to name but a few. Other court rulings, autobiographies, and census reports to name but a few. Other sources never intended for publication include private diaries and letters, typically published as part of a collection upon the death of the original author. Published primary sources are widely available, depending upon the time under study, and can be found in library collections, online databases, and as appendices or addenda to a number of secondary sources.

Secondary Sources

Secondary sources, or indirect sources, are publications produced from primary sources. They generally include books, essays, articles, and papers compiled by historians for subsequent investigators or public consumption. Such works can be scholarly or popular in nature and can include wide variations in scope and detail. In some cases, secondary sources can serve as primary sources when primary materials are no longer available. In most cases, as the study becomes broader, secondary sources are more heavily relied upon to incorporate and synthesize information. Secondary sources are ubiquitous and generally available on even the most esoteric of topics.

Analyzing a Source

As the artifacts for any historical study, sources must be carefully vetted prior to being employed in a study. All sources must be comprehensible, carefully located in place and time, and confirmed for authenticity. Verification of these characteristics of sources is vital, as misprints, misleading information, and forgeries are common in historical research.

Upon authentication of the source, the researcher can begin to evaluate the reliability of a document. This process can be an arduous endeavor that involves (a) tracing the genealogy or genesis of an artifact; (b) identifying its purpose; (c) placing the document into proper historical context; (d) determining the intent of the author; and (e) gauging the author’s authority, competence, and trustworthiness. The last item in this process typically is the most difficult to evaluate. Authors of both primary and secondary sources could be very selective in that which they chose to write and may have possessed prejudices or biases of which modern readers may not be aware. These authors may have had hidden agendas or ego-driven objectives that the researcher must explore to determine whether such factors play a role in the validity of the source. Seemingly objective sources such as census data, corporate financial records, or legal proceedings can contain deliberate inaccuracies or omissions that must be taken into consideration by the researcher before constructing a reliable set of sources to serve as the basis for a new historical interpretation.

Creating an Interpretation

Upon collecting and vetting a number of reliable sources, researchers must choose those they wish to emphasize, mention, or omit to construct a new historical narrative. A number of traditional and alternative methods exist to assist the researcher in this task.

Traditional Methods

Pioneered by the German historian Leopold von Ranke (1795–1886), traditional methods of historical research center upon primary sources, source criticism, and understanding historical context in an effort to define an accurate rendition of past events. The researcher must carefully compare many reliable sources to establish both consistencies and incongruities in the historical record. When sources are in agreement on an event, historians generally can consider it to be true. When disagreement occurs, care is needed to establish those sources that originated from a more authoritative author, eyewitness, or independent actor to determine more informative facts for the overall study. Researchers must carefully weigh each artifact, make prudent assumptions, and employ logical reasoning to create new historical interpretations.

Studies that employ traditional methods of historical research can be quite valuable, although researchers must proceed with caution when drawing conclusions based on the historical record. Many pitfalls are associated with these methods, including (a) generalizing the specific, (b) confusing correlation with causality, (c) using two unconnected facts to prove a third, and (d) adding unrelated evidence that has no established relevance to the issue under study. These errors are common in traditional historical interpretations, and many alternative methods have been developed to address them.

Alternative Methods

Although many historical studies continue to employ traditional methods related

Although many historical studies continue to employ traditional methods related to source criticism, an increasing number of modern historians have developed alternative methods to distill lessons from history. These methods are interdisciplinary in nature as they seek to combine the traditional approaches to the study of history with developments in other social science fields to produce a more robust understanding of the past.

Social History

Social history employs the major theories and methodologies of the social sciences (e.g., economics, political science, sociology, psychology) to describe historical events and trends. The theories of Karl Marx, Max Weber, Émile Durkheim, Sigmund Freud, and many others are routinely incorporated into social histories to explain the way in which individuals organize into groups, respond to influences, and change over time.

Beyond theories, the social sciences also have contributed a number of methodologies to the study of history. The social sciences primarily introduced quantitative research methods to historical research. Many historians now employ advanced statistical techniques to analyze historical records and to produce generalizable conclusions. The use of established qualitative research methods and techniques also has expanded. Some historians now conduct observational studies, wherein they focus on the actions rather than the writings of historical participants to address some of the difficulties associated with the vetting of primary sources. Although the social sciences have contributed many useful methodologies to advance the study of history, researchers should be careful to ensure that the historical record is robust enough to support an analysis using these techniques.

Cultural History

While social historians typically attempt to understand the past through the use of macrolevel theories drawn from advances in other disciplines, cultural historians attempt to reconstruct past events through the eyes of individual participants. Often referred to as “microhistory,” cultural histories employ methods largely developed by anthropologists, sociologists, literary critics, and critical theorists to develop thick, rich descriptions of the everyday lives of participants in historical events. These descriptions typically rely heavily upon qualitative research methodologies such as ethnography, phenomenology, semiotics, and linguistic analysis to illuminate the common experiences of a particular group under study. For example, oral history methods are employed to capture memories and insights that fill gaps in the written record. These methodologies allow researchers to draw conclusions from a potentially more accurate historical point of view, as the issue under study is approached from the eyes of its direct and indirect participants.

Cultural methodologies typically lack the generalizability that most social histories attempt to achieve, yet they retain strong ties to the traditional methods of historical research because they focus heavily upon source analysis and criticism. Studies that employ cultural methods of historical research typically depend heavily upon primary sources not intended for public consumption, such as records, letters, diaries, or oral histories passed from one generation to the next. As such, a lack of these sources can severely restrict the researcher's ability to construct a cultural history.

A Final Note on Historical Interpretation

With a wide range of methods available with which to interpret the past, it is important for researchers to remember that interpretations can and do change over time. Historians regularly ask new questions while offering different answers to old ones, and this process produces its own history of the study of history or historiography. All topics in history have their own historiography, and researchers can benefit greatly by understanding previous interpretations of their chosen topic. This aspect to understanding history often is vexing, as multiple interpretations of the same event can lead researchers to conclude that the study of history is a subjective exercise. However, it is important to note that events that occurred in the past are not fiction. While one's ability to fully understand previous events may be limited, a carefully constructed interpretation of the past can shed much light on the human experience.

Patrick Radigan and Sylvia L. Mendez

See also Critical Thinking; Objectivity; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Appleby, J., Hunt, L., & Jacob, M. (1995). Telling the truth about history. New

York, NY: W. W. Norton.

Brundage, A. (2013). Going to the sources: A guide to historical research and writing. West Sussex, UK: Wiley.

Davidson, J. W., & Lytle, M. (1998). After the fact: The art of historical detection (4th ed.). New York, NY: McGraw-Hill.

Hobsbawm, E. (1998). On history. New York, NY: The New Press.

Howell, M., & Prevenier, W. (2001). From reliable sources: An introduction to historical methods. Ithaca, NY: Cornell University Press.

Salevoris, M. J., & Furay, C. (2015). The methods and skills of history: A practical guide (4th ed.). West Sussex, UK: Wiley.

Southgate, B. (2005). What is history for? London, UK: Routledge.

Weinburg, S. (2001). Historical thinking and other unnatural acts. Philadelphia, PA: Temple University Press.

Yuk Fai Cheong Yuk Fai Cheong Cheong, Yuk Fai

HLM

HLM

785

788

HLM

Hierarchical linear and nonlinear modeling (HLM) is a specialized statistical software program for analyzing multilevel and longitudinal data. It is published by Stephen Raudenbush, Anthony Bryk, and Richard Congdon and distributed by the Scientific Software International, Inc. (Chicago, IL). The program's original versions came out in the early 1980s. The design of the program—its modeling modules and options, input specifications, and output—is in close coordination with the textbook written by Raudenbush and Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*. The following entry describes the analytical options, estimation approaches, inferential methods, and operational and output features of HLM.

Modeling Modules

HLM has eight modeling modules. They differ according to (a) the levels of hierarchy, (b) the type and the number of outcomes, and (c) the nature of the hierarchy.

The Levels of Hierarchy

HLM allows users to model data sets that have two to four levels of hierarchy. For each level, there is a submodel with its own structural and random components. The structural component represents, at that level, the relations among variables, and the random component denotes the residual variability. In a school effects study in which the sample consists of students clustered within schools, for example, there are two levels of nesting and subsequently two submodels. With $i = 1,\ldots,n_j$ students (Level-1 units) nested within $j = 1,\ldots, J$ schools (Level-2 units), the first sub-or Level-1 model for the achievement of student i in school j, Ach_{ij} can be represented as:

$$Ach_{ij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_{qj} X_{qij} + r_{ij},$$

where the structural component consists of the intercept, β_{0j}, and the Level-1 coefficients and predictors, β_{qj} and $X_{qij}(q = 1,\ldots,Q)$. The random component is denoted by r_{ij}, which is assumed to be normally distributed with a variance of σ^2. Some Level-1 predictor examples are family socioeconomic status (SES), gender, and prior achievement. A given β_{qj} relates the qth predictor to the achievement outcome. Letting the qth predictor be SES, the coefficient β_{qj} will thus index the strength and direction of the student SES-achievement association for school j.

The second sub-or Level-2 model predicts the intercept, β_{0j}, and the Level-1 coefficients β_{qj}. For the intercept and the qth Level-1 coefficient, the model can generally be represented as:

$$\beta_{0j} = \gamma_{00} + \sum_{s=1}^{S_0} \gamma_{0s} W_{sj} + u_{0j}$$

$$\beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} W_{sj} + u_{qj},$$

where the structural component consists of the intercepts, γ_{00} and γ_{q0}, and the Level-2 coefficients and predictors, γ_{0s}, γ_{qs}, and W_{sj} ($s = 1,\ldots,S_{0 \text{ or } q}$). The random component is denoted by u_{0j} and u_{qj}, which are assumed to be distributed as bivariate normal with dispersion **T**. The matrix **T** contains the variance and covariance components $\boldsymbol{\tau}_{00}$, $\boldsymbol{\tau}_{qq}$, and $\boldsymbol{\tau}_{q0}$. Some Level-2 predictor examples are school type (private vs. public), school SES, and curricular policies. The parameter γ_{0s} and a given γ_{qs} relate the *s*th school-level predictor to the Level-1 intercept and the *q*th Level-1 coefficient, respectively. Letting *s*th school-level predictor be school type, a researcher can use it to model the Level-1 intercept, β_{0j}, and assess how it relates to school achievement. The relationship will be captured by γ_{0s}. The same predictor can also be used to model a given β_{qj}, for example, the coefficient for SES, to see if the SES-achievement associations vary between the two types of schools or, equivalently, whether school type moderates the relationships between student SES and achievement. The estimate of γ_{qs} captures this moderating effect.

If repeated assessments are administered to these students in a given school over time, the students themselves become a clustering unit. A three-level hierarchy with repeated measures (Level 1) nested within students (Level 2) within school (Level 3) arises. With policy and administrative data collected on school districts in which the schools are nested, another hierarchy occurs and it results in an additional submodel with its own structural and random components at the district level (Level 4). The HLM2, HLM3, and HLM4 modules in HLM handle these models with different levels.

Type of Outcome at the Lowest Level

HLM can estimate models with continuous, binary, count, nominal, and ordinal outcomes at the lowest level, or Level 1, of the hierarchy via transformations of the outcomes with different link functions. The program uses the logit link function on binary outcomes such as school dropout, nominal outcomes such as the choice over different types of public schools, ordinal outcomes such as performance related to five categories of proficiency, and the log link function on count outcomes such as number of days absent from school. For continuous outcomes, the link could be considered as an identity one with no transformation performed. The HLM2 and HLM3 modules in HLM provide options for analyzing all five types of outcomes. The HLM4 module is capable of handling the first three types of responses only.

The Number of Outcomes at the Lowest Level

HLM allows users to model with single or multiple outcomes at the lowest level of the hierarchy. It allows users to jointly model multiple Level-1 response variables. For example, in a study on the academic growth of students with a fixed design with identical measurement occasions, users have the option to model the outcomes on the repeated measures simultaneously even if there are randomly missing data. The multivariate models offer the options for investigating and comparing the fit of models with different variance structures as well. For instance, users can study a model with a Level-1 first order autoregressive model covariance structure in which a Level-1 residual at time *t* is related only to its immediately preceding ($t - 1$) residual. With the students clustered within the same schools over time, an additional hierarchy arises, and it results in an additional submodel with its own structural and random components at the school level. The HMLM and HMLM2 modules in HLM estimate the parameters in these multivariate models.

The Nature of Hierarchy

HLM supports the analysis of data sets that do not have a strictly hierarchical data structure. Referring again to the previous example on the study of academic growth, the data structure ceases to be strictly hierarchical when certain students change their membership across schools over the course of study. The repeated assessments on achievement (Level-1 units) become nested within cells cross classified by two higher level factors, students and schools (Level-2 factors). Accommodating this data structure and modeling, student and school influences require the use of two-level cross-classified random effects models. HLM also handles cross-classified data sets with one of the two higher level factors clustered within an even higher level factor. For example, if the academic growth study collects achievement data on a set of students attending schools in their own districts over time, their repeated assessments on achievement (Level-1 units) become nested within cells cross classified by students and schools (Level-2 factors), and the schools are clustered within school districts (Level-3 factor). To investigate the influences of students, school, and districts, a three-level cross-classified random effects model is needed. Another different data structure for the longitudinal data occurs for a set of students who attend the same schools in the same school districts over time, but the schools they attend may or may not be in the their own districts. The repeated measures (Level-1 units) become nested within students (Level-2 units), and the Level-2 units are nested within cells crossed classified by schools and school districts (Level-3 factors). To analyze the data, the hierarchical linear model with cross-classified random effects can be applied. The HCM2, HCM3, and HLMHCM modules in HLM estimate these three different types of models, respectively.

HLM estimate these three different types of models, respectively.

Special Modeling Features

HLM has four additional special modeling features. First, HLM performs latent variable analyses in which coefficients at a lower level, such as β_{0j} and β_{qj} in Equation 1, are used as latent predictors. In a two-level longitudinal study of student achievement, for example, one could use the latent Level-1 coefficient tapping the initial academic achievement to predict the coefficient capturing the growth rate with SES as a covariate. Secondly, HLM performs automated analyses with multiply-imputed data, which are common in large-scale assessment programs to properly incorporate uncertainty brought about by imputation. Third, it has a V-known routine for researchers to perform research synthesis or meta-analysis in different scientific disciplines. Finally, HLM allows users to incorporate spatial dependence when analyzing geographical data.

Parameter Estimation

HLM computes three major types of estimates for model parameters. In a two-level model, as represented in Equations 1 and 2, the first type is the empirical Bayes estimates of the Level-1 intercept and the randomly varying Level-1 coefficients. A *q*th Level-1 coefficient, β_{qj}, is randomly varying when there is a random source in its variation. For example, the SES-achievement relationship previously discussed can randomly vary across schools. The empirical Bayes estimates are obtained based on information from the specific unit as well as data from other similar clusters. The second type consists of the generalized least squares estimates of the Level-2 coefficients, γ_{0s} and γ_{qs}, which allow for differences in the precision of the information each of the units offers. Maximum and approximate likelihood estimates of variance and covariance components, σ_2 and **T**, are also provided. To accommodate the commonly unbalanced nature of nested data due to varying cluster sizes and dissimilar observed patterns on the Level-1 predictors, iterative algorithms such as the EM algorithm and Fisher scoring are used to obtain the maximum likelihood estimates. For models with binary outcomes, HLM offers options to estimate the models using adaptive Gauss–Hermite quadrature and high-order Laplace approximations to maximum likelihood.

HLM provides users with single-and multiparameter hypothesis–testing procedures on these estimates. For instance, it provides likelihood ratio tests for model comparisons when the full maximum likelihood estimation method is used. With maximum likelihood estimation, the joint likelihood of the variance parameters and the coefficients at the highest level of the hierarchy is maximized. In addition, it allows users to assess the sensitivity of the results on the estimates of γ_{0s} and γ_{qs} against the violations of the distributional assumptions of the random effects at each level with generalized estimation equations sampling variability estimates.

Operational and Output Features

HLM requires users to prepare and input data files, and then it processes and prepares multivariate data matrix files for analyses. The program has three major modes of execution: (1) Windows, (2) interactive, and (3) batch. In the Windows mode, users rely on the graphic interface to create multivariate data matrix files and perform analyses. In the interactive mode, the users respond to prompts and choose options from menus provided. In the batch mode, the users submit the program at the DOS prompt. Users can also run models using a partly interactive and partly batch mode. In the Windows mode, model equations are expressed in a hierarchical format as represented in Equations 1 and 2. Users can also request the equations be expressed in a combined and mixed format. An example of a two-level model with one Level-1 and one Level-2 predictor can be represented in a mixed format as:

$$Ach_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{10}X_{ij} + \gamma_{11}X_{ij}W_{1j} + X_{ij}u_{1j} + u_{0j}$$

Although HLM does not handle data processing, it provides users with the option of listwise deletion of missing data at the lowest level during multivariate data matrix creation. It also provides users with the option to center variables using group or unit means or grand or overall means during model specifications to aid interpretations or to improve estimation of the effects of interests.

HLM offers various graphing options and output features for exploratory and diagnostic purposes. For example, HLM graphs scatterplots of Level-1 residuals by predicted values, which allows users to assess the constant error variance assumption and probe for outlying cases.

assumption and probe for outlying cases.

Yuk Fai Cheong

See also Generalized Linear Mixed Models; Hierarchical Linear Modeling; Longitudinal Data Analysis

Further Readings

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society. Series B (Methodological), 1–41.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2011). HLM7: Hierarchical linear and nonlinear modeling. Chicago, IL: Scientific Software International.

Robert L. Johnson Robert L. Johnson Johnson, Robert L.

Mason Lee Branham Mason Lee Branham Branham, Mason Lee

Holistic Scoring

Holistic scoring

788

790

Holistic Scoring

Holistic scoring provides an examinee with a single score regarding the quality of examinee work (i.e., performance) as a whole. Most commonly, holistic scoring is used to assess writing samples, though it may be employed to assess any performance task, for example, acting, debate, dance, or athletics. When scoring an essay holistically, the rater neither marks errors on the paper nor does the individual write constructive comments in the margins. Instead, the rater considers the quality of the entire paper and then assigns one holistic score. The SAT, ACT, and Advanced Placement tests all utilize a 6-point holistic scoring rubric to assess their respective writing sections. This entry discusses the use of holistic scoring guides and anchor essays and provides an example of a holistic scoring guide. It then discusses the differences between analytic scoring and holistic scoring.

Scoring Guides and Anchor Essays

In holistic scoring, well-organized essays with many grammatical errors and poorly organized essays with few mechanical errors may receive equivalent scores. This occurs because the rater must assess all strengths and weaknesses among various criteria before assigning a single holistic score. To prepare raters to score holistically, assessment practitioners will typically train raters using holistic scoring guides, also referred to as rubrics, and anchor essays. Anchor essays serve as examples of the performance levels at each score level of a rubric. For example, if the rubric uses a 6-point scale, raters will receive multiple examples (i.e., anchor essays) of a performance level of a six, multiple examples of a five, and so on. These anchor essays serve to assist raters in distinguishing the qualities of an essay worthy of a high rating or the qualities of an essay deserving of a low rating. Training raters to use the holistic scoring guide and anchor essays contributes to the reliability of their assessment.

Holistic Scoring Guide for a Narrative Writing Task

This section provides an example of a focused holistic scoring guide used to assess student writing fluency on a narrative task in the 1996 National Assessment of Educational Progress. The guide asked raters to focus on specific qualities of a paper—development, organization, sentence structure, mechanics, and overall ability—and then give one score across all criteria (U.S. Department of Education, 1999, pp. 45–46).

Scores

A six story demonstrates a high degree of competence in response to the prompt but may have a few minor errors. A story in this category generally has the following features:

- is well developed with a clear narrative structure,
- contains considerable detail that enriches the narrative,
- clearly demonstrates facility in the use of language, and
- is generally free from errors in mechanics, usage, and sentence structure.

A five story demonstrates clear competence in response to the prompt but may have minor errors. A story in this category generally has the following features:

- is developed with a clear narrative structure,
- contains details that contribute effectively to the narrative,
- demonstrates facility in the use of language, and
- contains few errors in mechanics, usage, and sentence structure.

A four story demonstrates competence in response to the prompt. A story in this category generally has the following features:

- is adequately developed but may have occasional weaknesses in narrative structure,

- contains details that contribute to the narrative,
- demonstrates adequate facility in the use of language, and
- may display some errors in mechanics, usage, or sentence structure but not a consistent pattern or accumulation of such errors.

A three story demonstrates some degree of competence in response to the prompt but is clearly flawed. A story in this category reveals one or more of the following weaknesses:

- is somewhat developed but lacks clear narrative structure,
- contains few details that contribute to the narrative,
- demonstrates inappropriate use of language, and
- reveals a pattern or accumulation of errors in mechanics, usage, or sentence structure.

A two story demonstrates only limited competence and is seriously flawed. A story in this category reveals one or more of the following weaknesses:

- lacks development and/or narrative structure,
- contains little or no relevant detail,
- displays serious or persistent errors in use of language, and
- displays serious errors in mechanics, usage, or sentence structure.

A one story demonstrates fundamental deficiencies in writing skills. A story in this category reveals one or more of the following weaknesses:

- is undeveloped,
- is incoherent, and
- contains serious and persistent writing errors.

Holistic Scoring Versus Analytic Scoring

In contrast to holistic scoring, analytic scoring provides distinct criteria for assessing examinees’ work. When scoring analytically, raters assign scores for each criterion. For example, in analytic scoring of writing, raters might assign separate scores for the narrative structure, sentence formation, and mechanics. This allows the rater to specify the performance level for each criterion, that is, on which criteria the examinee demonstrates strong performance (e.g., a score of 5 or 6) and which require improvement (e.g., a score of 1, 2, or 3). While holistic scoring results in assignment of 1 score, analytic scoring results in scores for scoring results in assignment of 1 score, analytic scoring results in scores for each criterion and a total score.

Although holistic scoring is generally easier and faster for the assessor, one major drawback is that it does not provide specific feedback to examinees about how to improve their performance. If providing formative feedback to examinees is important, it may be necessary to grade analytically. However, holistic scoring is useful if an assessor has thousands of essays or performances to grade or if an assessment center is grading within a tight schedule.

Robert L. Johnson and Mason Lee Branham

See also Analytic Scoring; Inter-Rater Reliability; Performance-Based Assessment; Reliability; Rubrics

Further Readings

Ballator, N., Farnum, M., & Kaplan, B. (1999, April). NAEP 1996 trends in writing: Fluency and writing conventions, NCES 1999–456. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). Assessing performance: Designing, scoring, and validating performance tasks. New York, NY: Guilford Press.

Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 387–431). Westport, CT: American Council on Education.

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single—Occasion, single genre, holistically scored writing assessments. In E. V. Steendam (Ed.), Measuring writing: Recent insights into theory, methodology and practices. Leiden, the Netherlands: Brill.

Yang Lydia Yang Yang Lydia Yang Yang, Yang Lydia

Holm’s Sequential Bonferroni Procedure Holm’s sequential bonferroni procedure

790

791

Holm’s Sequential Bonferroni Procedure

Holm’s sequential Bonferroni procedure is a statistical procedure used to correct familywise Type I error rate when multiple comparisons are made. A more robust version of the simple Bonferroni correction procedure, Holm’s sequential Bonferroni procedure is more likely to detect an effect if it exists. This entry describes the rationale of the Holm’s sequential Bonferroni procedure and the steps to conduct it.

In 1978, Sture Holm invented Holm’s sequential Bonferroni procedure, as an adjustment to the simple Bonferroni procedure. When a researcher achieves statistical significance on an overall test involving three or more groups, post hoc tests are used to determine which pairs are significantly different statistically. In post hoc tests, the critical levels for all paired comparisons have to be adjusted, so that the overall test doesn’t yield more significant differences than there actually are, which is known as an inflated familywise error rate. Holm’s sequential Bonferroni procedure is one type of post hoc tests. It follows the following steps: 1. Set the familywise significance critical level (α) for the overall analysis and count the total number of comparisons (m) in the analysis. For example, the familywise α is set at .05 in a study that compares three groups. There are a total of three paired comparisons, so $m = 3$.

2. List p values for all paired comparisons in an ascending order (from the smallest to the largest).

3. The first (the smallest) p value is compared with the critical value for the first paired comparison: α/m. In the aforementioned example, the first critical value is $.05/3 = .017$. If the smallest p value is equal to or greater than .017, then none of

the *p* values for paired comparisons are statistically significant. The procedure is stopped here. If the first *p* value is smaller than .017, the paired comparison is statistically significant, and the researcher examines the second *p* value.

4. The second *p* value is then compared with $\alpha/(m - 1)$, in this case, .05/2 = .025. If the second *p* value is equal to or greater than .025, then none of the remaining comparisons are statistically significant. The procedure is stopped. If the second *p* value is smaller than .025, the second paired comparison is statistically significant, and the researcher examines the third *p* value.

5. The third (largest in this example) *p* value is then compared with $\alpha/(m - 2)$, in this case, .05/1 = .05. If the third *p* value is smaller than .05, the paired comparison represented by that *p* value is statistically significant.

In general, the researcher repeats the same procedure until a comparison for each *p* value is equal to or greater than the critical value for a paired comparison, which indicates no further *p* values are statistically significant. The procedure is then stopped.

Yang Lydia Yang

See also Alpha Level; Bonferroni Procedure; *p* Value; Post Hoc Analysis; Type I Error

Further Readings

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65–70.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63, 655–660.

Beverly Pell Beverly Pell Pell, Beverly

Homeschooling

Homeschooling

791

793

Homeschooling

The term *homeschooling* refers to the practice of educating a child or youth at home rather than at a public or private school. Prior to the advent of compulsory education, homeschooling was the primary method of teaching a child. The National Home Education Research Institute estimates there are approximately 2.3 million homeschooled students in the United States, with a growth rate of 2%–8% per year from 2010 to 2015. Homeschoolers are a diverse population of various races and ethnicities and with differing religious and political views. This entry first looks at the reasons for homeschooling, its history and legality in the United States, and the methods used in homeschooling. It then discusses the academic performance of homeschoolers, their socialization, and the advantages and disadvantages of homeschooling.

There are two main ideas behind the practice of homeschooling; one contends that children learn better in a natural setting where they can control the depth and pace of their learning; the other emphasizes homeschooling as the best way to instill values and beliefs in children. Many homeschooling families share common beliefs about public schools: that they devote insufficient attention to academic instruction, that they reflect a decline in moral and religious values, and that the learning environment in them is increasingly erratic. In an era of accountability and school choice, homeschooling is perhaps the fastest growing form of alternative education.

History and Legality

The modern homeschooling movement started in the 1970s when American educator and school reformer John Holt began arguing that traditional education institutions focused too much on rote learning, rather than allowing children to learn naturally. Holt encouraged parents to free their children from formal education and instead school them at home. Holt used the word *unschooling* to describe child-directed learning where children are allowed to explore and learn on their own without criticism from adults. Holt's followers connected through his newsletter, *Growing Without Schooling,* founded in 1977.

In the 1980s, educational theorists Dorothy and Raymond Moore advocated delaying academics for children until they were developmentally ready. The Moores believed children should be schooled at home until age eight or nine in order to give them a firm educational, psychological, and moral foundation. The Moores' 1981 *Home Grown Kids* became the most popular book among new homeschoolers.

By the mid-1980s, thousands of evangelical Christians, concerned that public schools were teaching a narrow and secular worldview different than their own, started homeschooling their children. As more and more parents removed their children from public schools, relationships between homeschoolers and local schools grew contentious. The Home School Legal Defense Association was established in 1983, with a mission to legalize homeschooling and protect the constitutional right of parents to direct the education of their children. By 1993, homeschooling became legal in all 50 states.

Homeschooling laws vary from state to state. Eleven states require no notice of intent to homeschool. Fifteen states require parents to notify state or local education officials. Other states have a moderate to high degree of regulation of homeschooling; these states' requirements may include that parents produce test scores or submit a professional evaluation of student progress, that curriculum be approved by the state, or that state officials conduct home visits.

Motivations and Methods

According to the U.S. Department of Education's National Center for Education Statistics, the most cited reason for homeschooling is concern about the school environment, including factors such as safety, drug use, and negative peer pressure. Other reasons for homeschooling are to customize curriculum and learning for the individual child, to provide better academic instruction, to cultivate family relationships, and to teach and impart a particular set of values, beliefs, and worldview to children.

Teaching methods popular among homeschoolers include the Charlotte Mason method and the Montessori, Waldorf, classical, unschooling, and eclectic methods, with the eclectic method consisting of a mix of several methods. Many states hold yearly homeschool conferences where parents can attend workshops given by veteran homeschoolers and receive guidance and support from other homeschooling families. Homeschooling vendors offer vast selections of curriculum and instructional resources. Some parents teach using a purchased curriculum complete with textbooks, tests, and recordkeeping, whereas other parents choose to design their own curriculum.

Homeschooling parents may partner with university-style learning academies that cater to homeschoolers. Students can enroll in one or more courses they desire or take classes that parents may feel less competent to teach. Self-directed learning is common among homeschoolers as many students take online courses, work with virtual tutors, and participate in classes through distance learning programs. Many secondary-level homeschoolers enroll in community college courses to receive dual high school and college credit.

Academics and Socialization

Data collected by the National Home Education Research Institute show that homeschooled students consistently score 15 to 30 points above nonhomeschooled students on standardized academic achievement tests. In addition, homeschooled students typically score above average on the SAT and the ACT college entrance exams, and many colleges are increasingly recruiting homeschool graduates. There is no research to support the widespread belief that homeschooled students are less socialized than their nonhomeschooled peers. Homeschooled children are often active in their community and participate in several social activities, including homeschool co-ops, sports, dance and music lessons, church classes, field trips, community clubs, and other events outside the home. Numerous homeschooling families use the Internet for online social networking and connecting with students who share similar passions and interests.

Advantages and Disadvantages

The primary advantages to homeschooling include personalized learning, moral development, the absence of negative peer pressure, quality time spent with family, tutorial-style learning, and the ability to provide extra help to meet a child's needs. Disadvantages may include financial loss of income from the teaching parent, potential emotional hardship of being different, and taking on complete responsibility for a child's education. Opponents of homeschooling, such as the National Education Association, argue that homeschooled children may be placed at a higher risk for abuse, neglect, and other problems; many homeschooling parents have little to no accountability to the government; and homeschooling may impede funding allotted to public schools.

Thus far, homeschool research studies have sampled mainly conservative Christian populations, presenting a limited understanding of why parents homeschool. The decentralized nature of homeschooling makes it difficult to adequately measure the effectiveness of homeschooling. Overall, studies show homeschooled students perform well academically, and they are socially and emotionally well-developed. More research is needed to understand the impact homeschooling has on a democratic society.

Beverly Pell

See also Curriculum; Distance Learning; Montessori Schools; Self-Directed Learning; Waldorf Schools

Further Readings

Holt, J., & Farenga, P. (2003). Teach your own: The John Holt book of homeschooling. Cambridge, MA: Perseus.

Home School Legal Defense Association. (n.d.). Homeschooling research: Frequently asked questions. Retrieved from www.hslda.org/docs/faqs/

Homeschool.com. (n.d.). Homeschooling approaches. Retrieved from http://www.homeschool.com/Approaches/

Ray, B. D. (2016). Research facts on homeschooling. Retrieved from www.nheri.org/research/research-facts-onhomeschooling.html

Dean R. Gerstein Dean R. Gerstein Gerstein, Dean R.

Human Subjects Protections Human subjects protections

793

798

Human Subjects Protections

Since World War II, institutional structures and procedures for prior and ongoing review of research projects in compliance with professional ethical codes and government regulations have been developed to protect the interests of people engaged as living subjects of formal scientific studies. Studies of educational practices, including experiments, surveys, and analysis of school records, are among the types of research covered by these ethical regimes, although education studies in the United States have a somewhat special position under the regulatory Common Rule (CR). This entry discusses the reasons for the protection of human subjects in research and how these protections developed. It then discusses the CR, including its scope, institutional review board (IRB) processes, and criticisms of the application of the rule to social science research.

The first explicit ethical standards regarding human research subjects came in reaction to evidence of torturous medical experiments conducted in wartime Nazi prison camps. In subsequent decades, numerous other examples of inhumane and callous treatment of people inducted into scientific studies have been discovered and broadly publicized. These studies were often motivated by national security concerns and objectives.

A movement led principally by medical organizations has generated a worldwide institutionalization of research ethics committees, also known as IRBs or committees for human research protection (HRP), in universities, medical centers, public agencies, and other organizations employing medical and behavioral scientists or attracting their sustained interest. These entities are mandated to assure that researchers respect the dignity, autonomy, and best interests of research subjects and their communities through programs of systematic education, project inspection, and sanctioning authority. The pressure of changing technologies and research practices, and ongoing controversies in of changing technologies and research practices, and ongoing controversies in the area assure that the regulatory regime will continue to evolve.

History of Human Subjects Protections

In the aftermath of World War II, the U.S.-constituted Nuremberg Military Tribunal held criminal proceedings against nearly two dozen physicians and administrators in the so-called Doctors' trial, sentencing many of the defendants to imprisonment or death for conducting deadly experiments on inmates of concentration camps. Appalled by the extent of professional misconduct, the tribunal issued the 10-point Nuremberg Code, a code of ethics for protecting human research subjects, the main points of which were voluntary consent to inclusion in a study, advance provision of accurate and thorough information about the study procedures and potential effects, minimization of risk, proportionality of risks to benefits, and provisions for ending experiments. These points have been adopted in subsequent influential ethical statements, including the 1964 Declaration of Helsinki of the World Medical Association, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, authored by the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research and released in 1978, and the 2002 *International Ethical Guidelines of the Council for International Organizations of Medical Sciences*. The later reports elaborated special provisions for especially vulnerable or compromised populations, such as children, pregnant women, prisoners, and persons with mental disabilities, and extended the reach of such codes from biomedical interventions to withholding of interventions, behavioral interactions such as surveys, psychometric tests, use of confidential records, and research with human tissue specimens and genetic information.

The implementation of such broad ethical codes in institutional arrangements to monitor, encourage, and enforce compliance with them has been spurred by public revelations of innumerable disturbing experimental and observational studies starting in the 19th century and continuing through and beyond the 20th, most of them conducted under the auspices of democratic governments, often led by faculty at leading universities. Among the most influential exposés of these studies were an article by British anesthesiologist Henry Beecher in the *New England Journal of Medicine* in 1966, the reports of a U.S. Senate Select Committee led by Senator Frank Church in the mid-1970s, and Rebecca Skloot's *The Immortal Life of Henrietta Lacks*, published in 2010.

Among the most egregious examples of research that have been brought to light are the Tuskegee syphilis study, in which a large cohort of naturally infected poor Black men went untreated by the U.S. Public Health Service for decades after penicillin, a proven effective treatment, had been developed; the Willowbrook hepatitis study, in which scores of mentally disabled children were fed or injected with hepatitis virus; and Project MKULTRA, a decadelong program of torture, interrogation, and mental manipulation with drugs and stressors, secretly funded by the Central Intelligence Agency.

A U.S. Public Health Service report, *Surgeon General's Directives on Human Experimentation*, recommended in 1966 that all biomedical and biobehavioral studies supported by the Public Health Service (e.g., through National Institutes of Health grants) be vetted by institutional peer review groups, setting in place the basis for IRBs at universities and biomedical organizations in the United States. Over the next quarter century, under congressional and journalist pressure and a continuing series of high-level federal commissions, this basis coalesced into Title 45 Part 46 of the *Code of Federal Regulations* (aka 45CFR46), which includes the CR.

The CR

The CR was originally issued in 1974 only to cover activities of the Department of Health and Human Services but was then incorporated in 1981 into 15 other CFR titles (e.g., 34CFR97, 45CFR690), covering most federal departments and agencies. It now extends to 18 federal departments and agencies including the Department of Health and Human Services (with its National Institutes of Health and Food and Drug Administration), the National Science Foundation, the Department of Defense, and the Department of Education. The CR has played a huge role in promulgating and shaping the theory and practice of IRBs, research ethics committees, HRP programs, and consequently all human subjects research in the United States and most other countries.

A number of technical revisions to the CR have occurred since 1974. A more extensive set of revisions was published in January 2017 and was scheduled to take effect in January 2018. These revisions came after 6 years of rulemaking activity, which included extensive public commentary and response.

During the rulemaking process, the National Academies of Sciences, Engineering, and Medicine issued a report that detailed broad academic and industrial opposition to some of the proposed revisions and recommended outright derailment of the rulemaking process and appointment of a new independent national commission to give fresh consideration to the frameworks governing research involving human subjects. The CR issued in 2017 dropped some of the more controversial provisions published during the rulemaking period, streamlined part of the IRB review and monitoring process, contracted somewhat the purview of IRB domains, and promulgated new rules regarding "broad consent" for sharing of data banks containing genetic and other types of "big data."

Separate from but sharing common intellectual DNA with the CR, a more general influence over human subjects research protection has been exercised by the World Health Organization, which issued *Operational Guidelines for Ethics Committees That Review Biomedical Research* in 2002. In many countries outside of the United States, IRBs governed by national laws or regulations are organized as regional units affiliated with medical agencies or schools. The IRB movement or system is adaptive, reflecting national and local political and ethical culture, and dynamic, with changes over time driven in part by critics and controversies, as they have been in the United States.

Scope of the CR

The CR incorporates the tripartite ethical formulation of the Belmont Report: respect for persons, beneficence, and justice. Respect for persons focuses on self-determination or personal autonomy. This principle holds that persons deemed routinely capable of self-determination may enter research only voluntarily and after being adequately informed, but those with diminished capacity for self-determination due to diverse conditions such as immaturity, disability, illness, or imprisonment should be specially protected by bringing in third parties such as parents, guardians, or independent advocates as decision makers.

The principle of beneficence means ideally do *not* harm research subjects but practically, *minimize* harms while *maximizing* benefits, noting that benefits may at times be only to the greater good through enhanced social knowledge. The principle of justice refers to fair distribution of burdens and benefits of research, specifically, that the relative few who may be selected to carry the risks of harm not be different as a demographic class (incarcerated vs. free, poor vs. rich, foreign vs. domestic, Black vs. White) from the potential beneficiaries.

Information given in advance of research participation should include its offerors, purposes, procedures, risks, and benefits. Researchers must assure that participants understand the information. Questions may be asked and must be answered truthfully. Some kinds of information may be withheld at the outset to protect the validity of the research but must be disclosed afterward. Agreement must be made free of overt or subtle coercion or undue influence (excessive or improper rewards). If participants have reduced capacity for comprehension or vulnerability to pressure, both the participant and a protective third party must give informed consent.

The principle of beneficence, per the Belmont Report, “requires that we protect against risk of harm to subjects and also that we be concerned about the loss of the substantial benefits that might be gained from research” (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979, n.p.). IRBs must analyze the risks of harm as against the probabilities of benefits. Finally, justice requires that there be fair procedures and outcomes in the recruitment and selection of research subjects. Researchers cannot repair underlying social inequalities, but they should be mindful that the rich not benefit unduly from research with the poor, and “racial minorities, the economically disadvantaged, the very sick, and the institutionalized” may not be continually sought as research subjects “for administrative convenience, or because they are easy to manipulate” (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979, n.p.).

Addenda to the CR define additional protections for three subpopulations: children, prisoners, and the triad of pregnant women, fetuses, and newborns. Of special interest to education researchers is that behavioral studies of children in which the children are identifiable and the subject matter is sensitive, if not subject to the “normal educational practices” exemption discussed later in this entry, must under most circumstances be reviewed by an IRB under expedited or convened review, and both parental consent and children’s assent to participate in such research are usually required, with limited exceptions.

The CR requires that institutions controlled or funded in part by any of the federal agencies must setup IRBs to formally review every directly funded human research study in advance and in many cases to monitor subsequent study implementation. Following this lead, the same or similar rules and procedures have been applied by private research funders—commercial funders such as have been applied by private research funders – commercial funders such as pharmaceutical companies and not-for-profit foundations such as Spencer, Gates, and MacArthur. Moreover, the CR agenda has been applied in whole or in large part to all human subjects studies whether or not directly funded by government or other external sources.

The CR defines human subjects research by its methods, aims, and materials. The methodological criterion is that research involves the systematic collection or accumulation of information for scientific analysis, that is, for the purpose of developing or contributing to generalizable knowledge beyond the specific case or cases in the study. The research also must involve intervention, interaction, or communication by the researchers with living people or with identifiable biospecimens or records of their private behavior or attributes.

Questions have long arisen about behavioral studies such as ethnography, oral history, biography, community-based participatory research, secondary analysis of data sets collected from living populations, and Internet social media studies. In the 2017 update of the CR, oral histories, biographies, and journalistic studies that are meant to focus on specific individuals are explicitly not covered by the rule, but ethnographic work that involves collecting information from multiple persons and using a social science perspective to seek generalizable knowledge continues to qualify as human subjects research. Secondary analysis of data sets in which individual identities are indeterminate is not human subjects research, and the status of social media studies depends on whether the data being collected passively are "public," that is, whether those contributing to the sites have a reasonable expectation of privacy. If researchers participate actively in a website or try to manipulate activities on it, it constitutes human subjects research.

IRB Processes

The CR firmly guides and in many circumstances dictates IRB activities in detail, but most of the thousands of IRBs in the United States are local in scope and have somewhat individualized characteristics and procedures to adjudicate human subjects research conducted by the organization's employees or associates, including students. The CR and associated guidance covers IRB membership, procedures, decision rules, and record-keeping, and requires each IRB to register with the Department of Health and Human Services Office for Human Research Protections and affirm that it applies the CR to all federally funded research in the institution. Whether the same rules will be applied to all funded research in the institution. Whether the same rules will be applied to all other human subjects research conducted therein is left to the discretion of each institution. IRBs are broadly considered to be either biomedical or social–behavioral–educational in focus, depending on the main type of research covered and the corresponding expertise of the members.

The CR requires that federally funded researchers must apply for IRB review prior to beginning a study, describing the study design and its formal provisions for protecting human subjects, covering such issues as selecting the subjects; obtaining voluntary informed consent to participate; how researchers will treat, intervene, or interact with subjects in the research process; the risks of adverse physical effects, mental effects, discomfort, or social detriment; the potential benefits for the subjects, society, or scientific knowledge; conditions for early termination or withdrawal; confidentiality or anonymity; maintenance and sharing of research data; contact information for researcher and IRB; and, if applicable, sponsorship/funding of the study, temporary deception, material compensation, and alternatives to research participation.

IRBs or their designees may classify research as exempt from IRB review (this may be determined, per local IRB policy, by an IRB staffer or designated, suitably trained person other than the researcher), qualified for expedited (single member) review, or requiring discussion by a convened IRB (full board). There are a series of explicit grounds for exemption by type of study if the study involves minimal risk, where “risk” means the possibility of harm or discomfort, including criminal or civil liability or damage to a person’s financial standing, employability, or reputation; and “minimal” means not exceeding the level of risk encountered in everyday life, including routine physical or mental examinations or tests. The series of grounds for exemption includes normal education practices in a commonly accepted education setting; psychological tests, surveys, interviews, or observations of public behavior; public officials; deidentified existing data; evaluation of public services or benefits; taste, food quality, or consumer acceptance.

No study may be disapproved except by majority vote after discussion at a convened IRB meeting that satisfies quorum requirements. A study reviewed by a convened board may be approved for a maximum of 12 months and may be amended or renewed upon reapplication. A typical IRB reviews hundreds of new applications annually, plus renewals, amendments, final reports, and reports on adverse results or other problems. IRBs are required to report to the Office for Human Research Protections any serious unanticipated adverse effects of research or any pattern of noncompliance by an investigator under the IRB’s jurisdiction with the CR or the IRB’s policies and decisions.

Criticisms and Continuing Issues

Behavioral scientists have criticized the CR regulatory regime as tailored for experiments with drugs, toxic exposures, and medical devices and therefore not a good fit for behavioral research methods. The emphasis on written consent forms, often extending for many pages, has been said to protect institutions against litigation rather than assuring participant comprehension. The primary federal enforcement tactic of placing all of an institution’s human subjects research on hold is considered an unwieldy sanction.

Multiple IRBs reviewing virtually identical protocols have been observed mandating a wide variety of (slightly) different accommodations, wasting research time and money in reconciliation. But the consolidation of multisite collaborative protocols under a single IRB, although mandated in 2016 by the National Institutes of Health for its grantees, has been controversial on grounds that local circumstances differ, and centralizing IRB decisions will precipitate a “race to the bottom” as fee-seeking IRBs compete for the sole jurisdiction. The rise of detailed DNA mapping of very durable biospecimens and of social media platforms with massive information archives strained the original CR’s underlying assumptions about the discreteness of project review. The 2017 revisions attempted to deal with all of these issues, but the difficulties in gaining broad assent from key stakeholders such as major research universities make it difficult to foresee the extent to which criticisms may be allayed as the new rules are implemented.

Dean R. Gerstein

See also Belmont Report; 45 CFR Part 46; Human Subjects Research, Definition of; Institutional Review Boards

Further Readings

Beecher, H. K. (1966). Ethics and clinical research. New England Journal of Medicine, 274(24), 1354–1360. doi:10.1056/NEJM196606162742405

Council for International Organizations of Medical Sciences and World Health Organization. (2002). International ethical guidelines for biomedical research involving human subjects. Geneva, Switzerland: Council for International Organizations of Medical Sciences. Retrieved July 6, 2016, from http://www.cioms.ch/publications/layout_guide2002.pdf

Department of Homeland Security and Other Agencies. (2015, September 8). Federal policy for the protection of human subjects: Notice of proposed rulemaking. Federal Register, 80(173), 53933–54061. Retrieved May 29, 2016, from https://www.federalregister.gov/articles/2015/09/08/2015–21756/federal-policy-for-the-protection-of-human-subjects#h-41

Department of Homeland Security and Other Agencies. (2017, January 19). Federal policy for the protection of human subjects; Final rule. Federal Register, 82(12), 7149–7274. Retrieved March 2, 2017, from https://www.gpo.gov/fdsys/pkg/FR-2017–01–19/html/2017–01058.htm

Menikoff, J., Kaneshiro, J., & Pritchard, I. (2017, February 16). The Common Rule, updated. New England Journal of Medicine, 376, 613–615. doi:10.1056/NEJMp1700736

National Academies of Sciences, Engineering, and Medicine. (2016). Ethical, legal, and regulatory framework for human subjects research. In Optimizing the nation’s investment in academic research: A new regulatory framework for the 21st century (pp. 153–170). Washington, DC: National Academies Press.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Retrieved May 15, 2016, from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

National Research Council. (2014). Proposed revisions to the Common Rule for the protection of human subjects in the behavioral and social sciences. Washington, DC: National Academies Press.

National Science Foundation. (n.d.). Frequently asked questions and vignettes: Interpreting the Common Rule for the protection of human subjects for behavioral and social science research. Retrieved May 15, 2016, from http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp#third

Office for Human Research Protections. (2016). International compilation of human research standards. U.S. Department of Health and Human Services. Retrieved May 28, 2016, from http://www.hhs.gov/ohrp/sites/default/files/internationalcomp2016%20.pdf

Sparks, J. (2002). Timeline of laws related to the protection of human subjects. Retrieved June 6, 2016, from https://history.nih.gov/about/timelines_laws_human.html#19801983

Stewart, W. H. (1966, February). Surgeon general, public health service to the heads of the institutions conducting research with public health service grants, February 8, 1966 (Clinical research and investigation involving human beings) (ACHRE No. HHS-090794-A). Retrieved May 20, 2016, from http://history.nih.gov/research/downloads/Surgeongeneraldirective1966.pdf. Reprinted with addenda in Surgeon-General's Directives on Human Experimentation. (1967). American Psychologist, 22(5), 350–355, Retrieved from http://psycnet.apa.org/doi/10.1037/h0024885

World Health Organization. (2000). Operational guidelines for ethics committees that review biomedical research. Geneva, Switzerland: Author. Retrieved July 6, 2016, from http://www.who.int/tdr/publications/training-guideline-publications/operational-guidelines-ethics-biomedical-research/en/index.html

Dean R. Gerstein Dean R. Gerstein Gerstein, Dean R.

Human Subjects Research, Definition of

Human subjects research, definition of

Human Subjects Research, Definition of

Human subjects research is defined by its methods, aims, and materials. The core methodological notion is that research involves the systematic collection or accumulation of data, that is, numbers or texts subject to later analysis. The definition of human subjects research is more a practical matter than a philosophical, linguistic, or scientific matter because defining research as human subjects research means that it can be subject to special regulatory protections and review by institutional review boards. The need for a clear definition of human subjects research grows from a history of degrading, harmful experimental and observational studies that the lead researchers and their sponsors may have seen as justified by the pursuit of medical knowledge, national security, or insight into the human condition.

Although the nature of the regulatory and ethical regimes applied to human subjects research evolves over time, and technological change creates new puzzles and issues to be solved, a fairly broad and durable consensus emerged on the matter of definition during the final decades of the 20th century. These definitions are clearly stated in such foundational documents as the U.S. federal policy for protection of human subjects—the Common Rule.

Casual observation, pure theorizing, abstract model building, haphazard recording, or anecdotal reflection do not constitute human subjects research. In addition, the purpose of the activity must be to develop or contribute to generalizable knowledge, that is, the study should be meant and designed to find application or relevance beyond the specific case or cases in the study. Diagnosis for the purpose of determining an individual course of therapeutic or helping procedures or services and treatment or training for the purpose of enhancing individual health, skills, talents, or welfare do not constitute human subjects research.

Finally, the research must involve intervention, interaction, or communication by the researchers with living people or obtaining records of the private behavior or attributes of identifiable living people. Historical or postmortem studies of subjects beyond harm or care do not constitute human subjects research.

Questions arise about the status of several types of studies, including on the behavioral side such as ethnography, oral history, biography, community-based participatory research, secondary analysis of data sets collected from living populations, and Internet social media studies and on the biomedical side such as the use of stored biospecimens. In general, ethnographic studies and oral histories that involve collecting from multiple persons and using a social science perspective qualify as human subjects research.

Single oral histories and contemporary biographical studies are usually not considered human subjects research, but some institutional review boards take a more expansive view than others of what constitutes a “systematic” and “generalizable” research design. If data sets do not permit identification of respondents, then secondary analysis of data sets do not comprise human subjects research. Issues with studies involving social media studies that use website information revolve mainly around whether data being collected passively are “public” or those contributing to the sites have a reasonable expectation of privacy. If researchers participate actively in a website or try to manipulate activities on it, that constitutes human subjects research.

Dean R. Gerstein

See also Belmont Report; 45 CFR Part 46; Human Subjects Protections; Institutional Review Boards

Further Readings

National Science Foundation. (n.d.). Frequently asked questions and vignettes: Interpreting the Common Rule for the protection of human subjects for behavioral and social science research. Retrieved May 15, 2016, from http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp#third

Graham R. Gibbs Graham R. Gibbs Gibbs, Graham R.

HyperRESEARCH

HyperRESEARCH

799

802

HyperRESEARCH

HyperRESEARCH is a computer program that can assist researchers with the qualitative analysis of their data. It supports the use of a range of types of data, text, video, audio, and images, which can be coded. The coding frame or coding book can be developed and modified, and if required, given a hierarchical structure. Data at each code can be retrieved. A key function in the program is the ability to test and develop complex hypotheses about the data set by using the case-based structure of the project and its coding. This entry reviews the history of the software application, delves into the software's functionality with regard to cases and sources, coding, and filtering and hypothesis testing and concludes by highlighting additional functions and features of HyperRESEARCH.

History

Up until the 1990s, software that dealt with textual data, of which the earliest was General Inquirer, focused on quantitative analysis and provided tools like keyword in context and concordance generation to support a content analysis of the data. In the early 1990s, teams of academicians and programmers from several European countries, Australia, and North America began developing a new type of software for textual analysis that, following the networking activities undertaken by Nigel Fielding and Raymond Lee, came to be known as computer assisted qualitative data analysis (CAQDAS) programs. These programs drew their inspiration from writers such as Barney Glaser and Anselm Strauss, and Patricia Hentz Becker, who tried to formulate clear guidelines about how a qualitative analysis of data should be undertaken. A key to this was the idea of coding data, that is, assigning tags or code names to passages of text (or other segments of data). After developing a number of codes and applying them to the data by coding it, the researcher would retrieve all data coded to each code in turn and analyze it by identifying key ideas, patterns, and concepts in it. Many of the programs written in the early 1990s supported these code and retrieve functions. Soon, some started to develop further functionalities to help the researcher. Many of these focused on developing the idea of retrievals in more complex ways by, for example, linking together one set of coding with others in a variety of logical ways. These programs became known as theory builders.

HyperRESEARCH was one of the earliest programs of the theory builder type when it was developed in 1990 by Sharlene Hesse-Biber, T. Scott Kinder, and Paul Dupuis. In 1991, they incorporated a company, Researchware, to sell, develop, and market the program. Most software at the time was designed to run under MS-DOS on PCs, but some ran on a Macintosh and some (such as HyperRESEARCH and NUD.IST) ran on both a Macintosh and PC. Unlike most programs that soon supported only Windows versions, HyperRESEARCH has always been available in almost identical Windows and Macintosh versions. It is now developed using a programming language called Transcript (on the LiveCode software platform) that provides cross-platform development across Windows, Linux, Mac OS, Web, and various mobile operating systems, and it appears that it will maintain the program’s multiplatform support.

Cases and Sources

The distinctive feature of HyperRESEARCH is that data organization, coding, retrieval, and hypothesis testing analysis are based around a case-based structure. All these elements of a project are combined into a file called a study.

A range of different source files may be used and the computer files for these are stored separately from the study file. These include text files and Word docx files, images, videos, and audio recordings. Once coding has been started, these files should not be edited.

When planning a new study, the researcher must first define a set of cases because when coding is undertaken, the sources will be assigned to one or more of these cases. In social research, cases are commonly people and associated sources might be one or more interview transcripts, video of a focus group they were a member of, and their photograph and other relevant pictures. But cases may also be places (e.g., several different schools in an educational study) or events (e.g., a number of elections in a political study). This assignment of sources to cases is key in the form of analysis supported by HyperRESEARCH, as coding retrieval, filtering, and hypothesis testing all operate at the level of the case. In most CAQDAS programs, coding just means that a segment or several segments of the source are labeled with a tag, whereas in HyperRESEARCH coding also means that the whole case is labeled with that tag. In this sense, the type of coding is closer to the quantitative categorization of cases using values like male, female, aged 20–25 years, and so on. The code categorizes the case; however, it is always possible to inspect the source segment or segments on the basis of which the case has been categorized. This is a distinctive feature of HyperRESEARCH compared with other CAQDAS programs. The study window lists the coding associated with each case rather than sources, though the listing does show to which source the coding has been applied. A source document is only opened within the program; it is not imported. Therefore, the contents of a source are not assigned to a case until a segment of it has been coded. This is achieved by ensuring that the relevant case is showing in the study window when the source is being coded.

Coding

The codes being used in the study are shown in the code book window. This is an ordered listing of all codes being used in the study, and it can be created by importing a file defining *a priori* codes (perhaps based on existing theory or the commissioning body’s questions) or codes can be created, inductively, one at a time, as the source contents are inspected and coded. Initially, codes are in a simple list but they can be arranged into a hierarchical structure of code groups and subgroups, though the group names are not themselves codes and cannot be used in coding. The process of coding means applying one or more of the codes to a segment of the source. If this is a text file, this can be anything from one character to a sentence, a paragraph, or even the whole file. In the case of video and audio sources, what is coded may be any period in the time line, and for images, it may be any rectangular selection. Coding is done by selecting the segment of the source with the mouse and then either clicking on the appropriate code (or codes) in the code book window and clicking on the Apply Codes button, by double clicking the code name, or by dragging the code name to the selected segment of a textual source. As this is done, each code segment is listed for that case in the study window showing the code used and the source file. Clicking on one of these in the list immediately opens the relevant source file window showing the whole of that source with the coded text highlighted, the image coded rectangle showing, or it plays the coded segment of the video or audio file. Displayed to the left of the text source contents are the codes used on that source with angled shading showing the lines of the source that have been coded with each code. This means that the coding undertaken can easily be inspected in its context.

There is a facility to autocode textual documents based on a lexical word or phrase search facility. This means the researcher can specify a set of words and/or phrases, and the program will find all matches and automatically code them and a stipulated amount of text surrounding the match to a specified code.

The development of the coding and the refinement of the coding book are supported by a range of functions that enable codes to be duplicated and renamed (so that segments coded to different codes can be combined under one code), and coded segments may be recoded (under the name of a different code) or uncoded (by deleting its coding).

Filtering and Hypothesis Testing

Cases may be filtered in a variety of ways: by name, by the sources they use, or by criteria. The criteria used in filtering match those used in hypothesis testing and include functions such as “proceeded by,” “overlaps,” and “includes” that specify the relations between coded segments (e.g., cases where segments coded as A overlap with segments coded as B) or logical (Boolean) relationships between codes such as AND, OR, and NOT (e.g., cases coded as A and coded as B). Brackets may be used if specifying a very complex set of criteria. Filtering in this way can be used in an exploratory way to identify which cases match certain criteria as a basis for extending or narrowing the codes and their coding. It may also be used to try out the various stages of a complex hypothesis. The theory builder enables the construction and development of hypotheses about the cases based on combinations of codes specified by functions and Boolean relations. In the theory builder, cases that match the specified criteria can be temporarily labeled, and these labels can then be used in further stages of hypothesis testing. This approach relies very heavily on the appropriate coding of cases, and inevitably hypothesis testing in the theory builder commonly involves the researcher going back to the data set to modify or add to the codes and what they code. This approach is quite different from the query and code searching functions found in most other CAQDAS programs. Using these tools typically returns the coded segments that match the specified criteria. In contrast, in HyperRESEARCH, filtering and the theory builder produce a set of cases that match the specified criteria.

HyperRESEARCH is therefore not a standard code and retrieve program, although retrievals can be done. By filtering to one code and producing a report on all cases, it is possible to retrieve all segments that have been coded the same way. However, in HyperRESEARCH, when a code is applied to some source material, it also categorizes the case to which that source belongs. Once coding is done following this philosophy, then the powerful tools of filtering and the theory builder can be used to examine and test hypotheses about the data set.

Additional Functions

The program now comes with a range of additional functions. These include a report builder that can gather together all coded segments associated with a case or set of cases, a masking tool (found within the report builder) that can be used to anonymize the reports, a frequency report that shows how often codes are used in cases or selected cases, a Mixed Methods Importer that allows the importing of quantitative variable case data that are automatically allocated to the respective cases in the study and a word counter that produces a concordance of all words used in selected sources and their frequency of use (this can also produce a word cloud image). There is also a graphical code map tool that can also be used to filter codes.

As all source files used in a study are kept external to the study file itself, there is also a study packager tool to bring together all these files so they may be easily moved to another PC or copied as a backup. HyperRESEARCH supports researchers working as a team by making it possible to merge studies by importing one into another to produce a master study file that can be distributed back to the team.

Version 3 was released in 2010 and subsequent updates have added in a range of additional functions. These include a command line interface for installing that can be used by IT system managers to install the software in labs, the ability to install the program on a memory stick so that it can be moved from PC to PC, and so on.

There is a full range of materials to support learning how to use the programs, which include a manual, tutorials, videos, and training webinars. There is a trial version that is not time limited but will only handle up to 75 codes, seven cases, and 50 code references per case.

Graham R. Gibbs

See also Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Gibbs, G. R. (2013). Using CAQDAS programs. In U. Flick (Ed.), SAGE handbook of qualitative data analysis. London, UK: Sage.

Hesse-Biber, S. N., & Dupuis, P. (2000). Testing hypotheses on qualitative data: The use of HyperRESEARCH computer-assisted software. Social Science Computer Review, 18(3), 320–328.

Silver, C., & Lewins, A. (2014). Using software in qualitative research: A step-by-step guide (2nd ed.). London, UK: Sage.

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

Hypothesis Testing

Hypothesis testing

802

804

Hypothesis Testing

Inferential statistics are used to make decisions about the population based on sample information. In order to reach statistical decisions, broad statements or hypotheses are formed about the probability distribution of the population. Hypothesis testing is the formal statistical process used to evaluate the probability or likelihood a hypothesis is true. The two major procedures for hypothesis testing are the classical approach and the probability value or *p* value method.

In classical or traditional hypothesis testing, four major steps are employed: (1) statement of the null and alternative hypotheses (H_0 and H_A); (2) determination of an analysis plan; (3) calculation of the test statistic; and (4) evaluation of the hypothesis statements. In the first step, the H_0 and H_A identify the parameter(s) being tested (e.g., mean, proportion, total) in order to determine whether the sample is coming from the same or different population. The H_0 assumes any observed difference between the sample and population is due to chance (i.e., sampling error) and not influenced by a predictor variable. Conversely, the alternative or research hypothesis (H_A) states a change or difference exists between the sample and population that is due to a nonrandom cause. The two hypotheses are written as mutually exclusive and exhaustive such that if one statement is accepted then the other statement must be rejected.

Hypothesis statements are formatted as either directional or nondirectional (equal to or not equal to a parameter value). The correct format depends on whether the researcher is interested in testing for a range of values or an exact value. For example, a professor believes that students who sleep more than 6 hours per night have test scores higher than the overall average ($\mu = 75$). In a directional format, the H_0 is that the test scores for these students are equal to or less than the expected average ($H_0 \leq 75$), and the alternative is then what the professor suspects, higher than average test scores for this group ($H_A > 75$). If the professor only speculates that the test scores for this group are different than the average (rather than higher or lower), then a nondirectional format is used (H_0: $\mu = 75$ and H_A: $\mu \neq 75$). This example is set up for a one-sample test using a single-point parameter, the population mean (μ), for comparison. Other statistical tests have hypothesis statements that are formatted slightly differently based on the number of samples, relationships, and parameters examined.

The second major step in hypothesis testing is deciding the appropriate sample statistic from which the decision to reject or not reject the H_0. The proper statistical test is determined by the research question being asked and whether the data meet the test requirements. Parametric statistics assume a random sample drawn from a normally distributed population with variables measured at the interval or ratio scale; nonparametric statistics are for data sets that do not meet these requirements. The sample size also impacts the calculation and interpretation of most statistical tests. Small samples ($n < 30$) utilize the Student *t* distribution, whereas large samples ($n \geq 30$) employ the *z* (also known as Gaussian or normal) distribution.

After calculating the appropriate test statistics (the third major step), the final step is to evaluate the hypothesis statements. Thus, either the H_0 is retained or the alternate hypothesis is supported. In the example noted with the directional format, the H_0 would be retained if the students who slept more than 6 hours a night have average test scores equal to or less than the expected average. However, if the appropriate test statistic showed statistically significantly higher than average scores for this group, as the professor suspected, the alternate hypothesis would be accepted.

The determination of statistical significance and decision to reject or not reject the H_0 is dictated by the accepted level of sampling error. The two primary sources of error are: a Type I (α) error, when the H_0 is rejected as false when in fact true, and a Type II (β) error, when the H_0 is accepted as true when in fact false. Type I errors are considered the more serious of the sampling errors and statistical significance is linked to low alpha (α) values, such as .05 or .01. In classical hypothesis testing, the rejection of the H_0 in favor of the alternative at the .05 significance level means only a 5% chance or less exists of committing a Type I error.

Classical hypothesis testing has two major weaknesses. First, the significance level (α) is arbitrarily chosen using conventional probabilities, often without a theoretical basis or statistical rationale. Second, the final decision regarding the H_0 and H_A is dichotomous. In other words, the conclusion informs only on whether to reject or accept the H_0 based on the specified significance level and not the exact confidence in that decision. The *p* value approach addresses these shortcomings and is the preferred method in modern empirical research.

In the *p* value method, the hypothesis testing procedure construct differs from the classical approach in one key aspect: the a priori significance level selection is eliminated. Instead, the exact probability of a Type I error (or the exact α value) is reported along with the evaluation decision and not just whether the significance level threshold is met. The *p* value measures the relative strength in the decision to reject or not reject the H_0. As *p* values approach 1, confidence in not rejecting the null in favor of the alternative becomes greater; the sample indeed comes from the same population. Conversely, *p* values closer to 0 indicate less confidence in the H_0 and the alternative should be accepted as true.

Then, hypothesis testing is a form of statistical inference whereby a population parameter is presumed or hypothesized and sample data are used to assess the hypothesis plausibility. The four major steps in the classical and *p* value methods outline the process for evaluating hypotheses and establishing their statistical significance or confidence in the evaluation decision. However, statistical significance does not always equate with practical or scientific meaning, and caution must be exercised in the hypothesis construction and evaluation.

Jill S. M. Coleman

See also *p* Value; *t* Tests; Type I Error; Type II Error; *Z* Scores

Further Readings

Martin, W. E., & Bridgmon, K. D. (2012). Quantitative and statistical research methods: From hypothesis to results. San Francisco, CA: Wiley.

Weakliem, D. L. (2016). Hypothesis testing and model selection in the social sciences. New York, NY: Guilford Press.

I

IEP

See Individualized Education Program

IES

See Institute of Education Sciences

David Teira David Teira Teira, David

Impartiality

Impartiality

805

806

Impartiality

Impartiality, or considering information without bias, is important in research, particularly in the social sciences. Because of its subjective nature, social science research depends on the impartiality of researchers, especially in the interpretation of data. In its simplest form, there are multiple examples of impartiality (or lack of it) in the history of science.

During the 20th century, the understanding of scientific impartiality shifted dramatically. In the first half of the century, philosophers and sociologists of science, at least those outside the Marxist tradition, generally agreed on the impartiality of both science and scientists. Logical positivism considered true scientific knowledge as a set of factual judgments, uncontaminated by the scientists’ values. Sociologist Robert K. Merton wrote that scientists shared a social norm of disinterestedness, setting their professional boundaries.

With the rise of the social studies of science and standpoint theory from the 1970s onward, impartiality was openly challenged as either an unattainable or undesirable ideal. For members of the so-called Edinburgh school, scientific concepts are shaped by social interests. Social epistemologists such as Miriam Solomon and Helen Longino have argued, in various ways, that scientists should acknowledge their biases and promote pluralism within their communities in order to make their research outcomes generally acceptable.

A common presupposition in these approaches to impartiality is that science should be conducted in the public interest. The more universal the public interest, the more impartial the research. Therefore, in biomedical research, the interests of the patients, being the greater number, should prevail over the particular interests of the drug developers when it comes to deciding about the true effects of a particular treatment. However, in the 21st century, we are gradually discovering how the individual interests of the scientists are often in conflict with the purported common good.

Sometimes, conflict between the interests of scientists and the public good occurs because research is conducted for profit and the financial stakes are high, as can be the case in biomedical research. But even when the financial stakes are low, conflicts of interest may arise due to the increasing pressure of academic competition for funding and positions. This pressure can cause conflicts of interest in fields where the financial stakes are low and is thought to be one reason why some highly cited experimental findings in social psychology are difficult to reproduce for third parties. Scientists are more likely to advance their careers and receive grants when they have had articles published in prestigious journals and when those articles are cited, creating an incentive for scientists to conduct experiments with positive results that attract attention.

There is a growing concern among both philosophers and concerned scientists about individual conflicts of interests. However, the purely methodological corrections for this sort of partiality (e.g., public registration of experimental data) have not been very successful so far. An open question is whether scientific institutions such as universities or journals have the enforcement power required to correct scientific biases in an environment of self-interested scientists.

David Teira

See also Collaborative Evaluation; Conflict of Interest; Goal-Free Evaluation

Further Readings

Teira, D., & Reiss, J. (2013). Causality, impartiality and evidence-based policy. In H.-K. Chao, S.-T., Chen, & R. Millstein (Eds.), Mechanism and causality in biology and economics. Dordrecht, the Netherlands: Springer.

Maggie Quinn Hannan Maggie Quinn Hannan Hannan, Maggie Quinn

Jennifer Lin Russell Jennifer Lin Russell Russell, Jennifer Lin

Improvement Science Research Improvement science research

806

808

Improvement Science Research

Improvement science research is enjoying increasing popularity as a strategy for creating scalable educational improvement. Although educational research has often focused on identifying what works, many in the field have come to believe that evidence that an innovation *can* work is not sufficient for large-scale, sustainable change, due to the challenges that result when innovations are implemented in diverse local contexts. In contrast, improvement science research aims to create practical improvements while generating knowledge about what works, for whom, and under what conditions. This entry further describes improvement science and how it developed and then discusses its principles and applications in education.

Improvement science emphasizes learning from rapid, iterative tests of small innovations conducted by practitioners working under diverse conditions. The process of experimentation begins with rigorous examination of a defined problem of practice and the specification of a theory of improvement. Systematic experimentation with changes in practice, in tandem with tracking process and outcome data, allows for the identification of promising change ideas and increasingly robust theories about how changes lead to improvement.

Although several strands of organizational and management theory contribute to improvement science, management scholar W. Edwards Deming crystallized the approach in the mid-20th century, using improvement science methods to rehabilitate Japan’s faltering automotive industry after World War II. Deming described improvement science as characterized by a combination of expert subject knowledge and what he called “profound knowledge,” which refers to an understanding of how to make changes in a system that will result in understanding of how to make changes in a system that will result in improvement. After Deming established the benefits of improvement science in manufacturing, other fields took notice and began to take it up. In the late 20th century, Donald Berwick and colleagues formed the Institute for Healthcare Improvement in order to promote the use of improvement science methods to address problems of practice in health-care organizations, such as surgical errors. More recently, educators and researchers have gravitated to the approach as a way to accelerate learning and improvement.

Improvement Science Principles

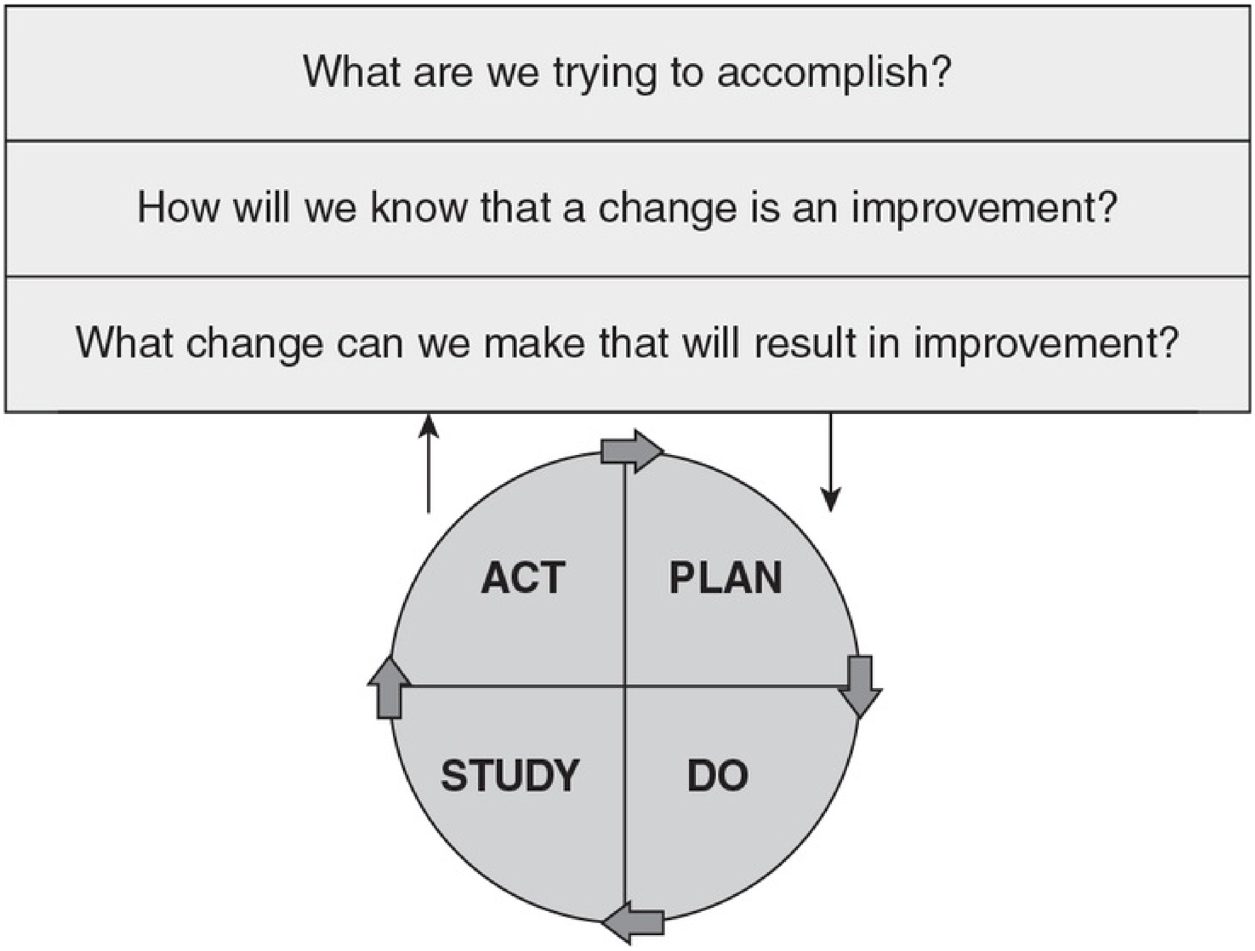
Improvement science offers a set of tools and processes for organization improvement based on Deming's concept of profound knowledge, which is comprised of four interrelated parts: appreciation for a system, understanding variation, building knowledge, and the human side of change. The first component, *appreciation for a system*, focuses on how processes interact with one another to create a system. A system is defined as an interdependent group of people, tools/objects, or processes working together toward a common purpose. Improvers' ability to "see the system" is crucial to creating meaningful change.

Understanding variation emphasizes the ability to interpret the variation observed in the performance of systems and processes. Learning about variation and its causes helps to guide appropriate actions for process improvement. Fast and practical measurement strategies, or leading indicators that are designed to be sensitive to diverse conditions, are essential to understanding variation.

The *building knowledge* component refers to the use of systematic methods for accumulating knowledge by making changes, observing or measuring results, and learning from the outcomes. The model for improvement (Figure 1) is a framework for testing theories and building knowledge, guided by three questions that focus improvement efforts:

> What are you trying to accomplish (aim)?
> How will you know a change is an improvement (measurement)?
> What changes can you make that will result in an improvement (theory of improvement)?

Figure 1 The Model for Improvement


What are we trying to accomplish?
How will we know that a change is an improvement?
What change can we make that will result in improvement?
ACT
PLAN
STUDY
DO

Source: Langley *et al.* (2009, figure 1.1, p. 24).

The model for improvement, coupled with a tool called the plan–do–study–act cycle, facilitates learning from carefully planned and measured changes in practice. The plan–do–study–act cycle is a way to test a process or practice change by *planning* to use it, *doing* it, observing and *studying* the results, and *acting* on what is learned. Another tool for operationalizing the model for improvement is the 90-day cycle, a routine for scanning the field relevant to a problem of practice and synthesizing knowledge held by scholars and practitioners about how to improve.

The final component of the system of profound knowledge is the *human side of change*, which emphasizes ideas, tools, and theories that help people better integrate changes into social systems. Drawing on research from psychology and change management, the human side of change emphasizes individual mind-sets, motivations, and preferences and how they intersect with change efforts.

In an authoritative text, *The Improvement Guide*, Gerald Langley and colleagues present practical guidelines on using improvement science in any organization. This volume delineates many strategies and tools that build on the system of profound knowledge to create positive organizational change.

Improvement Science Applications in Education

Following the productive use of improvement science in other fields, schools, districts, and departments of education have begun to establish quality improvement positions and offices and embed improvement science methodologies in core operations. The extent to which initiatives take up the improvement science principles outlined earlier in this entry varies widely, but some school systems have exemplified active use. For example, Sandra Park and colleagues have presented cases of the use of improvement science methods to support instructional improvement in classrooms in the Menomonee Falls School District in Wisconsin and to support district-level process improvement in service of better teaching and learning in the Montgomery County Public Schools in Maryland.

Researchers are also adopting improvement science methods in conjunction with traditional research methods in order to extend the impact of research and development. For example, the Carnegie Foundation for the Advancement of Teaching has been fostering a number of networked improvement communities that integrate improvement science methods in a networked structure in order to address pressing problems of educational practice. For instance, its Community College Pathways networked improvement community aims to improve community college students' educational attainment by supporting them through college-level math. The networked improvement community has enabled experimentation with curricula and small-scale classroom-level innovations in developmental mathematics classes in community colleges across the country.

Additionally, the Institute of Education Sciences offered funding to support the use of continuous improvement methods in education research. This line of funding has supported projects utilizing continuous improvement research, such as the work of researchers at the University of Pittsburgh working in collaboration with the Tennessee Department of Education to improve mathematics teaching and learning through instructional coaching throughout the state of Tennessee. Also supported by Institute of Education Sciences, the National Center on Scaling up Effective Schools at Vanderbilt University is using improvement science research to scale practices that have been shown to be effective for low-income and nondominant students in urban high schools. These and other emerging improvement science research projects in education are contributing to a growing knowledge base about scaling school improvements.

Maggie Quinn Hannan and Jennifer Lin Russell

See also Design-Based Research; Institute of Education Sciences; Program Evaluation

Further Readings

Berwick, D. M. (1989). Continuous improvement as an ideal in health care. The New England Journal of Medicine, 320(1), 53.

Bryk, A., Gomez, L., Grunow, A., & LeMahieu, P. (2015). Learning to improve. Harvard Education Press.

Cohen-Vogel, L., Tichnor-Wagner, A., Allen, D., Harrison, C., Kainz, K., Socol, A. R., & Wang, Q. (2014). Implementing educational innovations at scale transforming researchers into continuous improvement scientists. Educational Policy. doi:https://doi.org/10.1177/0895904814560886.

Deming, W. E. (1986). Out of the crisis. Cambridge, MA: Center for Advanced Engineering Study, MIT Press.

Hannan, M., Russell, J. L., Takahashi, S., & Park, S. (2015). Using improvement science to better support beginning teachers: The case of the Building a Teaching Effectiveness Network. Journal of Teacher Education, 66(5), 494–508.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). The improvement guide: A practical approach to enhancing organizational performance (2nd ed.). San Francisco, CA: Jossey-Bass.

Lewis, C. (2015). What is improvement science? Do we need it in education? Educational Researcher, 44(1), 54–61.

Park, S., Takahashi, S., & White, T. (2014). Developing an effective teacher feedback system. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Christine Ann Christle Christine Ann Christle Christle, Christine Ann

Inclusion

Inclusion

808

810

Inclusion

Inclusion refers to the practice of educating students with disabilities in a general education classroom with students without disabilities using specially designed instruction and supports. The term is generally distinguished from *mainstreaming,* which describes educating students with disabilities in the general education setting without specially designed instruction or supports. The emerging ideology of inclusion is that all children with or without disabilities have the right to participate actively in a general education setting as valued members of that learning community. This entry describes the background of inclusion in K–12 schools and discusses the advantages and disadvantages of inclusion.

A philosophical underpinning of the inclusion movement in the United States is a principle called *normalization* that emerged in Sweden in the 1960s. Normalization is a guiding principle for persons with disabilities that refers to making everyday life of the mainstream society available to them. However, individuals with disabilities in the United States were not allowed access to many aspects of mainstream society, and students with disabilities were not guaranteed access to public education until the mid-1970s.

In 1973, Section 504 of the Rehabilitation Act was passed and provided that any individual with a disability could not be excluded from any program or activity receiving federal funds, including public schools. Then, the passage of PL 94-142, the Education for Handicapped Children Act, was passed in 1975 and later reauthorized as the Individuals With Disabilities Education Act. This law afforded students with disabilities the right to a free and appropriate public education in the least restrictive environment. The least restrictive environment mandate states that students with disabilities should be educated in the general education setting with nondisabled peers to the maximum extent appropriate. The Individuals With Disabilities Education Act also requires that a continuum of placement options be available to meet the needs of students with disabilities. Although students with disabilities were allowed access to public schools, in the decade that followed, they were mostly educated in segregated classrooms.

The mid-1980s saw the launch of the Regular Education Initiative that encouraged a merger between general and special education to support greater academic gains for students with disabilities. This movement was a precursor to inclusion because it emphasized educating students with disabilities in general education classrooms and making adaptations to accommodate their needs. The Regular Education Initiative encouraged more mainstreaming for students with mild disabilities. However, it created much controversy and many educators rallied against it.

Parents and advocates of students with disabilities continued to push for more inclusive educational opportunities during the 1990s. These groups argued that segregated classes were exclusionary and stigmatizing, and there existed a lack of agreement as to the meaning of least restrictive environment. In 2001, another push for inclusion came in the form of the reauthorization of the Elementary and Secondary Education Act, known as the No Child Left Behind Act. The No Child Left Behind Act held states accountable for improving the quality of education for all students and expected schools to work toward 100% of their students meeting proficient status in reading, math, and science. Failure of any group of students in a school to meet the adequate yearly progress targets meant accountability measures for the entire school. With the No Child Left Behind Act’s focus on academic outcomes and access to the general curriculum, the pressure increased for educating students with disabilities in general education classrooms. Another legislative push for inclusion came in 2004 with the reauthorization of the Individuals With Disabilities Education Act, which mandated that students with disabilities be included in statewide assessments. This increased the focus on access to the general education curriculum for students with disabilities and on moving toward full inclusion.

The focus on accountability and high-stakes assessments accelerated the push for inclusion of the majority of students with disabilities in the United States now receiving most of their education in general education classrooms. The debate is no longer about access to the general education classroom but rather about the no longer about access to the general education classroom but rather about the benefits. The empirical evidence for inclusive education reveals mixed outcomes, with some research indicating better outcomes for students with disabilities in general education classrooms and other results demonstrating better outcomes in separate classrooms such as special resource classes.

The arguments both for and against inclusion are based on social and philosophical grounds as well as academic benefits. Some of the purported benefits of inclusion are (a) it better prepares students with disabilities for adult life in the community, (b) peers without disabilities help students with disabilities develop social skills and friendships, (c) peers without disabilities learn to interact with others who are different from them, and (d) both general and special education teachers benefit from collaboration and improve their teaching skills in inclusive classrooms.

Some of the arguments against inclusion are (a) the empirical evidence is inconclusive on the effectiveness, (b) many students with disabilities need more intensive interventions that cannot be provided in general education classrooms, (c) general education teachers lack the training and supports needed to effectively teach students with disabilities, and (d) many teachers and students have negative attitudes and beliefs about students with disabilities and inclusion.

Many argue that the major concern for students with disabilities should be improving their learning outcomes, not where or with whom they learn. The first consideration should be what constitutes an appropriate education for them, and the focus should be on the quality of instruction in any setting, which leads to higher achievement.

Christine Ann Christle

See also Accommodations; Adequate Yearly Progress; Individuals With Disabilities Education Act; Least Restrictive Environment; No Child Left Behind Act

Further Readings

Hicks-Monroe, S. L. (2011). A review of research on the educational benefits of the inclusive model of education for special education students. Journal of the American Academy of Special Education Professionals, Winter, 61–69.

Michaud, K., & Scruggs, T. E. (2013). Inclusion in the United States: Theory and practice. In C. Boyle & K. Topping (Eds.), What works in inclusion? (pp. 20–30). Berkshire, UK: Open University Press, McGraw-Hill.

Sailor, W., & Roger, B. (2005). Rethinking inclusion: Schoolwide applications. Phi Delta Kappan, 86(7), 503–509.

Christine Ann Christle Christine Ann Christle Christle, Christine Ann

Individualized Education Program Individualized education program

810

813

Individualized Education Program

Prior to the mid-1970s, students with disabilities had limited access to public education. In 1975, the U.S. Congress passed the Education for All Handicapped Children Act (later reauthorized as the Individuals With Disabilities Education Act [IDEA]) to provide a free appropriate public education for all students with disabilities. Congress mandated that an individualized education program (IEP) be developed to ensure that students with disabilities receive an individualized and appropriate education.

The IDEA established a process for school-based teams to develop IEPs that includes (a) assessing the student’s educational needs, (b) developing meaningful and measurable annual goals, (c) developing and implementing the special education program and services, and (d) monitoring progress toward the goals. Throughout this process, the IEP teams must ensure they follow both procedural and substantive requirements of the law. Procedural requirements include the process IEP teams use to develop a student’s special education program and the document itself. Substantive requirements refer to developing special education programs that lead to meaningful educational benefit for students with disabilities. This entry describes the procedural and structural requirements for the IEP.

Procedural Requirements

One procedural requirement is that the IEP meetings must be scheduled at a mutually agreeable time and place for the parents and school IEP team members. This requirement includes procedures for contacting parents, such as using at least two different types of contact (e.g., phone, letter, and e-mail). Another requirement involves the IEP team members. These include (a) the student’s parent or guardian, (b) the student’s special education teacher, (c) at least one of the student’s general education teachers, (d) a representative of the local education agency, (e) a person who can interpret the evaluation results, and (f) the student, if appropriate. Others with special knowledge or expertise also may be invited to the meetings.

An additional requirement spells out certain components that must be included in the IEP. These are (a) present levels of academic achievement and functional performance (PLAAFP); (b) measurable annual goals; (c) special education services; (d) the method for collecting and reporting the student’s progress; (e) the student’s participation in statewide or districtwide assessments; (f) the extent to which the student will not participate in general education settings; (g) the projected date for beginning services, anticipated frequency, location, and duration; and (h) transition services if the student will be 16 years old during the period for which the IEP is developed (many states require transition services at a younger age).

In general, the IEP must sufficiently describe the following: (a) the student’s specific educational needs, (b) the services needed to meet these needs, and (c) how to determine whether the needs were effectively addressed. The IEP is the centerpiece for IDEA and the process that ensures a free appropriate public education is provided for students with disabilities. Procedural violations of the IEP are frequent sources of litigation against school districts. It is crucial that the IEP team members are knowledgeable in the required components of the IEP.

IEP Components

PLAAFP

The component of the IEP that describes PLAAFP includes statements that provide the starting point and foundation for the student’s entire IEP. The PLAAFP information must be based on a full and individualized assessment to determine a student’s unique educational needs, and parental perceptions and concerns must be considered. These statements must (a) describe the impact of the student’s disability on the student’s performance in all areas that are affected by the disability (e.g., reading, math, and behavior), (b) be written in objective and measurable terms that are easily understood by all members of the IEP team, and (c) describe how the student’s disability affects the student’s involvement and progress in the general curriculum and progress in the general curriculum.

There are many sources of data to help ascertain the student’s academic and functional levels. Some of these include (a) statewide assessment results, (b) districtwide assessment results, (c) classroom performance information from all teachers, (d) previous intervention results, (e) observational data, (f) behavioral data (i.e., checklists, rating scales, and discipline referrals), and (g) curriculum-based measurements. Once the IEP team has gathered measurable data on the student’s academic and functional needs, it can make comparisons to students without disabilities. The effect of the student’s disability on involvement and progress in the general curriculum is crucial because it explains the need for special education services. Once the specific PLAAFP statements have been established, they become the baseline for measurable goals.

Measurable Annual Goals

The purpose of annual goals is to measure the student’s progress in the student’s special education program. The measureable annual goals in the IEP begin with measureable PLAAFP statements. There are four components to a measureable annual goal: (1) the observable behavior or skill to be taught (e.g., to write, to read aloud), (2) the condition or materials that will be used to measure change (e.g., the context, curricula, or environment for the behavior or skill), (3) the criteria for goal achievement (e.g., accuracy, speed), and (4) the time line (e.g., in 1 year, by the review date of the IEP). The following steps outline how to develop measurable annual goals:

1. Start with the PLAAFP data.
2. Determine the condition.
3. Describe the observable behavior.
4. Decide an ambitious rate of progress.
5. Determine the timeline.
6. Decide the criteria.
7. Write the goal statement.

The IEP teams can ensure that their goals are measurable by making a graph depicting the PLAAFP, the goal date and rate, and a goal line.

Special Education Services

Special education services describe the educational services the school will provide to help effectively address the student’s unique educational needs. The services must be aligned with the PLAAFP and the measurable annual goals in the IEP so that the students with disabilities will (a) advance appropriately toward meeting their annual goals, (b) advance in the general curriculum, and (c) be educated with their peers. Special education services may include related services, supplementary services, accommodations, and modifications to the curriculum. For example, a service statement may indicate “direct instruction in reading by the special education teacher in the resource room for 1 hour per day, 5 days per week.”

Related services are developmental, corrective, and supportive services required so that the student with disabilities can benefit from special education services. Examples of related services include (a) occupational and physical therapy, (b) parent counseling and training, (c) psychological services, (d) school health services, (e) social work services, (f) interpreting services, and (g) transportation. When related services are provided to a student with a disability, they must be included in the IEP, and they must be provided at no cost.

The IDEA stipulates that IEP teams make good faith efforts to include students with disabilities in general education settings by using supplementary aids and services. These are services provided in general education classes, other education-related settings, and extracurricular and nonacademic settings. Examples include (a) accommodations and modifications to the curriculum, (b) paraprofessional supervision and support, (c) a special education teacher coteaching with the general education teacher, and (d) staff training.

Reporting Progress

The IEP must state (a) how the student’s progress toward meeting the annual goals will be measured, (b) how the parents will be informed of the progress their child is making toward the goals, and (c) the extent to which that progress is sufficient to enable the child to achieve the goals by the end of the year. Parents are to be informed of their child’s progress on all goals at least as often as parents of children without disabilities are informed. For example, an annual goal may state, “By June 2015 when given a sixth-grade reading passage, the student will increase the number of correct words read aloud in 1 minute from 120 to 157.” A statement in the IEP regarding measurement and reporting progress may be, “The special education teacher will measure progress once per week using curriculum-based measurement and a progress report will be sent to parents every 4 weeks.” An example progress report may state, “The student has parents every 4 weeks." An example progress report may state, "The student has improved in oral reading fluency and currently reads 130 correct words per minute. At this rate, the student will likely reach the goal of 157 correct words per minute by the end of the year."

Participation in Assessments

The IDEA requires that students with disabilities participate in statewide or districtwide assessments and the IEP must indicate how the student will participate. If the IEP team determines that the student cannot be accurately assessed by taking the regular assessments, it must specify the appropriate accommodations needed. Any accommodation regularly used in instruction may be used on assessments for students with disabilities. If the IEP team determines that the student needs to be assessed with an alternative assessment, the IEP must include a statement of why the student cannot participate in the regular assessment and why the particular alternate assessment selected is appropriate for the student.

Participation in General Education Settings

The least restrictive environment mandate of the IDEA requires that students with disabilities receive their education in settings with students without disabilities to the maximum extent appropriate. When an IEP team decides a more restrictive option is needed, the IEP must specify the extent to which the student will not participate with students without disabilities in the general education placement, including academic and nonacademic activities.

Service Details

The IEP must also include the projected date for the beginning, frequency, location, and duration of services. This is to ensure that the services begin as soon as possible after the IEP has been developed. It also quantifies the school's commitment of resources to address the student's needs.

Transition Services

If the student is 16 years old (or younger in some states) during the period for which the IEP is developed, the IEP must include appropriate, measurable postsecondary goals related to training/education, employment, and where appropriate, independent living skills. There must also be a statement of appropriate, independent living skills. There must also be a statement of transition services, including appropriate courses of study to assist the students with disabilities in reaching the postsecondary goals.

Substantive Requirements

The substantive mandates of IDEA require that the student's IEP is relevant and appropriate to the student's needs and demonstrates meaningful benefit. To accomplish this, a complete and individualized assessment of the students' needs must address all areas of suspected disability to identify the relevant domains. The subsequent PLAAFP information should be written in objective and measurable terms to provide the baselines for the ambitious, clear, and measurable goals. The IEP teams can use curriculum-based measurements and other normative scales to determine ambitious IEP goals. These scales are based on research indicating the anticipated average progress per grade level and the level of progress that would represent ambitious growth at each grade level. The IEP teams can then compare the student's growth to that of the average peers and to the student's own baseline levels to demonstrate meaningful benefit.

IDEA also requires that special education services be based on peer-reviewed research to the extent practicable and that the services must be implemented as written. The intent of this mandate is to ensure that IEP teams select sound educational practices. Thus, teachers need to seek interventions that are supported by peer-reviewed research that are appropriate for the student and implement them accurately. The description of services must include location, frequency, and duration, so that the exact nature of the services provided will be clear to all involved. Finally, the IEP should be updated as needed but must be reviewed at least annually.

Christine Ann Christle

See also Curriculum-Based Measurement; Individuals With Disabilities Education Act; Least Restrictive Environment; Special Education Law

Further Readings

Bateman, B. D. (2011). Individual education programs for children with disabilities. In J. M. Kauffman & D. P. Hallahan (Eds.), Handbook of special education (pp. 91–106). Philadelphia, PA: Taylor … Francis.

Christle, C. A., & Yell, M. L. (2010). Individualized educational programs: Legal requirements and research findings. Exceptionality, 18(3), 109–123.

Yell, M. L. (2016). The law and special education (4th ed.). Upper Saddle River, NJ: Pearson.

Christine Ann Christle Christine Ann Christle Christle, Christine Ann

Individuals With Disabilities Education Act Individuals with disabilities education act

813

818

Individuals With Disabilities Education Act

The Individuals With Disabilities Education Act (IDEA) is a federal law that provides funding to states and localities to support them in educating students with disabilities by protecting their rights, meeting their individual educational needs, and improving results. The primary purpose of this law is to provide a free appropriate public education specifically designed to meet the individual needs of students with disabilities. This entry describes the background of this landmark legislation, the structure of the law, and the major provisions that school district personnel need to follow. The entry concludes with a discussion of the impact of IDEA since its first iteration in 1975 (then titled the Education for All Handicapped Children Act).

Background of IDEA

Before IDEA, few special education programs existed for children with disabilities, and most were private or residential. The availability of educational programs and their quality varied greatly within and between states, and many children with disabilities were denied access to any educational opportunities. During the 1950s and 1960s, the civil rights movement brought the issue of segregation in education in the United States to national attention. In 1954, the U.S. Supreme Court issued a landmark civil rights decision in *Brown v. Board of Education*, ruling that segregation in public education by race was unlawful. The Court held that when a state provides an education to some of its citizens, it must equally provide it for all of its citizens.

After the *Brown v. Board of Education* decision, parents of children with disabilities and advocacy groups began using the courts to push for equal educational opportunities for children with disabilities, arguing that by excluding these children, schools were discriminating against them on the basis of their disabilities. After several lawsuits were filed against individual states, Congress began an investigation into the status of education for children with disabilities. They found that millions of children with disabilities were not receiving an appropriate education.

Congressional efforts began to advance more educational opportunities for all children with disabilities until finally in 1975 the 94th Congress passed Public Law 94–142—the Education for All Handicapped Children Act. This landmark legislation guaranteed a free appropriate public education to every student with a disability aged 3 through 21. Additionally, the law contained procedural safeguards to protect the rights of students with disabilities and their parents and assistance to states and localities in order to provide a free appropriate public education to these students. Between 1975 and 2004, several important amendments were made to the law that have expanded the educational rights of students with disabilities.

In 1986, the Education of the Handicapped Amendments (P.L. 99–457) expanded services for infants and toddlers from birth to 2 years old with disabilities or those at risk for disabilities. Another law passed in 1986, the Handicapped Children’s Protection Act, allowed courts to award attorney’s fees to parents or guardians who prevailed in lawsuits under IDEA. In 1990, several changes were made to the law, including renaming Education of the Handicapped Amendments as Individuals With Disabilities Education Act. The terms *handicapped child* or *handicapped student* were changed to *child with a disability* or *student with a disability*, emphasizing the person first. Separate disability categories were recognized for students with autism and students with traumatic brain injury. The law was also changed to include a transition plan in the individualized education program (IEP) of students 16 years and older to begin planning for their posthigh school years.

In 1997, Congress made a number of additional changes to IDEA, such as requiring that IEPs focus on improving educational results and encouraging higher expectations for students with disabilities. This shifted a focus from the procedural requirements of the law to the substantive requirements of providing meaningful educational benefit for students with disabilities. A requirement was added that IEP teams must consider whether assistive technology devices and services are needed for the student to benefit from special education and related services are needed for the student to benefit from special education and related services. Other changes in 1997 promoted educating students with disabilities with their peers without disabilities in the general curriculum and including students with disabilities in state-and districtwide assessments. Disciplinary provisions were provided to ensure schools are safe and conducive to learning, while also protecting the rights of students with disabilities.

Congress reauthorized and amended IDEA in 2004 to increase the quality of special education programs for students with disabilities by increasing accountability for results. The following are some of the important changes made to the law to accomplish this. Districts now have more leeway in determining how students with disabilities are identified. For instance, districts are no longer required to use the discrepancy between IQ and academic performance to determine whether a student has learning disabilities. They may use a process that is based on the child's response to scientific, research-based interventions.

Another important change in IDEA 2004 allows districts to use funds for developing "early intervening services" for students in Grades K–12. These services are intended for students who have not been identified as needing special education or related services but who need additional academic or behavioral support to succeed in general education classes. The 2004 amendments also focus on making IEPs more relevant to student progress and reducing paperwork.

A significant change for teachers in IDEA 2004 is the requirement that special education teachers meet the law's highly qualified definition, by obtaining full state certification as a special education teacher or passing their state's special education teacher licensing exam and holding a license to teach in the state. In addition, special education teachers who teach core academic subjects must demonstrate subject-matter competency in each subject taught. Another change to the law provides guidance on disciplining students with disabilities, such as determining when a change of placement is appropriate and using manifestation determination guidelines. IDEA 2004 also specifies that all children with disabilities who need instructional materials in accessible formats receive them in a timely manner.

The Structure of IDEA

IDEA includes five major parts, Parts A, B, C, D, and E. Part A lays out the basic foundation for the rest of the law, including definitions of the terms used within it. Part B lays out the educational guidelines for students with disabilities aged 3 through 21 and describes the provisions that school districts must comply with in order to receive funding. Part C provides guidelines concerning the funding and services for infants and toddlers from birth through 2 years of age and their families. Part D describes a variety of national activities such as grants and other resources to improve education and transition services for students with disabilities. Part E establishes the National Center for Special Education Research to expand knowledge and understanding of the needs of children with disabilities in order to improve their developmental, educational, and transitional services and results.

Students are determined to be eligible to receive services under IDEA if they qualify under one of the 13 categories defined in the law. These categories are (1) autism, (2) deaf–blindness, (3) deafness and hearing impairment, (4) emotional disturbance, (5) intellectual disability, (6) multiple disabilities, (7) orthopedic impairment, (8) other health impairment, (9) specific learning disability, (10) speech or language impairment, (11) traumatic brain injury, (12) visual impairment, including blindness, and (13) developmental delay (preschool). A multidisciplinary team makes the determination of whether a student meets the definition of having one of these disabilities and whether the disability has an adverse impact on the student's education. Some students with disabilities may not qualify for services under IDEA if their disability does not adversely impact their educational achievement.

Major Provisions of IDEA in Part B

Part B of IDEA contains provisions that school personnel need to follow to provide special education and related services to students with disabilities aged 3 through 21. Although not delineated in the statutory language of IDEA as major provisions or principles, the following requirements are often discussed in the professional literature: (a) zero reject, (b) protection in evaluation, (c) free appropriate public education, (d) IEP, (e) least restrictive environment, (f) parent and student participation in decision making, and (g) procedural safeguards. School personnel who are involved with the education of students with disabilities should be familiar with these requirements.

Zero Reject

Zero Reject

This requirement means that each state education agency is responsible for locating, identifying, and evaluating all children, from birth to age 21, residing in the state who have disabilities or who are suspected of having disabilities. This provision is called the Child Find system and it applies regardless of the severity of the disability. School districts have an affirmative duty to seek out all students with disabilities in their jurisdiction. When students are identified in the Child Find system, the school district must determine whether the student qualifies for services under IDEA. This is done by a multidisciplinary team and includes a comprehensive evaluation.

Protection in Evaluation

Prior to providing special education services, a student must be evaluated to determine whether the student is an eligible “child with a disability” according to IDEA definition and, if so, to determine the educational needs of the student. Parents must give informed consent for the evaluation and for services, and an evaluation must be conducted within 60 calendar days of the parent giving consent. The student must be evaluated in all areas of suspected disability using a variety of tools and strategies to gather functional, developmental, and academic information.

The evaluation instruments and methods used must be (a) technically sound, (b) not culturally discriminatory, (c) in the language the child uses, and (d) administered by trained and knowledgeable personnel. In addition, a new or updated evaluation is to be conducted if there is reason to suspect a need or if the parents request one. A comprehensive reevaluation must be conducted every 3 years unless both the parent and school agree it is not necessary. Reevaluations may also occur when conditions warrant or when parents request them. Parents also have a right to request an independent evaluation at public expense if they disagree with the results of the school’s evaluation. Of course, parents may seek an independent evaluation at their own expense at any time.

Free Appropriate Public Education

Students who are eligible for services under IDEA have a right to a free appropriate public education that consists of special education and related services that (a) are provided at public expense under public supervision, (b) meet the standards of the state department of education, (c) are designed to meet the unique needs of each eligible student, and (d) are provided according to a written IEP. The special education program must be designed for the student to progress in the general education curriculum and provide meaningful educational benefit. This includes related services and supports and extracurricular activities. Related and supportive services may include, but are not limited to, speech and language therapy, psychological services, counseling services, physical and occupational therapies, and social work services.

IEP

An IEP is a program for each student with a disability described in a written document. The program is to be developed, reviewed, and revised at least annually by a team. The team members include (a) the student’s parent or guardian, (b) the student’s special education teacher, (c) at least one of the student’s general education teachers, (d) a representative of the local education agency, (e) a person who can interpret the evaluation results, and (f) the student if appropriate. Others with knowledge or expertise needed for the development of the student’s special education program may be invited to the meetings.

The IEP team must develop the individual instructional program without regard to where it will be implemented. That is, the focus must be on the program first and placement second. Parents and the student need to be meaningfully involved in the development and revisions of the program. Further, IDEA requires that the IEP include the following components: (a) present levels of academic achievement and functional performance; (b) measurable annual goals; (c) special education services; (d) the method for collecting and reporting the student’s progress; (e) the student’s participation in statewide or districtwide assessments; (f) the extent to which the student will not participate in general education settings; (g) the projected date for beginning services with anticipated frequency, location, and duration; and (h) transition services if the student will be 16 years old (many states require transition services at a younger age). Finally, IDEA mandates that the student’s IEP is relevant and appropriate to the student’s needs and demonstrates meaningful benefit.

Least Restrictive Environment

IDEA requires that students with disabilities are to be educated with their peers without disabilities to the maximum extent appropriate; and to achieve this, schools must provide supplementary aids and services in the general education classroom or other integrated settings. Any placement outside the general education classroom must be justified by the child's disability-related need. The intent of the law is to ensure that students with disabilities have meaningful access to same-age peers without disabilities. However, in cases where the severity of the student's disability is such that the student cannot receive an appropriate education in the general education classroom with supplementary aids and services, the least restrictive environment will not be the general education classroom. School districts must maintain a continuum of alternative placements to ensure that students are placed in the most appropriate and least restrictive setting. The continuum consists of the regular classroom, self-contained classrooms, special schools, and hospital or institutional settings.

Parent Participation in Decision Making

The provision that parents must participate in decision making is one of the cornerstones of IDEA, as parents of students with disabilities play an extremely important role in helping schools meet the educational needs of their children. Thus, parents and students (whenever appropriate for the student) are to be meaningfully involved in (a) determining what data need to be collected during evaluation; (b) educational placement decisions; (c) reviewing evaluation data; (d) the development, review, and revision of the IEP; and (e) transition planning and services starting by age 16.

Procedural Safeguards

Protecting the rights of students with disabilities and their families is a key purpose of IDEA and thus one of the most important principles of the law. IDEA includes an extensive set of procedures that school personnel must follow to ensure that parents are meaningfully involved in the process. These include (a) providing parents with notices of their rights, of meetings, and of programming or placement changes for their child; (b) obtaining informed parental consent for evaluation and programming; (c) providing parents access to their child's records; and (d) the right to request an independent educational evaluation if they disagree with the district's evaluation.

Another important aspect of this principle is a clear and systematic set of procedures to follow when there are disagreements between the school personnel and parents of students with disabilities. This process begins with informal problem-solving meetings and voluntary mediation. If these do not resolve the disagreement, either party may request a due process hearing. This takes place at the district level with a formal administrative hearing before an administrative law judge to decide the disputes between parents and educators that relate to the provision of special education. After conducting the hearing, the judge will issue a decision that can be appealed to the state level and then in a civil court.

The Office of Special Education and Rehabilitative Services within the U.S. Department of Education is responsible for implementing, monitoring, and enforcing IDEA. Each year states must show compliance with IDEA by submitting their state plans to the Department of Education. If a state is found not to be compliant with the provisions and regulations of IDEA, the Department of Education may withhold the funds provided through the law.

Impact of IDEA

Each year, the secretary of the U.S. Department of Education is required to submit an annual report to inform Congress and the public of the progress made in implementing IDEA. This report describes the nation’s progress in (a) providing a free appropriate public education for all children with disabilities and early intervention services to infants and toddlers with disabilities and their families, (b) ensuring that the rights of these children with disabilities and their parents are protected, (c) assisting states and localities in providing for the education of all children with disabilities, and (d) assessing the effectiveness of efforts to educate children with disabilities.

Before IDEA was enacted, many individuals with disabilities lived in state institutions and only received basic needs of food, clothing, and shelter rather than education and rehabilitation. Much progress has been made toward protecting the rights of, meeting the individual needs of, and improving educational results for infants, toddlers, children, and youths with disabilities. School personnel are using educational approaches, practices, and techniques grounded in research to include students with disabilities in general education classrooms. As a result, students with disabilities are experiencing more success in school. Many are attending colleges, universities, and other postsecondary programs. They are finding suitable employment, living independently, and accessing community services. Nevertheless, there is still much work to be done to improve outcomes for students with disabilities for them to realize their potential as citizens in their communities.

Access to education has been a critical civil rights issue for historically underrepresented groups, and by adopting this landmark legislation, Congress laid the foundation for the nation’s commitment to ensure that children with disabilities have opportunities to learn alongside their peers without disabilities, to reach their individual potentials, and to contribute to their communities. Although it is not perfect, IDEA has been the framework for advancing the education of students with disabilities and improving their lives.

Christine Ann Christle

See also Inclusion; Individualized Education Program; Least Restrictive Environment; Special Education Identification; Special Education Law

Further Readings

Council for Exceptional Children. (n.d.). IDEA. Retrieved from https://www.cec.sped.org/Policy-and-Advocacy/Current-Sped-Gifted-Issues/Individuals-with-Disabilities-Education-Act

U.S. Department of Education, Office of Special Education and Rehabilitative Services. (2017, February 16). Individuals With Disabilities Education Act. Retrieved from https://www2.ed.gov/about/offices/list/osers/osep/osep-idea.html

U.S. Department of Education, Office of Special Education and Rehabilitative Services. (2010). Thirty-five years of progress in educating children with disabilities through IDEA. Washington, DC: Author.

Wright, P. (2010, November 29). The history of special education law. Wrightslaw. Retrieved from http://www.wrightslaw.com/law/art/history.spec.ed.law.htm

Yell, M. L. (2016). The law and special education (4th ed.). Upper Saddle River, NJ: Pearson.

Yell, M. L., Katsiyannis, A., & Bradley, M. R. (2011). The Individuals With

Disabilities Education Act. In J. M. Kauffman & D. P. Hallahan (Eds.), Handbook of special education (pp. 61–76). Philadelphia, PA: Taylor … Francis.

Michael A. Seaman Michael A. Seaman Seaman, Michael A.

Inferential Statistics Inferential statistics

818

820

Inferential Statistics

The term *inferential statistics* refers to applying statistical analysis with observed data for the purpose of making inferences to that which cannot be observed. Although a descriptive statistic is an index that is calculated on a set of data to represent some property of that data, an inferential statistic is calculated from the data as a means of inferring more general properties that go beyond observable data. A common way to conceptualize inferential statistics is to consider that researchers are interested in understanding some property, such as center or variability, of data for a population (e.g., all fourth graders in the United States), yet there are constraints (e.g., access and cost) that keep them from collecting all of the data. Consequently, researchers would obtain data for a subset of the population, called the sample, and then use these data to make inferences to the larger population. The validity of such inferences depends on various factors such as how the sample data are obtained and whether the sample is representative of the population. In practice, it is often not possible to obtain data that strictly meet the requirements for valid inference, yet inferences can be useful approximations of properties of unobservable data. This entry describes methods for calculating inferential statistics, types of inferential statistics, types of inference, and philosophies of probability.

Inferential Methods

Technically, inferential statistics refer to numerical indices used for inference, yet the term often is used to apply to a collection of methods for calculating inferential statistics. These methods include *t* procedures, analysis of variance, chi-square procedures, the Wilcoxon method, the Kruskal–Wallis method, and many others. A particular method is appropriate for specific configurations of the study, such as the number of explanatory and response variables, and whether the data for each variable are obtained at a categorical or quantitative level of measurement. The method leads to the calculation of statistics that can be used for inference, which is why both the methods and the statistics themselves often fall under the general heading of inferential statistics.

Types of Inferential Statistics

An inferential method applied to a set of data will result in an index that is an inferential statistic. For example, applying analysis of variance techniques to a set of data will result in an *F* statistic. Although this is an important inferential statistic, it is an intermediate step in the inferential process. In order to make inferences, the researcher must compare this observed statistic to the larger collection of all possible values that could have been obtained for the statistic. If researchers hypothesize characteristics of a population or multiple populations in a study (e.g., it might be hypothesized that the characteristics of multiple populations are all equal), they can calculate not only the possible values of an inferential statistic but also how likely it is to obtain a value of the inferential statistic that is within some specified range. For example, researchers might calculate that there is only a 5% chance that they will observe an *F* statistic that is greater than 4.0. This ability to make such calculations allows them to calculate inferential statistics that are linked to probability statements.

Two of the most common probability-type inferential statistics are *p* values and confidence intervals. A *p* value is the probability of obtaining an inferential statistic from a set of data that is at least as large as the observed statistic (e.g., the observed *F* statistic) if a specified hypothesis (e.g., multiple populations have equal characteristics) is true and the data were obtained through random selection. As such, researchers use a type of "reverse reasoning" with *p* values, reasoning that obtaining an inferential statistic that is associated with a small *p* value is evidence that the initial hypothesis is false. The association of *p* values with this type of reasoning, as well as the fact that *p* values do not answer the most important questions in research, which have to do with the differences in populations and the effectiveness of treatments, has led many researchers to dispose of *p* values, or at least supplement them, in the research process.

A more widely accepted inferential statistic is a confidence interval. A confidence interval is a range of possible values for an index of some characteristic of a population that is constructed, so that there is a specified probability that this interval will capture the actual value of the index in the population. For example, there is a 95% chance that the mean of a population will be contained in a 95% confidence interval for the mean of that population. Thus, confidence intervals provide a statement of the characteristic of the population that is of interest and attach to it a probability or level of confidence.

Types of Inference

The most common way that most researchers think of inferential statistics is as a way to make an inference about a population from a sample of data. This is referred to as population inference and is appropriate if data are randomly selected from the population of interest. Another type of inference is causal inference. With causal inference, inferential statistics lead to a probability statement about a causal relationship among two or more variables. Although population inference depends on randomly sampling from a population, causal inference depends on randomly assigning study participants (known as randomizing) to study conditions. Even if study participants are not selected at random, if they are randomly assigned to conditions, calculating inferential statistics can lead to probability statements about observed effects representing actual cause–effect relationships, rather than the result of chance assignment to conditions.

In studies that include both random selection and random assignment, inferential statistics can lead to both population and causal inference. It is often the case that a study does not have either random selection or assignment. In these cases, inferential probability statements are approximations that depend on either the representativeness of sample data to the larger domain of interest or the similarity of the study participants across the conditions of the study. Even when such representations are doubtful, inferential statistics can be useful to help researchers distinguish among real and perceived patterns in the data.

Philosophies of Probability

The statements about inference in this entry are based on the frequentist perspective, which is one of the two major philosophies about probability and thus about inference. The frequentist view of probability is still the most common and is almost universally taught in beginning statistics courses in education. For frequentists, there are unknown characteristics of populations, and researchers can make a limited number of observations in order to construct statements of probability regarding the chances that they have captured these characteristics. The other major philosophy is Bayesian. Bayesians use prior knowledge about a population to inform inference and then collect data as a means of improving or updating this information. For a Bayesian, probability is about the degree of belief. It is a subjective statement. By contrast, a frequentist statement of probability is based on a large number of repeated samplings. Statisticians argue about which approach is most useful, but all understand that the different perspectives lead to different ways of using and talking about inferential statistics.

Michael A. Seaman

See also Bayesian Statistics; Confidence Interval; Descriptive Statistics; *F* Distribution; *p* Value

Further Readings

Agresti, A., & Finlay, B. (2014). Statistical methods for the social sciences (4th ed.). Essex, UK: Pearson.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. Journal of Educational and Behavioral Statistics, 20, 115–147.

Moore, D. S., Notz, W. I., & Fligner, M. A. (2015). The basic practice of statistics (7th ed.). New York, NY: Freeman.

Trochim, W., Donnelly, J. P., & Arora, K. (2016). Research methods: The essential knowledge base. Boston, MA: Cengage.

Claudia A. Gentile Claudia A. Gentile Gentile, Claudia A.

Information Processing Theory Information processing theory

Information Processing Theory

Information processing theory, which arose in the 1940s and 1950s, seeks to explain how the mind functions and encompasses a range of processes, including gathering, manipulating, storing, retrieving, and classifying information. While information processing theories are used to inform instructional design and approaches to learning, these theories tend to emphasize the understanding of how information is processed rather than how learning happens. This entry examines the core beliefs of information processing theory as well as its applications to theories of intelligence and development and to learning and instruction.

Unlike the behaviorist perspective about how the human mind functions, which focuses on how people respond to stimuli, information processing theory posits that the human mind is like a computer or information processor. Information is gathered through the senses (the brain’s input devices) and processed via short-term memory (the brain’s CPU), resulting in storage in long-term memory (the brain’s hard drive storage). Long-term memory includes three types of knowledge: declarative (knowing that), procedural (knowing how), and episodic (personal stories). Some researchers argue that our memory for images differs from our memory for words and that our memory for other senses may differ as well. Other researchers have focused on the mechanisms we use to control how we process information (metacognitive processes and strategies).

Researchers have expanded upon the basic metaphor of “brain as computer” in several ways. Some researchers have focused on the sequential nature of information processing (e.g., stage theory’s three-stage model: input→processing→output). Others explored the relationship between how information is processed and our ability to later access the information (e.g., the level of processing model). Research using the level of processing model has level of processing model). Research using the level of processing model has found that the degree of elaboration affects how well information was learned and that information was more easily retrieved if the way it was accessed was similar to the way it was stored. The connectionist model, supported by neuroscience research, focuses on how information is stored simultaneously in different areas of the brain and is connected as a network. Research using this model found that the ease of retrieval of a piece of information was related to the number of connections it had.

Applications to Theories of Intelligence and Development

Information processing theory is a component of several major theories of human intelligence and development. Notably, Robert Sternberg’s theory of intelligence includes information processing as a key component and posits that information processing is comprised of three parts: (1) meta-components that involve planning and evaluating problems, (2) performance components that involve implementing the plans, and (3) knowledge-acquisition components that involve learning from the planning and implementation phases.

While information processing theory is often viewed as an alternative to Piaget’s theory of cognitive development, in Jean Piaget’s theory, the four stages of growth are characterized, in part, by the type of information processed and by distinctive thought processes. The sensory motor phase (from birth to 2 years) involves the use of the five senses to process information, with responses based on reflexes. The preoperational phase (2–6 years) involves learning through imitation and the inability to view situations from another’s viewpoint. The concrete operational phase (6–11 years) involves the development of the ability to use logic and consider multiple factors to solve problems. The formal operational phase (11 years and older) involves planning, processing, and understanding abstract concepts as well as the ability to create arguments and evaluate risks and benefits.

However, while Piaget’s theory of cognitive development conceives of development in stages, information processing theory views the process of development as continual: As we grow, our brains mature, leading to advances in our ability to process and respond to increasingly more information and more complex information.

Applications to Learning and Instruction

Researchers and educators across disciplines use information processing theory to explore students’ learning and to design instruction across all subject areas and grade levels. Instructional design models often incorporate two perspectives from information processing theory on how the human mind works: (1) we constantly process information using a complex series of systems and (2) our processing systems modify the information we gather in systematic ways. They also focus on three types of skills to enhance the development of students’ abilities to process information: focusing skills, information-gathering skills, and remembering skills.

Educators also use information processing theory approaches to design curriculum and adapt instruction to meet students’ varying needs, including students classified as English learners and those receiving special education services. With the perspective that learning represents the process of gathering information and organizing it into mental schemata, instructional programs are developed that employ learning strategies to improve students’ retention and the retrieval of information, such as the use of multimedia to engage students’ attention and increase their memory for information presented in videos.

Overall, information processing theory provides educators with several key implications for designing instruction. First, because sensory and working memory stores are limited, instructional programs need to focus students’ attention on important information and engage in as much automated processing as possible (e.g., mastering basic skills for reading and mathematics). Second, relevant prior knowledge has been found to facilitate encoding and retrieval processes by providing easily accessed retrieval structures in memory. Programs that have students connect their prior knowledge to new information help promote learning. Third, the use of learning strategies such as organization (how information is sorted and arranged in long-term memory), inferencing (making connections between concepts), and elaboration (increasing how meaningful information is by connecting new information to ideas already known) have been found to improve learning. Thus, including the development of information processing–based learning strategies in instructional programs will foster students’ development.

Recent research involving information processing theory includes the use of brain imaging technology and techniques from neuroscience research to further explore the human memory, the ways the brain processes a wide range of information and types of information, and how our extensive use of technology and multimedia influences how we process information.

Claudia A. Gentile

See also Cognitive Development, Theory of; Triarchic Theory of Intelligence

Further Readings

Fleck, B. K. B., Beckman, L. M., Sterns, J. L., & Hussey, H. D. (2014). YouTube in the classroom: Helpful tips and student perceptions. The Journal of Effective Teaching, 14(3), 21–37.

Piaget, J., & Inhelder, B. (1958). The growth of logical thinking from childhood to adolescence. New York, NY: Basic Books.

Shannon, C., & Weaver, W. (1963). The mathematical theory of communication. Urbana: University of Illinois Press.

Sternberg, R. J., & Sternberg, K. (2016). Cognitive psychology (7th ed.). Belmont, CA: Wadsworth.

Sarah Parsons Sarah Parsons Parsons, Sarah

Informed Consent

Informed consent

821

823

Informed Consent

Informed consent is a formal agreement made by individuals to participate in research, having been fully advised of the potential benefits, risks, and the procedures or activities of research participation. In educational research, informed consent is almost always sought from potential participants before collecting any data unless there are specific reasons for not disclosing the details of a research project up front. Informed consent is considered an essential aspect of good research ethics practice and is mandated by universities, research funders, and organizations in many countries. This entry describes the origins of the term *informed consent*, how it is applied in educational research, and some of the debates that exist about whether and how informed consent can be achieved effectively.

Core Principles and Practice of Informed Consent

Informed consent is underpinned by three core principles: (1) respect for persons, (2) beneficence, and (3) justice. These principles are known as the Belmont principles, after the 1979 report from the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, which was convened in the Belmont Conference Center in Maryland in the United States.

The principle of respect for persons states that human research participants should be treated as autonomous agents, be fully informed about the nature of the research they are asked to take part in, and that individuals with diminished autonomy should be protected. At the core of this principle is a moral judgment about the proposed research activities based on the question, “would you be happy to be treated in this way?” The principle of beneficence focuses on the consequences of research participation and asks researchers to ensure that the benefits of participation outweigh any harm. The principle of justice states that the selection of research participants should be fair and that those who are asked to bear the burden should also benefit.

In education research, these principles are typically translated into practice through written project information sheets that are given to potential research participants at the beginning of a research study. Project information sheets are designed to inform participants about key information including why the project is being carried out; who is doing the project, and how it is funded; why individuals are specifically being approached for participation; what participation in the research will entail; any risks and benefits for those taking part and for others; anonymity and confidentiality of participation and how personal data will be protected; and contact details of key personnel.

Information sheets are often accompanied by a consent form that participants are asked to sign if they agree to participate. The consent form generally asks individuals to confirm that they have read the information sheet and had the opportunity to ask questions, agree to take part voluntarily and can withdraw at any time, and understand their rights to confidentiality and anonymity. Together, the information sheet and consent form tend to be the key communication tools through which informed consent is gained. Information sheets, consent forms, and the research project protocols that guide their content are usually reviewed by panels of relevant stakeholders before being sent to potential participants. Common names for these panels are institutional review boards or research ethics committees.

Historical Development

The Belmont principles underpin the ethical conduct of contemporary research with human participants and are based on two highly influential sets of guidelines: the Nuremberg Code in 1947 and the Declaration of Helsinki in 1964. The Nuremberg Code established for the first time the fundamental rights of human research participants and the responsibilities of researchers wishing to include humans in their research. It followed from the Nuremberg trials where perpetrators of Nazi abuses in World War II were convicted, and executed, for their experimentation on human beings. Among other key principles, the

Nuremberg Code established the importance of research participation being voluntary, with research participants able to withdraw from participation, and that benefits must outweigh risks. The Declaration of Helsinki, made by the World Medical Association (and updated periodically since 1964), supported these core principles and added that research should be subject to review by an independent committee and that the privacy and confidentiality of participants must be assured.

Debates About Informed Consent

At the time when the Belmont principles were developed, the primary concern was with biomedical research, where researchers planned specific medical studies to intervene or implement a specific procedure and evaluate the consequences. In social science research, including education, many different research designs are used that do not follow the traditional assumptions of biomedical research. This has led to critiques of the application of research ethics principles derived from biomedical research to social science research.

In qualitative interpretive research, it is often not possible to know at the outset how research participation may change or develop over time or to judge what the risks or benefits might be; some participants may also wish for their identities to be made public. Some social science researchers argue that the overall risk of harm in social science research is likely to be much lower than in biomedical research and so the same level of scrutiny is not needed.

Another debate includes whether and how “vulnerable” participants should provide informed consent to participate in research and if proxy respondents can consent on their behalf. The concept of vulnerability is contested but is typically applied to children, the elderly, and those who may have reduced capacity to consent (e.g., individuals with intellectual disabilities). Proxy respondents often include parents, carers, or teachers. Other researchers have questioned whether the provision of information in primarily written forms is the most accessible way of supporting the comprehension of all research participants.

A further issue relates to *opt in* versus *opt out* consent: *Opt in consent* requires formal, usually signed, agreement to take part before data collection begins; *opt out consent* assumes that consent has been given if a consent form is not completed and returned to the research team. There are concerns as to whether opt out consent meets the core principle of respect for persons because it is not possible to know whether someone has actively agreed to participate in the research.

Sarah Parsons

See also Assent; Belmont Report; Declaration of Helsinki; 45 CFR Part 46; Human Subjects Protections; Human Subjects Research, Definition of; Institutional Review Boards; Nuremberg Code; Qualitative Research Methods

Further Readings

Brooks, R., te Riele, K., & Maguire, M. (2014) Ethics and education research. London, UK: Sage.

Hammersley, M., & Traianou, A. (2012). Ethics and educational research. In British Educational Research Association on-line resource. Retrieved April 3, 2016, from https://www.bera.ac.uk/wp-content/uploads/2014/03/Ethics-and-Educational-Research.pdf

Macfarlane, B. (2009). Researching with integrity: The ethics of academic enquiry. Abingdon, UK: Routledge.

Morrow, V., & Richards, M. (1996). The ethics of social research with children: An overview. Children … Society, 10(2), 90–105.

Chi Yan Lam Chi Yan Lam Lam, Chi Yan

Inputs

Inputs

823

824

Inputs

The term *inputs* refers to the resources made available to a program, policy, or curriculum to enable its operation. More precisely, inputs provide the antecedent conditions from which some programmatic activities are to occur and, as a consequence, achieve some predetermined objectives. Put simply, inputs are what get invested to do the work.

Inputs are important to make explicit because they play a limiting function in the implementation program, policy, or curriculum. For instance, the reach of a program is dependent on its inputs, such as the funding allocated to the program, the size of the venue in which the program is delivered, or the availability of program staff with expertise in the area. Without sufficient input, the efficacy and/or the effectiveness of a program may suffer. Yet, the opposite is not necessarily true. Overinvesting in a program, policy, or curriculum does not necessarily yield greater or better outcomes, if processes are unable to take advantage of abundant inputs. Hence an accurate accounting of inputs is important to understanding the effects of a program, policy, or curriculum.

Forms of Inputs

Inputs can take multiple forms. Recognizing the different forms that inputs can take is important to their use.

Financial Inputs

Financial inputs are typically valued in monetary units and can be exchanged for other resources. Examples include funding allocation from a government, per-participant program funding, and charitable donations from foundations.

Physical Inputs

Physical inputs typically refer to the physical infrastructure and other tangible resources made available to a program. Examples include the venue in which a program operates and any specialized equipment necessary to administer the program.

Human Inputs

Human inputs typically refer to the labor and the expertise available to a program, policy, or curriculum. Examples include access to consultants, key staff, and program administrators who bring specialized knowledge and/or experience to allow for program operation. The capacity of available human inputs is also important to consider; the expertise may be there, but without a corresponding reduction in existing workload, the expert may not be able to contribute to a program, policy delivery, or curriculum optimally.

Legislative Inputs

Finally, legislative inputs typically refer to any enacted laws and/or other official policies that place either legal obligations or restrictions on the performance of particular action. Legislative inputs trigger the most immediate change in the types of input. They often provide the easiest justification for adopting a new program, policy, or curriculum.

Applications of Inputs

Inputs are generally articulated over the course of logic modeling of a program or policy. Making explicit inputs to an intervention fosters a common understanding among stakeholders regarding the resources available to an effort and which resources are critical to the operation and attainment of intended results. Accurate and precise accounting of inputs in a study or evaluation is crucial, particularly when effectiveness and accountability are of concern.

Emerging Development

In recent years, policy makers and social researchers are increasingly recognizing the complex nature of social systems and interventions. This has led to growing recognition of the dangers around oversimplifying systems, in general, and the limitations of conceptualizing interventions in linear ways, in particular. The notion that inputs precede activities and outcomes logically is starting to give way in some cases. In its place, policy makers and social researchers are beginning to accept that outcomes may feed back into a system or intervention in a recursive manner, serving as inputs.

Chi Yan Lam

See also Formative Evaluation; Logic Models; Program Evaluation; Summative Evaluation

Further Readings

Funnell, S. C., & Rogers, P. J. (2011). Purposeful program theory: Effective uses of theories of change and logic models. San Francisco, CA: Wiley.

Knowlton, L. W., & Phillips, C. C. (2013). The logic model guidebook: Better strategies for great results. Thousand Oaks, CA: Sage.

Michael F. Hock Michael F. Hock Hock, Michael F.

Institute of Education Sciences

Institute of education sciences

824

826

Institute of Education Sciences

Created as part of the Education Sciences Reform Act of 2002, the U.S. Department of Education’s Institute of Education Sciences (IES) was charged with providing scientific evidence that informs education policy and practice. In addition, IES was expected to share research information in ways that are useful and accessible to educators, parents, policy makers, researchers, and the general public.

The IES is divided into four research and statistics centers: (1) the National Center for Education Evaluation and Regional Assistance (NCEE), (2) the National Center for Education Research (NCER), (3) the National Center for Special Education Research (NCSER), and (4) the National Center for Education Statistics (NCES). Each of the centers has various programs that focus on specific areas related to the overall mission of IES. The remainder of this entry outlines the centers and programs in place within IES as of 2016.

The NCEE

conducts large-scale evaluations of federal education programs and policies. The evaluations are designed to address complex issues of national importance. For example, NCEE has evaluated the impact of alternative pathways to teacher preparation, teacher and leader evaluation systems, school improvement initiatives, and school choice programs. NCEE also provides resources to increase use of data and research in education decision making through a program called the What Works Clearinghouse (WWC). The WWC conducts independent and rigorous reviews of research on what works in education and determines what programs and interventions have been found to meet the WWC determines what programs and interventions have been found to meet the WWC standards without reservations, to meet the standards with reservations, or to not meet the standards. States and districts often require the WWC “stamp of approval” as they evaluate program adoption.

Another program within NCEE is the Regional Educational Laboratories. The Regional Educational Laboratories support all 50 states and 16 territories in the United States and offer opportunities to learn what works as well as providing coaching, training, and other support for research use. Finally, NCEE’s statewide longitudinal data system grants enable states to more efficiently track education outcomes and provide useful, timely information to decision makers.

The NCER

funds development and rigorous testing of new approaches for improving education outcomes for all students. NCER supports development of practical solutions for education with a five-tier goal structure that supports exploration, development and innovation, efficacy and replication, effectiveness, and measurement. NCER supports researchers who wish to learn what works for improving instruction, student behavior, teacher learning, and school and system organization. NCER also supports advancement of statistics and research through specialized training and development of methods and measures. In addition, NCER funds predoctoral and postdoctoral training programs, as well as database training and short courses on cutting-edge topics for working statisticians and researchers. This focus on work that researches new methods and measures ensures continued advances in the accuracy, usefulness, and cost-effectiveness of education data collections and research.

The NCSER

is similar to the NCER in goal structure and funding mechanisms but funds development and rigorous testing of new approaches for improving education outcomes for students with disabilities or at risk of having disabilities. NCSER funds development of solutions for education with the same five-tier goal structure as NCER: exploration, development and innovation, efficacy and replication, effectiveness, and measurement. Thus, NCSER supports development of practical solutions for education from the earliest design stages through pilot studies and rigorous testing at scale. With NCSER support, researchers are studying what works and under what conditions for improving instruction, student behavior, teacher learning, and school and system instruction, student behavior, teacher learning, and school and system organization designed to impact outcomes for students with disabilities or at risk of having disabilities.

The NCES

provides data that describe how well the United States is educating its students. NCES projects collect and analyze official statistics on the condition of education, including adult education and literacy; support international assessments; and carry out the National Assessment of Educational Progress. The National Assessment of Educational Progress is NCES’s primary assessment of what American elementary and secondary students know and can do in academic subjects. The National Assessment of Educational Progress assessment disaggregates data on the performance of subgroups of students, enabling educators to gauge how well these groups are performing in reading, mathematics, science, and writing skills in relation to the mean performance of students in general.

NCES also assesses the proficiency of adults in performing basic literacy and mathematical tasks through the National Assessments of Adult Literacy. In addition, NCES provides insight into the educational outcomes of the United States by comparing them with those of other countries. This is achieved through the International Activities Program at the NCES, which provides statistical information comparing the educational experiences and trends in other countries to those of the United States. This work is designed to provide comparable indicator data about the outcomes of educational systems and institutions in other nations.

The NCES has a mandate to report to Congress on the condition of education by June 1 of each year. The Condition of Education annual report summarizes important developments and trends in education using the latest available data. The 2016 report presents 43 key indicators on the status and condition of education grouped under four main areas: (1) population characteristics, (2) participation in education, (3) elementary and secondary education, and (4) postsecondary education.

The IES is led by a director nominated by the president and confirmed by the Senate for a 6-year term. A national board advises the director on the work of IES. The National Board for Education Sciences has up to 15 members who are nominated by the president and confirmed by the Senate. The main work of the

National Board for Education Sciences is to approve the research priorities of the IES, ensure that IES completes quality work, and review various activities of IES to ensure that the IES proposal review process reflects the IES's research priorities.

IES has spent over $950,000,000 to support the research activities of contractors and researchers since its inception in 2002. Most of that support has gone to efforts to support education and training, data study/analysis, education services, and technical assistance.

Michael F. Hock

See also National Assessment of Educational Progress; National Science Foundation; Office of Elementary and Secondary Education; U.S. Department of Education

Websites

Institute of Education Sciences: http://ies.ed.gov/

National Center for Education Evaluation and Regional Assistance: http://ies.ed.gov/ncee/

National Center for Education Research: http://ies.ed.gov/ncer/

National Center for Education Statistics: http://nces.ed.gov/

National Center for Special Education Research: http://ies.ed.gov/ncser/

What Works Clearinghouse: http://ies.ed.gov/ncee/wwc/

Dean R. Gerstein Dean R. Gerstein Gerstein, Dean R.

Institutional Review Boards Institutional review boards

826

828

Institutional Review Boards

This entry defines institutional review boards (IRBs), discusses their history, and explains their structure and operation. An IRB is a committee formed and authorized by an organization to decide the ethical acceptability of research involving human subjects (participants) conducted by the organization's employees or associates, including students. IRBs worldwide are based on ethical principles codified in a few influential documents authored since World War II by national and international agencies in reaction to abusive treatment of human research subjects by biomedical and behavioral researchers.

Government regulations firmly guide and in many circumstances mandate IRB activities, but most of the thousands of IRBs in the United States are local in scope and have somewhat individualized characteristics and procedures. In other countries, IRBs (usually called research ethics committees outside the United States) are often organized as regional units affiliated with medical agencies or schools, governed by national laws or regulations. The IRB movement or system is dynamic, with changes driven in part by critics and controversies.

History of Human Research Protection

The Doctors' Trial of 23 German physicians and administrators, conducted from 1946 to 1947 by a U.S. military tribunal, resulted in legal sanctions, including seven death sentences, for those convicted of carrying out deadly experiments inside the Nazi slave labor and extermination prisons. The tribunal ultimately issued the 10-point Nuremberg Code, a code of ethics for human subjects research. The points of the code, including voluntary and well-informed consent, minimization of risk, proportionality of risks to benefits, and provisions for ending experiments, have been adopted in subsequent official ethical statements, including the 1964 Declaration of Helsinki of the World Medical Association and the 1978 Belmont report of the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

The inadequacy for practical purposes of broad ethical statements, if not accompanied by formal arrangements for monitoring and enforcing compliance, became clear with the publication of British anesthesiologist Henry Beecher’s 1966 article in the *New England Journal of Medicine*. Beecher’s work and subsequent documentary exposés brought to light numerous studies in many countries, including the United States, that grossly violated these ethical codes, before and after the Nuremberg and Helsinki reports. Across the globe, public and private organizations involved in medical and behavioral research shortly began formalizing structures, processes, and rules for ethical oversight of individual research investigations.

Since 1991, a policy known as the Common Rule has guided protections for human subjects in federally funded research in the United States. The Common Rule is part of Title 45 Part 46 of the Code of Federal Regulations and is mirrored in other Code of Federal Regulations titles. It mandates that institutions funded by any of 18 federal agencies must assure that ethical principles are followed in human subjects research, and as part of that assurance, institutions must set up IRBs to formally review each funded study in advance and in many cases to monitor during and after study implementation.

IRB Structure and Operations

Most U.S. colleges and universities, medical centers, school districts, federal and state human services agencies that have research programs, and organizations that regularly conduct human subjects research have developed their own IRBs (often called human research protection committees or the like) or have arranged with other organizations, including commercial “central IRBs,” to vet their projects. The Common Rule and associated regulations cover membership, procedures, decision rules, and record keeping of IRBs, which are required to register with the federal Office for Human Research Protections and affirm that they apply the regulations to all federally funded research in their institutions and, at their option, to all other human subjects research conducted therein.

Most research organizations have one IRB, but the largest research universities have as many as 10. Some IRBs' members are divided for review and voting purposes into distinct, separately registered panels. An individual IRB must have a minimum of five members, with at least one nonscientist, one scientist, and one member not affiliated with the host institution. Most IRBs are considered either biomedical or social–behavioral–educational, depending on the main type of research covered and the corresponding expertise of the members.

Prior to beginning a study, researchers must describe their project and its provisions for protecting human subjects in a formal application for IRB review, covering such issues as selecting the subjects; obtaining voluntary informed consent to participate in research; how researchers will offer treatment, intervention, or interaction with subjects in the research process; the risks of adverse physical effects, mental effects, discomfort, or social detriment; the potential benefits for the subjects, society, or scientific knowledge; conditions for early termination or withdrawal; confidentiality or anonymity; maintenance and sharing of research data; contact information for researcher and IRB; and if applicable, sponsorship/funding of the study, temporary deception, material compensation, and alternatives to research participation.

Based on the formal application and subsequent communications, IRBs or their designees under the Common Rule may classify research as exempt, expedited, or "full board." Exempt means minimal in risk and not requiring further review after initial vetting. The determination to grant an exemption may be made, per local IRB policy, by an IRB staffer, member, or designated, suitably trained individual other than the researcher. An application that qualifies for expedited review may be reviewed and approved by the IRB chair or one or more members designated by the chair without a discussion by the full convened board.

Studies that do not meet the criteria for exemption or expedited review, and those that are not approved under an expedited review, cannot go forward unless they are discussed and approved or disapproved by majority vote of the IRB, after discussion at a convened meeting that satisfies quorum requirements. Approval under the Common Rule by a convened board is for no more than 12 months and may be amended or renewed upon reapplication, usually a much lighter process. A typical IRB reviews hundreds of new applications annually, plus renewals, amendments, final reports, and reports on adverse results or other problems.

The shape of IRB activity has changed over time, partly in response to criticism and technological change. Behavioral scientists have criticized the IRB and technological change. Behavioral scientists have criticized the IRB regulatory regime, designed especially for experiments with drugs, toxic exposures, and medical devices, as a poor fit for behavioral research methods. The emphasis on written consent forms has been said to favor protecting institutions against litigation rather than assuring participant comprehension. A primary federal enforcement tactic, namely to disqualify an IRB and place all of an institution's human subjects research on hold, is considered an unwieldy sanction. Multiple IRBs reviewing virtually identical protocols have been observed mandating a variety of different accommodations.

Dean R. Gerstein

See also Belmont Report; Compliance; 45 CFR Part 46; Human Subjects Protections; Human Subjects Research, Definition of

Further Readings

Beecher, H. K. (1966). Ethics and clinical research. New England Journal of Medicine, 274(24), 1354–1360. doi:10.1056/NEJM196606162742405

Department of Homeland Security and other agencies. (2017, January 19). Federal Policy for the Protection of Human Subjects; Final Rule. Federal Register, 82(12), 7149–7274. Retrieved March 2, 2017, from https://www.gpo.gov/fdsys/pkg/FR-2017-01-19/html/2017-01058.htm

National Research Council. (2014). Proposed revisions to the Common Rule for the protection of human subjects in the behavioral and social sciences. Washington, DC: National Academies Press.

Office for Human Research Protections. (2016). International compilation of human research standards. U.S. Department of Health and Human Services. Retrieved May 28, 2016, from http://www.hhs.gov/ohrp/sites/default/files/internationalcomp2016%20.pdf

Schrag, Z. M. (2010). Ethical imperialism: Institutional review boards and the social sciences, 1965–2009. Baltimore, MD: Johns Hopkins.

Linda S. Behar-Horenstein Linda S. Behar-Horenstein Behar-Horenstein, Linda S.

Instructional Objectives

Instructional objectives

828

831

Instructional Objectives

Instructional objectives, also referred to as objectives or student learning outcomes, are statements that indicate the behavioral changes in knowledge, skills, or attitudes sought from students as a result of instruction or teaching activity. Instructional objectives indicate what students are expected to be able to do that they could not have done prior to teaching exercises. Written with the intent of learning as the outcome, teaching is the process used to accomplish that endpoint. In this regard, instructional objectives are outcome based and learner centered, focusing on the critical information, skill, or attitudinal outcomes sought.

Well-written instructional objectives highlight what learning should result by the end of a class session or another given period of time, such as a semester-length course. It is also recommended that each instructional objective use only one active verb and that it refer to an action that can be measured or observed. The active verb can be related to or linked with the type of assessment used in the course. The remainder of this entry discusses how instructional objectives are classified and their advantages for instructors and students.

Instructional objectives are classified into one of three domains: cognitive, psychomotor, or affective. The cognitive domain refers to intellectual or thinking skills. The psychomotor domain refers to learned procedural or physical skills pertaining to the performance of an action. The affective domain encompasses emotions, empathy, attitudes, and values. Each domain is also guided by a particular taxonomy that specifies levels, key words, and related questions that can be used to develop the instructional objectives.

can be used to develop the instructional objectives.

The cognitive domain consists of six levels: remember, understand, apply, analyze, evaluate, and create. Active verbs have been linked to levels of Bloom’s revised taxonomy to specify the categories of intellectual outcomes. Bloom’s taxonomy provides a set of key words and types of questions that may be asked to aid in the development of lower- (remembering, understanding, and applying) and higher-level (analyzing, evaluating, and creating) thinking skills. Verbs such as list, state, explain, relate, compare, contrast, differentiate, illustrate, analyze, categorize, classify, formulate, imagine, decide, defend, determine, evaluate, and judge are among many that can be used to denote lower and higher level thinking skills. Also, these verbs can be measured or observed, unlike verbs such as understand, be familiar with, know, learn, or appreciate. When teaching takes place at higher levels of learning, lower order behaviors such as remembering, understanding, and applying are subsumed within instruction.

The psychomotor domain consists of seven levels: perception, set, guided response, mechanism, complex overt response, adaptation, and origination. The affective domain consists of five levels: receiving, responding, valuing, organization, and internalizing values.

Cognitive Domain Objectives

Behavioral or cognitive changes are identified by the acronym ABCs, with A referring to the level of achievement, B to the behavior sought, and C to the conditions under which the behavior is to be observed. The following example describes the ABCs that could be identified when teaching a young child how to read. In this example, when presented with a card set of 25 single syllable sight words, the objective is that the student will correctly identify 80% of these words, or 20 words.

> The level of achievement specifies the degree of change sought that the student correctly identifies 80% or 20 single syllable sight words.
> The behavior refers to the skill that the student is expected to illustrate, correctly identifying sight words.
> The conditions under which behaviors will be observed refers to being presented with a card set of 25 single syllable sight words.

Instructional objectives are used by instructors to communicate expectations that they hold for students and changes in behavior that are sought as a result of teaching–learning interactions.

Drawing upon teaching activities in the professions, each level of the cognitive domain of Bloom’s revised taxonomy is illustrated in the remainder of this section using representative dental learning activities and sample verbs.

Level 1 (remembering)

While using a list of 10 options, students are asked to match 5 items that are associated with a periodontal pocket. This exercise requires that students know how to define, distinguish, draw, find, match, read, record, acquire, label, and list.

Level 2 (understanding)

Students in an introductory endodontics course are given a quiz on isolation. For the first question, they are asked to state 4–5 reasons that a rubber dam isolation is essential during endodontic procedures. This exercise requires that students differentiate, fill in, find, group, outline, predict, represent, trace, compare, demonstrate, and describe.

Level 3 (applying)

After completing textbook readings about the basics of periodontology, students are asked to explain the progression of periodontal disease from the perspective of pathogenesis. Students may be asked to convert, demonstrate, differentiate between, examine, experiment, prepare, produce, record, discover, discuss, or explain.

Level 4 (analyzing)

Students are presented with the following scenario: A 32-year-old White male arrives at your office and presents with pain and swelling over the upper right eye tooth for the past 3 days. His medical history is remarkable for gastroesophageal reflux disease for which he takes omeprazole (Prilosec) daily and he is allergic to penicillin. Your exam reveals signs of periodontitis and tooth decay. What are your concerns? How would you treat and prescribe? Students will need to be able to determine, discriminate, form, generalize, categorize, illustrate, select, survey, take apart, transform, and classify.

Level 5 (evaluating)

Students learn that the same dental treatment plan has been developed for two patients aged 18 months and 10 years old who have cleft palate. Neither patient has been previously seen by health professionals or treated for this condition until now. Students are given a complete summary of the dental, medical, social, and psychological health of each child and asked to critique the soundness of each treatment plan using authoritative and credible sources. Students will need to critique, defend, interpret, judge, measure, test, select, argue, award, or verify to respond correctly to this activity.

Level 6 (creating)

Students are presented with the following scenario. A 62-year-old prosthodontics patient has two fixed mandible bridges that have deteriorated over the last 2 years due to poor hygiene control and lack of brushing. These bridges now need replacement. You are a newly graduated dentist in the practice of two senior partners. The senior members suggest taking impressions and replacing the fixed bridges with new ones but have come to you and asked for your earnest opinion. You have read the recent literature on dental implants and would like to offer the patient this option; however, the senior partners are not really familiar with information about implants. You also recognize that implants are more appropriate to the patient’s needs and that over time, they represent a cost savings especially if the patient needs to have his bridges replaced within 2 years. Develop a plan for responding to the senior partners, in which you will provide an evidence-based rationale for suggesting the use of dental implants. The student will need to synthesize, organize, deduce, plan, present, arrange, blend, create, devise, rearrange, or rewrite information in response to this dilemma.

Psychomotor Domain Objectives

The dental learning activities described earlier can also be used to illustrate objectives in the psychomotor domain. For example, for Level 1—perception,—the dental student adjusts the height of the dental chair to ensure appropriate reach into the patient’s oral cavity. For Level 6 of the psychomotor domain—adaptation,—a new dentist demonstrates adaptable proficiency by modifying his motor skills to fit a new dental practice.

Affective Domain Objectives

For Level 1 (receiving)

the dental student demonstrates listening attentively to a patient’s description of her oral health problem and shows appropriate sensitivity to the patient’s presenting problem.

For Level 2 (responding)

the dental student listens to a patient’s description of her oral health problem and demonstrates attentiveness by asking questions aimed at seeking clarification and also by restating what the patient has stated.

For Level 3 (valuing)

the dental student demonstrates sensitivity and awareness of cultural differences, thus valuing diversity.

For Level 4 (organization)

the dental student recognizes the need for balance between patient autonomy and responsible practitioner behavior and understands the role of systematic planning in diagnosis and treatment planning.

For Level 5 (internalizing values)

a first-year dental student assigned to work with an interprofessional health-care student team demonstrates concern with personal, social, and emotional adjustment; displays self-reliance in working independently; and cooperates in group activities, thus illustrating a capacity to engage in teamwork.

Advantages of Instructional Objectives

Instructional objectives have distinct advantages for both instructors and students. Instructional objectives present a clear picture of the intended outcome as a result of teaching–learning interactions. They provide communication about the content and intent of a teaching activity or the curriculum. For students, instructional objectives communicate the target behavior that the instructor wants students to show as a measure of successful attainment of behavioral change, hence acquisition of new knowledge, skill attainment, values orientation, or professional attributes. Similarly, instructional objectives help focus student attention on what they need to study and master promoting their concentration on what is essential. When measured along a continuum, instructional objectives let students know the degree of success they have attained and the additional behavioral changes or new learning, skill attainment, or attitude formation that are still are needed.

When used wisely, instructional objectives can be used to develop corresponding measures of assessment to test the sustainability of behavioral changes, skill development, or performance actions over time. Instructional objectives provide a means for faculty to communicate with one another across courses, programs, or institutions. In addition, instructional objectives give purpose and focus to teaching in ways that can ensure alignment between the selection and organization of content, learning activities, assessments, and outcomes. Determining instructional objectives prior to instruction can assist instructors with selecting the material that is likely to be of greatest value to students.

Linda S. Behar-Horenstein

See also Bloom's Taxonomy; Classroom Assessment; Common Core State Standards; Curriculum Mapping; Curriculum-Based Assessment; Formative Assessment; Goals and Objectives; Response to Intervention

Further Readings

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). Taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (2nd ed.). Boston, MA: Allyn … Bacon.

Webb, E. M., Naeger, D. M., Fulton, T. B., & Straus, C. M. (2013). Learning objectives in radiology education: Why you need them and how to write them. Academic Radiology, 20(3), 358–363.

Lee Teitel Lee Teitel Teitel, Lee

Instructional Rounds

Instructional rounds

831

832

Instructional Rounds

Instructional rounds is a school and district improvement process, loosely based on the medical rounds model used by doctors, that brings educators together to look at classroom instruction in a focused, systematic, purposeful, and collective way. These observations are not intended to provide supervision or evaluation for specific teachers but instead to look closely at what is happening in classrooms and to work together systematically to make improvements. In this way, rounds visit is different from traditional outside-in or top-down administrative “walk-throughs.” They are generally more collegial and participatory, with host schools and districts playing important roles in every step of the improvement cycle.

This starts well before the actual visit, when educators who will be hosting the visit identify a stuck point or “problem of practice” on which they wish to get help. The more that the problem is tied to ongoing school or district improvement efforts, and the more engaged the staff is in the development of the problem of practice, the greater the likelihood that the rounds visit will contribute to school improvement.

On the day of the rounds visit, either observers will come from peer schools in a network or colleagues conducting school-based rounds will assemble to observe one another. After getting oriented to the school and the problem of practice, observers divide into teams of four or five and spend about 20 minutes per classroom in three to five classrooms, paying attention to what students and teachers are actually doing, and what kind of content they are working on. They write detailed observational notes related to the problem of practice, focusing on what they actually saw—not their reactions, judgments, or inferences.

what they actually saw—not their reactions, judgments, or inferences.

After these observations, the observers share data from their notes that they think would be particularly useful in focusing on the problem that the school has identified, and together, they look for patterns within and across the classrooms they saw. Host teams choose which of these patterns seem “high leverage”—meaning that if they could make good progress in addressing that pattern, they would anticipate making significant progress on their original stuck point. Observers and host school staff analyze the high leverage patterns, figuring out what might be their root causes, and together, develop specific suggestions for the learning that can be done at the school and, if appropriate, at the district level as well.

A key part of the improvement cycle takes place after the visit. Teachers and administrators at the host school and district make sense of the observations, patterns, and suggestions that emerged from the visit day. They make plans to translate the learning that they have engaged in that day into action in ways that connect with and support the ongoing improvement efforts at the school and in the district. They also figure out how to engage all members of the faculty in these improvement efforts—not just the handful who were on the visit.

Although the rounds’ process is not a silver bullet that will single-handedly lead to better test scores or increased learning for students, it can be a powerful accelerant of school and district improvement efforts, providing tools for helping districts and schools improve teaching and learning at scale.

Lee Teitel

See also Teacher Evaluation

Further Readings

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). Instructional rounds in education: A network approach to improving teaching and learning. Cambridge, MA: Harvard Education Press.

Teitel, L. (2013). School-based instructional rounds: Improving teaching and learning across classrooms. Cambridge, MA: Harvard Education Press.

Teitel, L. (2015). Instructional rounds. In L. Easton (Ed.), Powerful designs for professional learning, learning forward. National Staff Development Council.

Neal Kingston Neal Kingston Kingston, Neal

Instructional Sensitivity

Instructional sensitivity

832

834

Instructional Sensitivity

A test item is deemed instructionally sensitive if, when controlling for other factors, students who receive high-quality instruction on the content of the item do better than students who have not received high-quality instruction. Although some specialized achievement tests, such as the National Assessment of Educational Progress, Trends in International Mathematics and Science Study, and Progress in International Reading Literacy Study, are designed to estimate the distribution of a state or nation’s scores rather than the scores of individuals, in general, academic achievement tests are designed to measure the knowledge of the individual students to whom the tests are administered. This entry discusses the increasing focus on instructional sensitivity as part of teacher evaluation systems that incorporate student test scores, data collection designs and data analysis approaches for determining instructional sensitivity, and research on instructional sensitivity.

In recent years, student test scores have been used as a measure of teacher effectiveness, in some cases as part of state accountability systems. Many have argued that this latter use requires an assumption that tests are sensitive to the instruction. That is, to attribute student success on achievement tests to teacher quality requires a belief that high-quality instruction reliably leads to higher student test scores than low-quality instruction and even more so when compared to test scores of students who did not receive in-school instruction at all. There is limited evidence for this claim.

Counterarguments have been made that the items most sensitive to instruction are likely to be those at low levels of cognitive complexity, for example, items measuring factual knowledge or low-level comprehension (per Bloom’s measuring factual knowledge or low-level comprehension (per Bloom's taxonomy). Items measuring higher levels of cognition, such as analysis, evaluation, and synthesis, are conjectured to be less instructionally sensitive due to their greater complexity. There is no empirical evidence to support this.

Although the term *instructional sensitivity* was first used in the 1970s, the concept that the quality of achievement test items should be judged by their ability to reflect improvement in student achievement following instruction stems from the origins of objective student achievement testing circa 1920. At that time, item quality was typically judged by comparing the item scores of students in the grade in which related curriculum was taught to the scores of students in the previous grade on the same item. That is, item quality was defined by the increase in the percentage of students who answered correctly in the grade at which instruction on the topic occurred to the percentage of students in the previous grade who responded correctly.

Data Collection Designs for Determining Instructional Sensitivity

Methods of detecting the instructional sensitivity of test items can be divided into four broad data collection designs based on (1) item data from two representative groups, one that was exposed to the content and one that was not, (2) pretest–posttest administration to the same group, (3) item data from a single group where some had been exposed to the content and some had not, and (4) expert judgment.

As mentioned earlier, the use of item data from two representative groups has been practiced since the 1920s, before the use of the term *instructional sensitivity* or the use of test scores for evaluating educator quality. In such studies, item data are collected from all or a random sample of students in the grade for which the item is intended as well as in a random sample of students in the previous grade. Because the groups are representative of entire grade levels, it is usually reasonable to assume that most of the variability in item difficulty is due to instruction. Typically, item sensitivity is measured as the difference in percent correct between the two groups. However, percent correct has certain undesirable statistical characteristics. For example, the standard error of percent correct is dependent on the value of *p*. If *p* values are analyzed using least squares regression, this violates the assumption of heteroscedasticity.

An alternative instructional sensitivity metric with better statistical properties is the difference between the probit transformed *p* values for the instructed and uninstructed groups. The probit transformation places item difficulty on a normal metric. The probit is the lower tail *z* value that corresponds to the percent correct. For example, for a *p* of .50, the *z* value would be 0, and for a *p* of .84, the *z* value would be 1.0.

The second data collection design, pretest–posttest administration to the same group, is not common because such data are not typically available for large-scale testing programs. In general, it is the most powerful research design for empirical instructional sensitivity analysis because the students serve as their own control. To the extent that the pretest immediately precedes instruction and the posttest immediately succeeds instruction, there is little or no opportunity for extraneous nonrandom factors to influence the results. Analysis approaches include differences between pre and post *p* values or probit-transformed *p* values.

The third data collection design uses a single administration but seeks additional information regarding the exposure of each student to instruction. Although commonly used, this is a weaker design because there is no basis for the argument that the groups are equivalent other than with regard to instruction. That is, there might be a correlation between having had instruction and other factors that might impact student performance, such as socioeconomic status, English language proficiency, or disability status. Known factors can be considered as part of an analysis plan, but unknown factors might also have an impact. Any such factors could not be controlled directly during item analysis.

The fourth category, judgmental approaches, relies on a panel of curriculum specialists and/or teachers’ judgments of items’ instructional sensitivity. Empirical approaches identify instructionally insensitive items based on student responses to test questions.

James Popham, the originator of the judgmental method, posed three questions that a group of expert educators should be asked about each item whose instructional sensitivity is being considered:

1. Would a student’s likelihood of responding correctly to this item be determined mostly by the socioeconomic status of the student’s family?
2. Would a student’s likelihood of responding correctly to this item be

determined mostly by the student’s innate verbal, quantitative, or spatial aptitudes?

3. Responsiveness to instruction. If a teacher has provided reasonably effective instruction related to what’s measured by this item, is it likely that a substantial majority of the teacher’s students will respond correctly to the item?

Items for which most experts answered “no” to the first two questions and “yes” to the third question would be deemed instructionally sensitive.

The major advantage of judgmental strategies is that they are less expensive and can be implemented before the items are used in a test. The major disadvantage is that no published studies have shown whether the results of judgmental approaches agree with the results of empirical approaches.

Data Analysis Approaches for Determining Instructional Sensitivity

Several analytical approaches can be used to identify items that are sensitive to instruction. In addition to the two approaches previously described (difference in percent correct or difference in probit transformed percent correct), which would be inappropriate for unmatched groups, two conditioning approaches are common: Mantel–Haenszel and logistic regression.

Conditioning methods adjust for differences in overall proficiency between the two groups that would interfere with determining whether test items are instructionally sensitive. The minimum conditioning variable is overall test score. Without conditioning on overall test score, if the students who received instruction were of overall higher proficiency, then the impact of instruction would be overestimated. If they were of lower overall proficiency, the impact of instruction would be underestimated.

Summary of Studies of Instructional Sensitivity

No research studies, regardless of the research methods used, have shown that most of a test’s items show a statistically significant level of instructional sensitivity. Some studies have been based on small sample sizes and thus may not have been powerful enough to detect instructional sensitivity. However, for the most part, results of large-sample size studies have been consistent with the small studies.

Only a handful of studies have explored why some test items are more likely to be sensitive to instruction than others. Given the small number of such studies, it is not surprising that there is no consistent, replicable information about what factors lead an item to be sensitive or insensitive to instruction.

Neal Kingston

See also Accountability; Achievement Tests; Bloom’s Taxonomy; Power; Race to the Top; Standardized Tests; Teacher Evaluation; Value-Added Models

Further Readings

Naumann, A., Hochweber, J., & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. Educational Assessment, 21(2), 89–101.

Polikoff, M. S. (2016). Evaluating the instructional sensitivity of four states’ student achievement tests. Educational Assessment, 21(2), 102–119.

Charles M. Reigeluth Charles M. Reigeluth Reigeluth, Charles M.

Minkyoung Kim Minkyoung Kim Kim, Minkyoung

Instructional Theory

Instructional theory

834

838

Instructional Theory

Instructional theory concerns anything that is done purposely to facilitate learning. This entry describes the nature of instructional theory, including its major components. It identifies the major kinds, paradigms, and domains of instructional theory. Finally, it describes a variety of research methods to improve instructional theories.

What Is Instructional Theory?

Instructional theory is a design theory rather than a descriptive theory because it is goal oriented or instrumental: Its purpose is to offer the best known methods of instruction to accomplish given goals under given conditions. This stands in contrast to learning theories, such as schema theory and information processing theory, whose purpose is to describe learning processes. They do not offer guidance about how to facilitate learning. Design theory is concerned with creating something, whereas descriptive theory is concerned with what already exists (typically cause–effect relationships or natural processes).

Instructional theory is also often confused with the instructional development process. Instructional theory provides an image of the instruction, whereas the instructional development process is concerned with the process of creating the instruction. This is similar to the distinction between an architectural blueprint and the process for constructing the building. Other areas not strictly within the scope of instructional theory are assessment theory, which offers guidance about how to evaluate student learning, and curriculum theory, which offers guidance about what should be learned.

Every instructional theory has two major components: instructional methods and the situations in which those methods are believed to be preferable to the alternatives.

Methods

Instructional methods can vary in the following ways.

Scope of a Method

This is the amount of instruction that a method encompasses. It is a continuum that spans from micro (for an individual skill or understanding) through meso (for a cluster of related skills and/or understandings) to macro (for an entire course or curriculum). Most instructional theories address only one or sometimes two levels.

Generality of a Method

This is the breadth of situations for which a method is recommended. It is a continuum that ranges from universal (or pervasive, common) to local (or narrow, restricted). Most instructional theories claim more generality for their methods than is warranted. It is important to look for situations in which each method is not preferable to the known alternatives.

Precision of a Method

This is the level of detail of the *description* of a method—again on a continuum. More precision can be added to a method by describing its *parts*, by describing alternative ways of doing the method (*kinds*), and/or by providing *criteria* for making a decision regarding the method. The methods in most instructional theories are described at a relatively imprecise level. It is helpful to understand that the more precisely a method is described, the less general (more situational) it is likely to be.

Power of a Method

This is the degree to which a method contributes to attain the learning goal for which it was selected. This continuum is based on the increase in probability that the goal will be achieved as a result of using the method, all else being equal. Most instructional theories focusing on the methods the theorist believes are the most powerful.

Consistency of a Method

This is the reliability with which a method contributes its power to attain the learning goal in the situations for which it is appropriate. This continuum describes how much variability there is in the effectiveness of the method. Most instructional theories provide little information about the consistency of their methods.

Situations

Like methods, the instructional situations can vary in several ways. They include the following.

Values

These are aspects of instruction that are considered important by an instructional theory. They are a matter of opinion rather than a matter that can be proven. The latter are principles of instruction, rather than values. The set of values underlying an instructional theory represent an educational philosophy. It is important to make sure that the values of a theory align with the values of the stakeholders of the instruction (instructors, learners, and their institution), so each instructional theory should explicitly state the values upon which it was created. Kinds of values include the following:

Values About Learning Goals

These are statements about which learning outcomes are valued (philosophically). These are contrasted with learning outcomes that are identified empirically through a needs analysis.

Values About Priorities

These are statements about priorities that should be used to judge the success of

These are statements about priorities that should be used to judge the success of the instruction. They concern the relative importance of the effectiveness, efficiency, and appeal of the instruction.

Values About Methods

These are statements about the instructional methods that are valued from a philosophical point of view. They are contrasted with methods that have been empirically proven to be successful.

Values About Power

These are statements about who is given power to make decisions about goals, priorities, and methods.

All these kinds of values tend to vary depending on the situation.

Conditions

In addition to values, there are other factors that can influence the selection of instructional methods. They include the following:

Content

This is defined broadly as the nature of what is to be learned, which includes such things as metacognitive skills, emotional and social development, and values.

Learner

This is the nature of the learner, including prior knowledge, learning styles, learning strategies, motivations, interests, and more.

Learning Environment

This is the nature of the environment in which learning will occur, which includes human resources, material resources, organizational arrangements, and more.

Instructional Development Constraints

Instructional Development Constraints

These are the resources available for creating and implementing the instruction, including money, calendar time, and person hours.

Major Instructional Theories

There are many instructional theories, and they can be categorized according to kinds, paradigms, and domains.

Kinds of Instructional Design Theories

There are several major kinds of design theories related to instruction, including:

Instructional Event Design Theory

This theory addresses what the instruction should be like. This is what most people think of first when the term *instructional theory* is used. This theory deals with instruction.

Instructional Analysis Design Theory

This theory addresses the process of gathering information for making decisions about what the instruction should be like. This information includes information about the learners, what is to be learned, and constraints for the instruction. This theory involves analysis.

Instructional Planning Design Theory

This theory addresses the process of creating the plans for the instruction. This theory deals with design, but *design* is often used to refer to all these kinds of instructional theory collectively—the entire instructional development process.

Instructional Building Design Theory

This theory addresses the process of creating the instructional resources. This theory involves development, but *developmen*t is often used to refer to all these kinds of instructional theory collectively—the entire instructional development process.

Instructional Implementation Design Theory

This theory addresses the process of implementing the instruction, including instructor training, equipment procurement and installation, and even organizational change. This theory deals with implementation.

Instructional Evaluation Design Theory

This theory addresses processes for both formative and summative evaluation of the instruction (not the learner). This theory involves evaluation.

All but the first of these kinds of design theories are parts of what are often called *instructional design* (or *development*) process. Because all these kinds of design theories are about aspects of instruction, they are all instructional theories, though they are not what typically comes to mind when that term is used. Nevertheless, they are all important to instructional practice and research. In fact, it is important to understand that useful guidance for practitioners must integrate *all* of them.

Paradigms of Instructional Design Theories

There are also several paradigms of instructional theories, each fundamentally different from the others. The paradigms are based on Alvin Toffler's description of three waves of societal evolution, each one pushing aside the previous type of society.

The Agrarian Age Paradigm

During the Agrarian Age, the predominant paradigm was based on tutoring and apprenticeships; this paradigm is still used in some settings today. It typically entails one-on-one instruction and activity-based learning. Instructional methods include doing (with guidance and feedback) and showing (with explanations), in that order of frequency. Instructional theories in this paradigm date back centuries (if not millennia) before Socrates.

The Industrial Age Paradigm

During the Industrial Age, the predominant paradigm was based on “batch processing” with lecture, time-based student progress, and norm-referenced assessment. It typically entails teacher-centered, one-to-many, standardized instruction with no peer collaboration. Instructional methods are mostly telling (in person or through texts), with some showing (demonstrations) and doing (for largely inauthentic tasks). Instructional theories in this paradigm date back to the early 1900s.

The Information Age Paradigm

This paradigm is based on active, collaborative, personalized, competency-based, and self-directed learning that takes place in many schools today (and is how students have learned in Montessori schools for over a century). It typically entails task-based instruction (including project-and problem-based learning) with competency-based student progress and assessment. Instructional methods typically include interdisciplinary, authentic, collaborative projects with just-in-time tutorial support from the teacher, peers, or digital systems. The tutorials use doing (practice until a criterion for mastery is met), showing (demonstrations), and telling (explanations, typically in combination with demonstrations). Instructional theories in this paradigm are relatively new, developed largely within the past decade or two.

Domains of Instructional Design Theories

Finally, there are three domains in which instructional design theories have been developed: cognitive, psychomotor, and affective.

Cognitive Domain

Instructional theories in this domain focus on methods to help learners acquire mental skills and knowledge. Although taxonomies have been developed for different purposes by Benjamin Bloom, Robert Gagné, and others, the major differences in instructional methods for the cognitive domain fall into three categories based on type of learning: memorization, understanding, and application.

Theories for Memorization

These address both recall and recognition. They include instructional methods

These address both recall and recognition. They include instructional methods derived from behavioral learning theories (practice with reinforcement/feedback, repetition, chunking, and prompting) and cognitive learning theories (presentations and mnemonics).

Theories for Understanding

These address the development of conceptual, causal, and process understanding. Instructional methods were derived primarily from schema theory. For conceptual understanding, the methods relate new concepts to a learner’s prior knowledge. Methods include context, compare and contrast, analysis, analogy, instantiation, and others. For causal and process understanding, the methods include generality (expository or discovery/confirmatory) and demonstration (observation or exploration/manipulation). For all kinds of understanding, practice applying the understanding in diverse situations is also helpful.

Theories for Application

These address skill development, including concept classification, the use of rules (procedures and principles), and the use of metacognitive skills. Methods of instruction focus on telling how to do it (generality), showing how to do it (demonstration or example), and doing it (practice) with feedback.

Psychomotor Domain

Instructional theories in this domain focus on methods to help learners acquire both reproductive and productive physical skills.

Theories for Reproductive Skills

These are physical movements, such as touch typing, that have little or no variation and are memorized, making them automatic. Their instructional methods are similar to those for memorization in the cognitive domain: practice with reinforcement/feedback, repetition, chunking, and prompting.

Theories for Productive Skills

These are physical movements that have moderate to great variation and require concentration, flexibility, and strategic thinking. Their instructional methods concentration, flexibility, and strategic thinking. Their instructional methods include (a) impart knowledge of what should be done, using experiential, discovery techniques; (b) demonstrate the skill and provide verbal cuing of the steps; (c) provide long, continuous practice sessions with feedback; and (d) develop flow, automatization, and generalization.

Affective Domain

Instructional theories in this domain address emotional, moral, social, spiritual, aesthetic, and motivational development. Each of these kinds of learning has multiple components, including knowledge, skills, and attitudes.

Research Methods to Improve Instructional Theories

In contrast to descriptive theory, which is concerned with validity and truthfulness, design theory is concerned with preferability and usefulness. Thus, it requires very different research methods. Design theory can be advanced by both research to *prove* (confirmatory research, which is most appropriate in the later stages of development of a design theory) and research to *improve* (exploratory or developmental research, which is most appropriate in the earlier stages).

Research to Prove

These research methods include both experimental and quasi-experimental designs.

Experimental Designs

These entail random assignment of students to groups that vary in the instructional methods used. This is typically achieved in controlled or laboratory environments.

Quasi-Experimental Designs

These entail nonrandom assignment to such groups, but typically use statistical methods to adjust for differences in the students from one group to another.

Research to Improve

These research methods focus on developing a new design theory or improving an existing design theory through implementation in authentic contexts. Methods include grounded theory development, design-based research, and formative research.

Grounded Theory Development

Pioneered by Barney Glaser and Anselm Strauss, this research method focuses on inductive processes for theory development, hence the term “grounded.” Much of the guidance concerns coding of data.

Design-Based Research

Unlike grounded theory development, this method is driven by theory and prior research. It requires collaboration between researchers and practitioners in the real-world settings and entails flexible adaptation to improve both theory and practice as the research iteratively unfolds.

Formative Research

This is a kind of developmental or design-based research intended to develop or improve a design theory using a case study approach and formative evaluation techniques. It can be used with designed cases, as with design-based research, or with naturalistic cases (either in vivo or ex post facto) that are within the scope of the design theory.

Charles M. Reigeluth and Minkyoung Kim

See also Active Learning; Constructivist Approach; Design-Based Research; Information Processing Theory; Learning Theories; Mastery Learning; Montessori Schools; Social Learning

Further Readings

Gagné, R. M. (1985). The conditions of learning and theory of instruction. New York, NY: Holt, Rinehart and Winston.

Reigeluth, C. M., & Carr-Chellman, A. A. (Eds.). (2009). Instructional-design theories and models: Building a common knowledge base (Vol. 3). New York, NY: Routledge.

Stufflebeam, D. L., & Coryn, C. L. S. (2014). Evaluation theory, models, and applications (2nd ed.). San Francisco, CA: Jossey-Bass.

Toffler, A. (1970). Future shock. New York, NY: Bantam Books.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), Cambridge, MA: Harvard University Press.

Hamish Coates Hamish Coates Coates, Hamish

Instrumentation

Instrumentation

838

839

Instrumentation

A large range of instruments are used in educational research and practice. Instrumentation refers to the nature and use of such instruments. Consideration of instrumentation is imperative to ensure that evidence has desirable psychometric properties. Measurement invariance is among the properties most important for good assessment and has received attention in the scientific literature. This entry provides a brief overview of educational instruments, discusses measurement invariance, and reviews examples and key research.

Instrumentation varies depending on the phenomena being assessed and the purpose of assessment. Questionnaires may be deployed in surveys to assess practical or subjective things. Schedules may be used to structure interviews that generate qualitative data. Tests may be used in exams to assess knowledge or skill. Scenarios may be used in performance assessments to probe skill. Instruments may be deployed verbally, in paper, online, in person, or in a mixed range of modalities. Education is expansive and eclectic, and measurement in education may also draw on instruments from other fields such as medicine, engineering, or finance.

Part of the broader notion of validity is measurement invariance, which pertains to whether instruments are measuring the same phenomenon across place and time. Instruments that have variable measurement properties across place or time are a clear threat to the validity of measurement and hence to the assessment more broadly.

By way of example, evaluation of a mathematics intervention may be confounded if the test used to assess student competence before and after the intervention is biased against females (variation across place) or is somehow easier for people postintervention irrespective of their mathematics competence (variation across time)—perhaps due to prior exposure to test format or topics. In this instance, the intervention will appear more effective than it really is because of a larger difference between preintervention and postintervention scores arising from instrument invariance and irrespective of actual competence. What has happened is a change in the instrument's measurement properties rather than a change in the participants.

A variety of validation procedures can be used to assure instrument invariance. In terms of validation design, multitrait/multimethod evaluation is the most comprehensive, though invariance can also be tested across people and time. Common psychometric procedures include item response modeling and a host of covariance analyses such as covariance modelling, structural equation modeling, reliability replication, and even exploratory factor analysis. With sufficient validation, it is feasible and even desirable to use different but psychometrically linked instruments across place and time within the same study. What is important is having known validation and calibration properties.

There is a tradition of research on instrumentation and invariance in particular. This work has focused on building methods ensuring that instruments measure the same constructs and sustain calibration across different episodes of application. Ensuring the constancy of instrument measurement properties—the "instrumentality"—is of fundamental importance to ensure internal and hence external and consequential assessment validity.

Hamish Coates

See also Analysis of Covariance; Exploratory Factor Analysis; Interviews; Measurement Invariance; Multitrait–Multimethod Matrix; Structural Equation Modeling; Surveys; Tests; Threats to Research Validity

Further Readings

Cook, T. D., Campbell, D. T., & Day, A. (1979). Quasi-experimentation: Design … analysis issues for field settings, Vol. 351. Boston, MA: Houghton Mifflin.

Shadish, W., Cook, T., & Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton

Mifflin.

Stanley, D., & Campbell, J. (1966). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin.

Lyman L. Dukes Lyman L. Dukes Dukes, Lyman L. III

L. Danielle Roberts-Dahm L. Danielle Roberts-Dahm Roberts-Dahm, L. Danielle

Intellectual Disability and Postsecondary Education Intellectual disability and postsecondary education

839

843

Intellectual Disability and Postsecondary Education

An intellectual disability (ID) is characterized by significant limitations in both intellectual functioning and adaptive behavior as expressed in conceptual, social, and practical adaptive skills originating prior to the age of 18. This definition and the term *ID* designates the population of people who were previously diagnosed as having mental retardation. That term has been replaced both colloquially and in legislation with the term ID.

Educational opportunities for people diagnosed with ID have changed over the past 40 years. Since the 1970s, students with ID have been provided life skills training, employment training, and, many times, have been educated in academic settings alongside peers without disabilities. These changes, in concert with advocacy efforts and amendments to legislation, resulted in many students with ID, and their families, having an interest in educational opportunities that extend beyond high school into postsecondary educational settings. This entry describes the changing nature of disability terminology, legislation involving this terminology and legislation focusing on postsecondary education opportunities for individuals with ID, and educational systems and expectations impacting students with ID. In addition, employment and independence are discussed throughout the entry, which concludes with a discussion of research on postsecondary education and individuals with ID.

The Changing Nature of Disability Terminology

The Changing Nature of Disability Terminology

Using terminology that is respectful of people with disabilities is generally intended to avoid long-held stereotypes. Both advocacy efforts and legislative mandates have aided in the shift to greater integration and independence for people with disabilities. Therefore, it is appropriate that society use language that reflects sensitivity for and respect of persons with impairments. People with disabilities and their advocates can serve as guides when learning to use terminology that portrays people with disabilities in a manner that is dignified.

It is recommended that individuals consider the following when speaking to, referring to, or writing about people with disabilities: avoid outdated disability terminology (e.g., mentally retarded); avoid the use of terminology that carries an emotional message such as “suffers from”; generally speaking, use person-first language such as “person with an ID”; and avoid making reference to disability unless it is relevant to the message being communicated.

Legislation and ID

Rosa’s Law, signed in 2010, legislated that the federal government eliminate the terms *mentally retarded* and *mental retardation* from federal education, labor, and health policy and replace them with ID. Moreover, references to mental retardation have been changed to ID. Additionally, references that are not person first have been altered to reflect the use of more current language. For example, *mentally retarded person* has been changed to *person with an ID*.

Rosa’s Law was named after Rosa Marcellino, a child with Down syndrome, whose mother advocated for a change in terminology after finding out that Rosa had been labeled retarded at school. Rosa’s mother collaborated with other parents and legislators to introduce a law to change the terminology used in her home state of Maryland, which influenced senators in Maryland and elsewhere, to propose legislation to change the language used in federal documentation as well.

The opportunity for students with ID to participate in postsecondary education alongside their peers is increasingly a reality in the United States due to provisions within the Higher Education Opportunity Act (HEOA) of 2008. The HEOA amendments were the first federal guidance provided relating to higher education services for students with ID and sought to remedy the high variability of services provided in postsecondary education for students with ID through defining comprehensive transition and postsecondary education programs for students with ID. Since 2008, there has been a substantial increase in the number of postsecondary programs for students with ID.

With respect to employment, the 2014 Workforce Innovation and Opportunity Act increases individuals with disabilities’ access to high-quality workforce services and preparation for competitive integrated employment. One focus of Workforce Innovation and Opportunity Act is to increase individuals with disabilities’ access to workforce services in order to foster competitive integrated employment, including preemployment transition services to high school youth.

Educational Options for Students With ID

Recent interest in postsecondary education for students with ID is due in part to parental expectations, the increase of students with ID in K–12 education, and a societal focus on postsecondary education as a desired outcome for all. There is an expectation that most, if not all, students graduate from high school fully prepared for college and careers. The College and Career Readiness initiative is also aligned with another national movement for systems change based on the philosophy that employment is an expectation for all people, including individuals with significant disabilities, called Employment First.

Students with ID have historically been excluded from postsecondary education. There is often an assumption that students with ID do not have the ability necessary to benefit from college attendance. As such, students with ID have experienced dismal postschool outcomes and, as a disability group, are the least likely to participate in postsecondary education.

Prior to 2008, there were college and university programs for this student population; however, they were few in number. In 2008, the HEOA authorized funding for a model demonstration program called Transition and Postsecondary Programs for Students with Intellectual Disabilities for the development and expansion of postsecondary education programs. One of the core beliefs of the Transition and Postsecondary Programs for Students with Intellectual Disabilities initiative is that students with ID will have improved employment and other life outcomes as a result of their college experience. Outcomes for students with ID participating in postsecondary education are continually being examined, and recent results have determined that paid employment coupled with opportunities to live independently while in school ensured the acquisition of life skills. Based upon preliminary evidence, it can be concluded that carefully selected program components and other transition services provided in an age-appropriate environment, such as an integrated college campus, results in improved adult outcomes.

The purposes of postsecondary programming for students with ID include employment, inclusion with same-age peers, independent living skills, and participation in college classes. These purposes are fundamentally different from those of many other students attending college, whose primary focus is often academic work. While an academic component is typically included in programming for students with ID, the programs are often nondegree or certificate based and usually do not lead to a college degree.

Some postsecondary institutions have a residential component for students with ID, such as residence halls, on-campus apartments, off-campus apartments, fraternity and/or sorority houses, and special sections of campus housing exclusively for students with ID. There are sometimes additional services offered for students with ID not typically provided to other students, such as independent living training, 24-hour staff support, and paid roommates.

Dual Enrollment Programs

College-based dual enrollment programs for students with ID are programs that enroll students in secondary education and postsecondary education simultaneously, which usually occurs for secondary students to use local education funds to support participation in postsecondary education. These programs provide opportunities to explore college while still receiving support from high school education staff. A great level of collaboration is required in order for a program to exist and flourish between two different systems: K–12 and higher education. To assist with the collaborative efforts of dual enrollment programs, it is helpful to clearly articulate the role of each partner. Such initiatives are becoming more common nationwide.

The Role of Schools and Colleges

Although programs differ depending upon the participating postsecondary institution, local educational agency, disability-related agencies, and community partners, most dual enrollment programs include these four aforementioned partners in some capacity. The local educational agency provides the students on-campus support staff such as a special education teacher and sometimes a paraprofessional. The special education teachers' role is different than the role of traditional high school teachers as they also serve as curriculum coordinators, assisting students in building the independence necessary to successfully complete the program.

There are other important members of the team as well. The local educational agency secondary transition specialist provides programmatic support, from finding students who are a good fit for the program, to allocating resources to the program such as job coaching and support from social workers. The postsecondary program provides access to college courses and campus activities and other amenities. Student disability support offices are willing to provide support to participating students on most campuses.

Transition Programs and Community-Based Vocational Education

Students with ID are increasingly being included in the general education classroom in K–12 settings; however, separate placements still exist and include part-time resource support or self-contained classrooms. Because students with ID may receive a free appropriate public education usually to age 22, they may participate in extended transition programs on a high school campus or in a community or workplace setting. For example, during high school and extended transition, students may participate in community-based instruction, which is instructed in the community in naturally occurring environments with the goal of gaining real-world experience. Similarly, students can participate in career vocational education and training through community-based vocational education, where education is provided in a typical community work setting instead of in a school environment.

Postsecondary Education and Employment Outcomes

Students with ID participating in postsecondary education have demonstrated improved employment outcomes. Historically, they have had higher rates of unemployment or underemployment and earn lower wages than those in other disability categories or people without disability. As a vital component in career development and expanding earning potential over a lifetime, higher education is essential whether one has a disability or not, and attending postsecondary education increases the potential for competitive employment being paid at minimum wage or above.

Research has found that students with disability who attend postsecondary education are more likely to be competitively employed and obtain higher earnings over time than those who do not attend. Although students with ID do not typically earn a college or university credential, those who were exposed to postsecondary education have been more likely to find jobs with higher wages compared to those without any postsecondary education experience. This finding is specific to students with ID, which reinforces the uniqueness of support needs that such students bring to postsecondary education.

Employment First, a multiagency approach to improve employment outcomes for individuals with disability, is one of the factors influencing the changing nature of education for this population. Employment First is a philosophy ensuring that employment in integrated settings within the community is an expectation and a priority for people with disabilities. This initiative also includes supported and customized employment as well as self-employment.

Community Participation and Independence

Postsecondary education is typically considered essential to achieving positive adult outcomes. It provides the possibility of developing successful employment skills and the opportunity to acquire a network of friends and acquaintances. Thus, college addresses not just student academic needs but also the opportunity to make friends and develop community contacts. It is often through this family and friend community network that employment opportunities develop. These are, in fact, some of the overarching goals of many postsecondary education programs for students with ID. Currently, there are limited data available regarding the independent living outcomes for students with ID participating in postsecondary programming; however, recent studies involving mentoring and building social capital have shown promising results.

Research on Postsecondary Education and Individuals With ID

People with ID are considered a marginalized population that has often been excluded from educational opportunity. With the passage of the HEOA of 2008, there has been an increase in postsecondary access for students with ID and research describing the types of programs developed, including student outcomes.

Employment and independent living are important outcomes of higher education for all students, including those with ID, and preliminary evidence indicates that postsecondary opportunities are having a profoundly positive impact upon the lives of people with ID and their communities. These benefits include increased levels of employment at higher wages, self-determination, self-sufficiency, and quality of life. Further, the voices of individuals with ID are being included as part of the dialogue today through inclusive methods and member checking and also by providing training on and measuring levels of self-determination and self-advocacy. The efforts of people with ID and the scholars examining their progress will continue to be monitored.

Lyman L. Dukes III and L. Danielle Roberts-Dahm

See also Americans with Disabilities Act; Developmental Disabilities; Inclusion; Individualized Education Program; Individuals With Disabilities Education Act; Least Restrictive Environment

Further Readings

American Association on Intellectual and Developmental Disabilities. (2013). Definition of intellectual disability. Retrieved from http://aaidd.org/intellectual-disability/definition#.VvSMkPkrK00

American Psychiatric Association. (2013). DSM-5 intellectual disability fact sheet. Retrieved from http://www.dsm5.org/psychiatrists/practice/dsm/educational-resources/dsm-5-fact-sheets

Center for Parent Information and Resources. (2015). Intellectual disability. Retrieved from http://www.parentcenterhub.org/repository/intellectual/

Downing, J. E. (2010). Academic instruction for students with moderate and severe intellectual disabilities in inclusive classrooms. Thousand Oaks, CA: Corwin.

Folk, E. D. R., Yamamoto, K. K., & Stodden, R. A. (2012). Implementing inclusion and collaborative teaming in a model program of postsecondary education for young adults with intellectual disabilities. Journal of Policy and Practice in Intellectual Disabilities, 9(4), 257–269.

Grigal, M., & Hart, D. (2010). Think college! Postsecondary education options for students with intellectual disabilities. Baltimore, MD: Brookes Publishing.

Grigal, M., Hart, D., & Weir, C. (2012). A survey of postsecondary education programs for students with intellectual disabilities in the United States. Journal of Policy and Practice in Intellectual Disabilities, 9(4), 223–233.

Thoma, C. A., Lakin, K. C., Carlson, D., Domzal, C., Austin, K., & Boyd, K. (2011). Participation in postsecondary education for students with intellectual disabilities: A review of the literature 2001–2010. Journal of Postsecondary Education and Disability, 24(3), 175–191.

Wehman, P., & Kregel, J. (2012). Functional curriculum for elementary and secondary students with special needs. Austin, TX: Pro-Ed.

Winzer, M. A. (1993). The history of special education: From institution to integration. Washington, DC: Gallaudet University Press.

John Mark Froiland John Mark Froiland Froiland, John Mark

Intelligence Quotient Intelligence quotient

843

845

Intelligence Quotient

Historically, an individual’s intelligence quotient (IQ) was calculated by taking the individual’s “mental age” divided by the chronological age and multiplied by 100. An individual’s IQ is now derived through advanced statistical analysis of how that individual performs on multiple aspects of intelligence tests, such as verbal comprehension, visual–spatial ability, working memory, fluid reasoning, and processing speed. Namely, psychologists discussing IQ are now usually referring to the overall intelligence scores but not to an actual IQ.

Knowing an individual’s Full Scale IQ (or estimate of individual’s overall intelligence) contributes to the evaluation of giftedness, intellectual disability, autism spectrum disorder, attention-deficit/hyperactivity disorder, traumatic brain injury, and learning disabilities. Furthermore, IQ is often found to be the strongest predictor of academic achievement, suggesting it should be considered in a comprehensive evaluation of contributors to achievement. IQ is also positively related to many other important factors that promote learning and success in school, such as social cognition, intrinsic motivation to learn, and executive functioning (e.g., attention and inhibition of negative responses). This entry further defines IQ and discusses its applications.

What is IQ?

IQ, often referred to in intelligence tests as Full Scale IQ, is thought to represent general intelligence (or *g*) and be comprised of multiple index scores (or factors), such as the following: visual–spatial (evaluating and integrating visual–spatial information), verbal comprehension (vocabulary, understanding of complex verbal information, and verbal reasoning), working memory (mentally holding and manipulating information in real time), fluid reasoning (determining how visual objects are inherently related and thinking through the application of intricate rules), and processing speed (rate of visual recognition and efficiency in decision making). Together, these types of factors combine to form one’s overall IQ.

Although knowing one’s overall IQ is valuable, it can also be helpful for individuals to understand their cognitive strengths and weaknesses. For instance, one person with an average IQ may be quite strong in verbal comprehension but have a below average processing speed. Another person with the same overall IQ might have above average working memory, yet struggle with integrating and applying complex visual information. Although these two students have equivalent IQs, they may have significantly different approaches to learning.

Some intelligence tests emphasize attention as a part of IQ, whereas many theorists and psychological scientists consider attention to be an executive function that is moderately positively related to intelligence but not part of the same psychological construct. The current use of the term *intelligence quotient* is an artifact of the past, in which IQ was determined by dividing individuals’ mental age by their chronological age. Although the norms from which IQ scores are derived take age into account, the IQ score is now derived in a much more complex statistical fashion than applying a relatively simple quotient. It is important to point out that IQ tests individually administered by psychologists have greater reliability and validity than group-administered IQ tests, which often involve multiple students taking an IQ test at once with relatively little interaction with the examiner.

Applications for IQ

The role of IQ testing in the schools has changed somewhat because the emphasis on determining a learning disability used to be the IQ-achievement discrepancy. More recently, the psychological and clinical assessment of learning disabilities has placed a greater emphasis on the cognitive processes that contribute to strengths and weaknesses in achievement as well as how students respond to science-based interventions. In response, a greater emphasis is often placed on index scores and specific cognitive strengths and weaknesses (e.g., processing speed vs. working memory) rather than overall IQ. Furthermore, intelligence tests have been developed that focus on how students solve problems, such as the Wechsler Intelligence Scale for Children–Fifth

Edition, Integrated, published in 2015. This test can help school and educational psychologists gain a deep understanding of how students approach academic and intellectual problems as well the conditions in which students perform better. In numerous studies, the strongest predictor of achievement is IQ; thus, schools that are attempting to elevate reading, science, and math achievement would be wise to include IQ within their models of achievement.

Often, the response to intervention movement has led to an emphasis on quick and repeatable academic measurements. However, there is great potential for intelligence testing to inform the development of strategic science-based interventions, especially because IQ is positively related to other important factors in school, such as the following: social cognition (the ability to process and reason about social and emotional situations), intrinsic motivation to learn (a love for learning or enjoyment of learning often accompanied by recognizing its beauty or purpose), and executive functioning (e.g., capacity to pay attention). Furthermore, knowing a child’s IQ can help educators to provide an optimal challenge in various learning environments based on the child’s level of cognitive development rather than boring the child with easy material or frustrating the child with material that is excessively difficult.

John Mark Froiland

See also Ability–Achievement Discrepancy; Cattell–Horn–Carroll Theory of Intelligence; *g* Theory of Intelligence; Kaufman-ABC Intelligence Test; Motivation; Raven’s Progressive Matrices; Response to Intervention; Social Cognitive Theory; Wechsler Intelligence Scales

Further Readings

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge, UK: Cambridge University Press.

Kaufman, A. S., Rairford, S. E., & Coalson, D. L. (2016). Intelligent testing with the WISC-V. Hoboken, NJ: John Wiley.

Wechsler, D. (2014). Wechsler Intelligence Scale for Children (5th ed.). San Antonio, TX: Pearson Clinical Assessment.

Tracy Paskiewicz Tracy Paskiewicz Paskiewicz, Tracy

Kathryn Doherty Kurtz Kathryn Doherty Kurtz Kurtz, Kathryn Doherty

Melissa Pearrow Melissa Pearrow Pearrow, Melissa

Intelligence Tests

Intelligence tests

845

848

Intelligence Tests

Intelligence testing is the complex process of measuring an individual’s ability to understand ideas and words, think in abstract terms, solve problems using different forms of reasoning, learn from feedback and experience, and process information in different modalities, such as visually or aurally. The intelligence quotient (IQ) is usually the result of this assessment process. Intelligence testing is also known as intellectual or cognitive assessment or IQ testing, with these terms being used interchangeably. This entry first describes intelligence and theories of intelligence before giving a brief history of intelligence testing in the United States. It then looks at why intelligence testing is used; psychometric issues related to measurement of intelligence; intelligence testing across the life span; strengths and weaknesses of, and misconceptions about, intelligence testing; and potential future directions for intelligence testing.

Cognition and intelligence are both associated with and refer to functions of the human brain. These functions reflect genetic endowment as well as environmental stimulation. Although there are both hereditary and environmental influences on intelligence, these are difficult to disentangle. It seems that both hereditary and environmental factors can influence *each other*. Therefore, it is not possible to conclude that an individual’s intelligence is influenced by genetics *or* the environment; it’s both.

Theories of Intelligence

Theories of Intelligence

Intelligence is generally thought of as a collection of attributes such as the ability to learn, adapt to the environment, and think abstractly. There are several different theories of the structure and concepts that make up one’s intelligence. Some theories emphasize that intelligence is a unitary concept; other theories propose that intelligence is multifaceted. The theories that emphasize a unitary concept refer to general intelligence as *g*. Theorists who advanced general intelligence are Charles Spearman, Philip Vernon, and John Carroll.

Multifaceted theories describe different aspects of intelligence, such as verbal reasoning ability, working memory, or visual–spatial thinking. Researchers advancing multifactor theories are Edward Thorndike, Louis Thurstone, J. P. Guilford, and Raymond Cattell and John Horn. Other theories are focused on information processing, such as the processes involved in taking in information, holding on to that information, and processing information for other purposes. Howard Gardner posed a theory of multiple intelligences that is composed of several independent competencies that interact to produce a diverse mixture of human talents.

A widely accepted contemporary theory of intelligence, the Cattell–Horn–Carroll theory, is based on a psychometric approach—that is, the structure or dimensions of intelligence are established through statistical procedures. This model describes a general factor (*g*) at the top of a hierarchy, several broad abilities in the middle, and narrow abilities at the bottom.

History of Intelligence Testing in the United States

Intellectual assessment has a long history of being both beneficial and controversial in terms of its impact on society, in the United States, and internationally. This section focuses on the history of intelligence tests in the United States. James McKeen Cattell, a student of Wilhelm Wundt, established a psychological laboratory at the University of Pennsylvania and published an article in 1891 where the author used the term *mental test*. The authors’ work shifted the discussion of the assessment of mental ability in the United States from its philosophical origins to one that strove to be grounded in empiricism.

The 1905 Binet-Simon Scale (developed by Alfred Binet, Victor Henri, and Theodore Simon in France) was introduced to the United States by Henry Goddard in 1908. The 1905 Binet-Simon Scale differed from earlier intelli

Goddard in 1908. The 1905 Binet-Simon Scale differed from earlier intelligence tests in its emphasis on mental processes instead of sensory functions. Goddard developed the 1908 Binet-Simon Scale by norming the 1905 Binet-Simon Scale on a sample of children from the United States. This particular intelligence test was used for a long period of time with its primary use being to identify individuals with intellectual disabilities.

Lewis Terman of Stanford University published a revision of the 1908 Binet-Simon Scale, naming it the Stanford-Binet in 1916. In this revision, Terman introduced the term *IQ*. This 1916 development of the Stanford-Binet marked a significant event in the field of intelligence testing. The 1919 Army Alpha and Army Beta test made intelligence testing more popular in the mainstream. The Wechsler Scales, which continue to be widely used today, began with the publication of the Wechsler-Bellevue Intelligence Scale, Form I in 1939. Researchers and practitioners continued to develop different assessments to measure mental ability during the decades following this 1939 development.

Uses of Intellectual Assessment

An individualized intelligence test follows standardized administration procedures in a one-on-one setting with a trained professional. Contemporary IQ tests generate scores for a variety of cognitive processes (i.e., language fluency, working memory, and three-dimensional thinking). Individual subtest scores are calculated to create an overall summary score, commonly termed the Full Scale IQ; these subtests tend to correlate with one another, even though the content seems dissimilar. The major purpose of intelligence testing is to identify learning strengths and needs, inform diagnosis of exceptionalities, and design interventions for children and adults. It is used in a variety of settings aimed to inform educational, medical, and mental health care including but not limited to schools and other educational settings, clinics, hospitals, community agencies, and offices of health-care professionals. In addition to being used for clinical purposes, intellectual assessment is often used for research purposes in university or agency settings.

In educational settings, intelligence testing is typically used as part of requirements for special education eligibility or to identify giftedness. The Individuals With Disabilities Education Act was originally enacted by Congress in 1975. The law was reauthorized in 2004, and regulations based on the 2004 amendments were published in September 2011. There are 13 disability categories specified under Individuals With Disabilities Education Act in which students can demonstrate needs to qualify for special education. Educational disability categories are based on the criteria outlined in the Individuals With Disabilities Education Act, with individual states establishing regulations for how these criteria are met. Most states require some formal intelligence testing in order to assess for intellectual disabilities, and many still require formal intelligence testing in assessments for specific learning disabilities.

Psychometric Issues Related to Intelligence Testing

In testing, the standards by which to judge appropriateness of an instrument include estimates of *reliability* and *validity*. A reliable test is one that shows consistency in measurement. The validity of a test refers to whether it measures what it is supposed to measure. These concepts, as well as other psychometric variables such as the assumptions of a normal curve, norming processes, errors in measurement, and potential sources of bias, help users gauge the strengths and weaknesses of different testing instruments. Even with the most sophisticated methods, it is important to remember that testing instruments are imperfect; there is no test that is completely free of measurement error. Users of intelligence tests must try to minimize sources of error by making sure the test is administered in a standardized manner and by striving for optimal testing conditions for the individual.

The types of scores typically derived from intellectual testing include standard scores, percentile ranks, and age-equivalent scores. These scores help a consumer comprehend an individual’s relative standing or comparison to the general population. Most standardized IQ tests utilize a common scale, with a standard score mean (average) of 100 and a standard deviation (a measure of how much scores vary from the mean) of 15. Thus, a person is said to have “average” intelligence if the person’s IQ score falls within the range of 85–115.

After about a century of research, intellectual assessment is one of the best studied procedures in the social sciences. The research shows that IQ tests are reliable and valid for predicting a number of important variables such as academic achievement in school as well as life outcomes such as occupational level and success.

Intellectual Assessment Across the Life Span

Intelligence tests are used in practice and research throughout the life span, beginning in infancy through late adulthood. The reasons for conducting intelligence testing in young children include to gauge developmental progress or to identify areas of need for intervention. In older adults, intelligence tests may be used to assess cognitive decline. Some of the most commonly used measures of intelligence are the Wechsler Scales, Stanford-Binet, Woodcock-Johnson Tests of Cognitive Abilities, Differential Abilities Scale, and the Kaufman instruments. To reflect changes in the demographics of the United States, intelligence tests are updated and renormed approximately every 10 years. In addition to tests that assess the construct of general intelligence, there are also brief/abbreviated intelligence tests and nonverbal intelligence tests that attempt to assess an individual’s cognitive ability using nonverbal administration and response formats.

Intelligence Tests and Culture

One of the main controversies of intelligence testing is whether these instruments are biased against individuals from minority backgrounds. Opponents of intelligence testing assert that culturally and linguistically diverse individuals (those who are from historically marginalized ethnic and racial minority groups and/or are nonnative English speakers) cannot be fairly assessed using traditional IQ tests. There is a significant body of research that addresses this topic. Some research beginning in the 1990s has focused on the impact of experience and background on IQ test performance. Researchers such as Samuel Ortiz have explored the validity of standardized intelligence tests for use with children from culturally and linguistically diverse backgrounds. This research suggests intelligence test data should be interpreted cautiously when working with culturally and linguistically diverse students because of their potentially different exposure to information included in cognitive tests such as vocabulary terms.

Sophisticated statistical and test construction methods are used in the development of intellectual assessment instruments. However, tests do contain items and content that reflect the mainstream culture of the United States. To the extent that an individual’s background experience is comparable to the mainstream in the United States, IQ tests are appropriate to measure that individual’s intelligence. When an individual’s culture and language background differs from that of the norm group for a test, test results are likely to have questionable validity. Even though the vast research on this topic shows IQ tests questionable validity. Even though the vast research on this topic shows IQ tests predict outcomes such as academic achievement equally well for diverse groups of individuals (e.g., different racial backgrounds, genders, age, and geographical region), tests measuring verbal abilities such as vocabulary and general conceptual knowledge should not be used to estimate the cognitive ability of individuals whose primary language is not English.

Where bias can come into play is in the *use* of intelligence test scores by practitioners. A test may be valid for a particular purpose but still result in biased decisions if the results are used improperly or if the results are not interpreted correctly. Therefore, the validity of intelligence tests lies in the decisions that are based on the results. As of 2016, the consensus in the field was to use multiple methods of assessment (e.g. interviews, observations, curriculum-based assessment, and rating scales) with all individuals to gain a broader sense of their functioning instead of solely relying on data gathered from intelligence tests.

Strengths and Weaknesses of, and Misconceptions About, Intelligence Tests

Intelligence testing is one of the most significant contributions to the field of psychology. Despite controversy over their use, psychologists and educators continue to rely on intelligence testing for many reasons. Intelligence tests predict important outcomes, such as academic achievement, occupational level, and economic success.

IQ tests provide a standardized way of comparing an individual with other individuals of the same age. IQ testing can help individuals understand processing strengths and weaknesses and, as mentioned, can assist psychologists in understanding individuals with disabilities. The critical function of IQ may be as a threshold; more specifically, below some very low point on the IQ distribution (<65 or 70), individuals may need additional supports to function as adult members of our society. However, IQ tests provide only a limited understanding of intelligence; many behaviors considered by our society to be “intelligent” behaviors are not tapped by traditional IQ tests. No existing test is capable of adequately measuring the ability to deal with all kinds of situations that require intelligent resolutions.

IQ tests can be used to classify or “label” children, possibly limiting their potential freedom to choose courses of study. The IQ can sometimes be misused as a measure of innate or inborn capability. Criticisms of IQ cite that intelligence is a multifaceted concept that cannot be easily “boiled down” to a single test score. IQs are of limited value in predicting nonacademic intellectual activities, such as social savvy, leadership potential, and resilience in the face of adversity. An IQ score cannot capture the complexity of the real-life situations involving the use of intelligence because intelligence tests sample only a limited number of behaviors.

Finally, there are many misunderstandings about the concept of intelligence, including that IQs are unchangeable and fixed. It is a misconception to believe that intelligence tests provide perfectly reliable and absolute scores, as IQ testing provides merely an estimate of a person’s ability under certain circumstances. Intelligence tests are constructed to reflect abilities valued by the mainstream culture of the United States; therefore, they contain culturally loaded content.

Future Directions

The increasing use of public health–oriented frameworks such as multitiered systems of supports, response to intervention, and positive behavioral interventions and supports in human service fields raises questions around what the role of intelligence tests will be in the future. Because intelligence testing has largely been used to identify individuals with disabilities in order for them to access disability-related services, will they continue to be used in a framework where many individuals are receiving varying levels of support to make progress? Or will curriculum-based measures that assess learning and academic progress replace standardized intelligence tests? Another consideration will be the role that rapidly advancing technology will play in the utilization of paper-and-pencil intelligence tests, as they have historically been conducted. These variations may influence diagnostic understandings as well as service delivery models.

Tracy Paskiewicz, Kathryn Doherty Kurtz, and Melissa Pearrow

See also Aptitude Tests; Standardized Tests; Stanford-Binet Intelligence Scales; Triarchic Theory of Intelligence; Wechsler Intelligence Scales; Working Memory

Further Readings

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). Essentials of cross-battery assessment (3rd ed.). Hoboken, NJ: Wiley.

Gould, S. J. (1981). The mismeasure of man. New York, NY: W. W. Norton … Company.

Ortiz, S. O. (2011). Separating cultural and linguistic differences (CLD) from specific learning disability (SLD) in the evaluation of diverse students: Difference or disorder? In D. P Flanagan & V. C. Alfonso (Eds.), Essentials of specific learning disability identification (pp. 299–325). Hoboken, NJ: Wiley.

Sattler, J. M. (2008). Assessment of children cognitive foundations (5th ed.). La Mesa, CA: Author.

Catherine O. Fritz Catherine O. Fritz Fritz, Catherine O.

Peter E. Morris Peter E. Morris Morris, Peter E.

Interaction

Interaction

848

853

Interaction

The causes of behavior are many and varied. Factors (or variables) that affect behavior do not act in isolation; sometimes, the effect of one factor differs depending upon the value of another. In order to understand the factors influencing a behavior, it may be necessary to examine them in the context of other factors. Factors are said to *interact* with one another when the effects of one factor change depending on the value of another factor. For categorical factors, such as working at home versus in the office, the analysis often involves analysis of variance or chi-square test. For continuous variables, such as a personality score, the analysis is likely to involve either multiple linear regression or analysis of covariance (ANCOVA). To illustrate the meaning and importance of interactions, some examples and commentary follow.

Example 1: Categorical Factors and Data—The Number of Cases in Each Category

For a simple example, consider the fictional data in Table 1, reporting the number of children who complete their homework versus those who do not. In the entire sample of 200 boys and 200 girls, children were equally likely to complete their homework regardless of gender. But when another factor—year of schooling—is considered alongside gender, as in Table 2, an interesting, underlying pattern emerges. In Year 3, girls were far more likely than boys to complete their homework; but in Year 8, the pattern was reversed. This difference in the pattern for gender, depending upon which year group is studied, is an interaction.

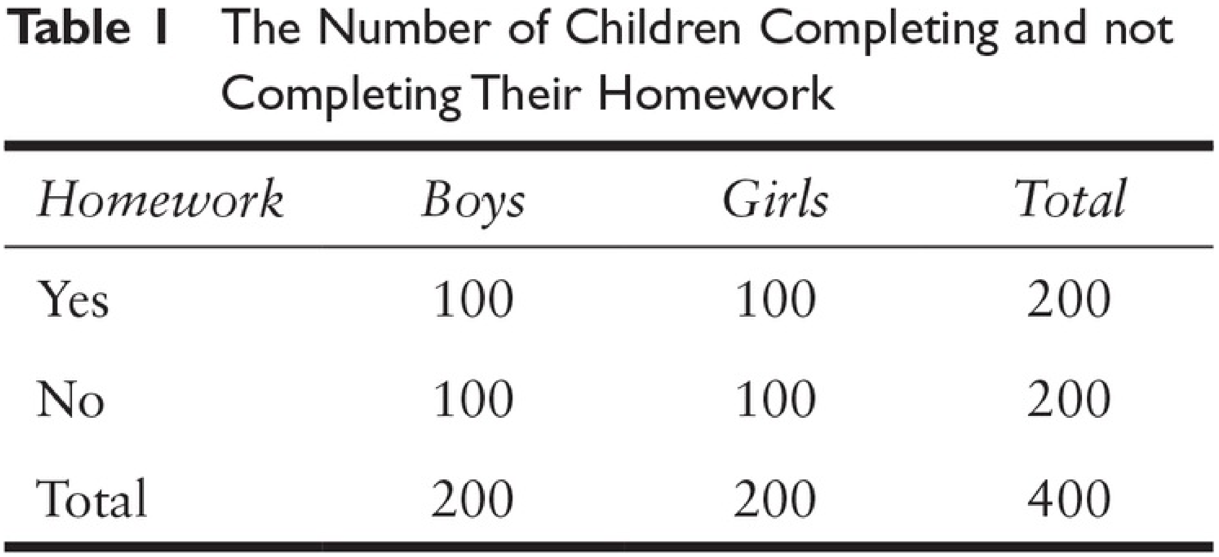
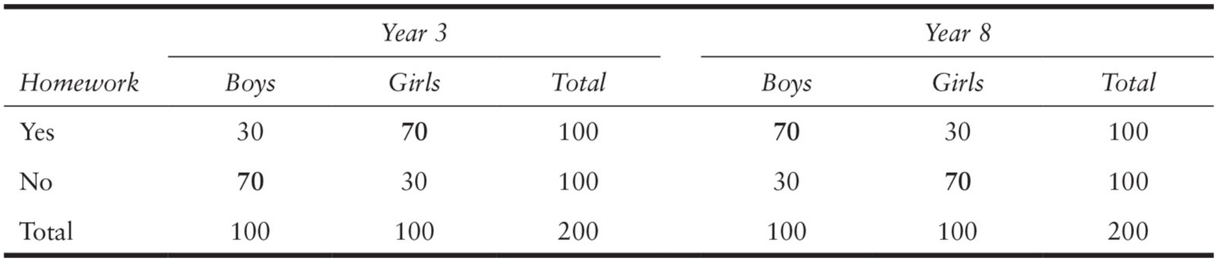


Table 1 The Number of Children Completing and not Completing Their Homework

Homework	*Boys*	*Girls*	*Total*
Yes	100	100	200
No	100	100	200
Total	200	200	400



	Year 3			*Year 8*		
Homework	*Boys*	*Girls*	*Total*	*Boys*	*Girls*	*Total*
Yes	30	**70**	100	**70**	30	100
No	**70**	30	100	30	**70**	100
Total	100	100	200	100	100	200

An interaction is usually described in terms of the *simple main effects* that make it up. A simple main effect is the effect of one factor when just one category of the other factor is considered. Because there are 2 year groups, there will be two simple main effects for gender, one for each year group. For Year 3 students, the simple main effect of gender is that more girls than boys complete their homework, as shown in the left side of Table 2. For Year 8 students, the simple main effect of gender is that more boys than girls complete their homework, as shown in the right side of Table 2. It is equally valid to describe the interaction using the simple main effects of year; there are two simple main effects for year because there are two genders. For boys, the simple main effect of year is that most of them complete homework in Year 8 than in Year 3 (70 vs. 30). For girls, the simple main effect of year is that fewer of them complete homework in Year 8 than in Year 3 (30 vs. 70).

Table 2 shows an interaction between gender and year—homework was completed by 30 boys versus 70 girls in Year 3 and 70 boys versus 30 girls in

Year 8—but there is no main effect of gender (homework was completed by 100 boys vs. 100 girls) and no main effect of year (homework was completed by 100 students from Year 3 and 100 from Year 8).

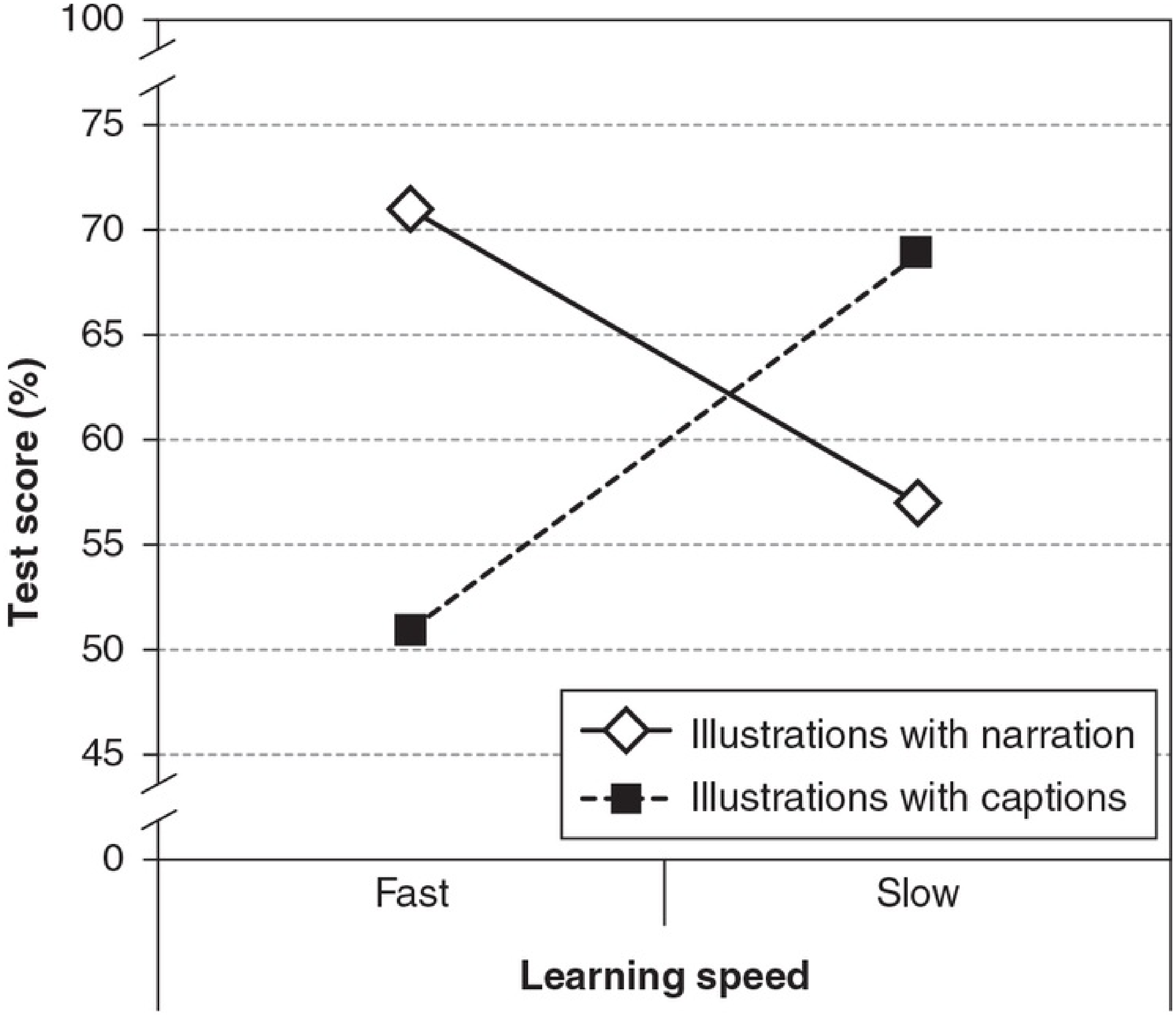
This example provides a clear picture of an interaction and demonstrates the importance of considering interactions and designing research that can test for them. If the research design had considered only gender, without year, the results would have incorrectly demonstrated that gender was not related to homework completion. Similarly, if the research design had considered only year, without gender, the results would have incorrectly shown that homework completion was not related to year in school. It is only by considering both factors together, in a single study, that the actual, more complex picture emerges.

Example 2: Categorical Factors and Continuous Data

When studying a lesson that uses illustrations, is it better to accompany them with audio narration or written captions? To answer this question, in 2007, Egbert Harskamp, Richard Mayer, and Cor Suhre conducted a field experiment, wherein secondary school students studied an illustration-based lesson at their own pace. Half of the students were randomly assigned to hear audio narration accompanying the illustrations; the other half saw written captions with the illustrations. Students were tested on the topic afterward. Test scores from the two groups were quite similar: The sample size (*N* = 27), mean test score (*M* = 70), and standard deviation (*SD* = 14) of the audio narration group were similar to the written text group (*N* = 28, *M* = 74, *SD* = 19). The difference was quite small, so the main effect of lesson format was negligible. But previous research suggests that students who completed their lesson quickly would benefit more from audio narration than written captions. The left side of Figure 1 shows the test scores for fast-finishing students: Audio narration (*N* = 14, *M* = 71, *SD* = 19) led to substantially better test scores than did written captions (*N* = 16, *M* = 51, *SD* = 21). The simple main effect of lesson format for fast-finishing students was that test scores for audio narration were higher by 20, with a 95% confidence interval [5.04, 34.97]. Audio narration was clearly beneficial to these students and is likely to be beneficial to other similar students under similar circumstances.

Figure 1 Mean test performance for fast-and slow-learning students following a lesson wherein illustrations were either accompanied by audio narration or

written captions. The “crossover” interaction between learning speed and lesson modality is evident in the graph. Error bars are not included in order to make the illustration clearer.



The right side of Figure 1 tells a different story. These data are from the students who were slow in completing the lesson. Among these students, test scores were poorer for audio narration (*N* = 13, *M* = 57, *SD* = 19) than for written captions (*N* = 12, *M* = 69, *SD* = 27). The simple main effect of lesson format for the students who finished the lesson very slowly was that test scores following audio narration were somewhat poorer than following written captions. Test scores for audio narration were lower by 12, with a 95% confidence interval [−7.58, 31.48].

According to Example 1, the interaction may be described by either pair of

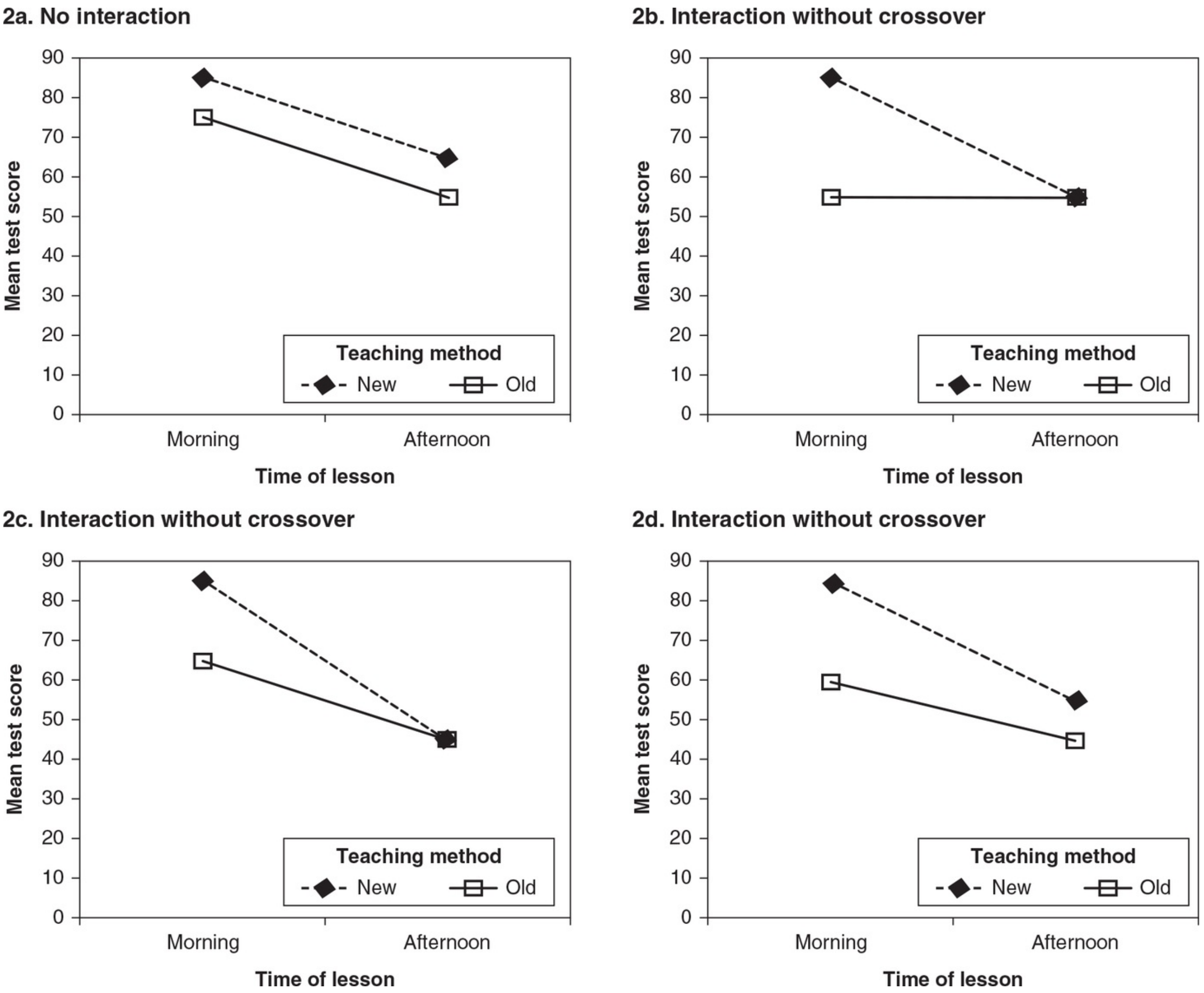
According to Example 1, the interaction may be described by either pair of simple main effects: the effects of lesson format for each type of student or the effects of student type for each lesson format. When one factor is manipulated (e.g., lesson format) and the other factor is something inherent in the people (e.g., speed in completing the lesson, gender), it often makes more sense to think about and report the interaction in terms of the effects of the manipulated variable for each group of people.

The “X” on the graph signifies a “crossover” interaction, in which the effect for one group is reversed in the other group, as is seen in this and in the previous example.

Example 3: Interactions That Do Not Cross Over

Not all interactions cross over, as in this example. A new teaching method might be tested against an established one by allocating one group of students to the new method and one to the old one. Because students respond differently in the morning and afternoon, for half of each group, the teaching method might be employed in the morning, and for the other half, it might be used in the afternoon. To evaluate the effectiveness of each method, all students would later be tested on the taught material. Some possible test score patterns are illustrated in Figure 2.

Figure 2 Some possible patterns for test results when comparing the effectiveness of a new teaching method as compared with an old one, using the teaching method either in the morning or the afternoon.


2a. No interaction

2b. Interaction without crossover

2c. Interaction without crossover

2d. Interaction without crossover

[Figure 2a](#) contains parallel lines that describe a pattern of results with no interaction. Both of the main effects are clear: Morning teaching was always more effective than afternoon teaching and the new method was always better than the old one. The new method should be adopted because it was better than the old method regardless of time of day. Note, though, that the teaching method might interact with other factors that were not examined in this research design, such as gender, year in school, or motivation.

[Figure 2b](#) shows a different possible pattern of results, with an interaction that does not cross over. In the morning, the simple main effect of teaching method is a clear advantage for the new method, but in the afternoon, the two methods are equivalent. The simple main effects are different, describing an interaction. If these factors were investigated in two separate studies, the conclusions would not have provided as much information.

The possible pattern of results in Figure 2c shows an interaction that is similar to Figure 2b except that here (as in Figure 2a) the simple main effects of time of day for each teaching method is that morning lessons yield better test scores for both teaching methods, but the time of day benefit is smaller for the old method than the new one. The clearest way to describe this interaction might be in terms of the simple main effects of teaching method for each time of day because there is a clear difference in the morning and no difference in the afternoon.

The interaction in Figure 2d is similar to Figure 2c in that it can be described with the simple main effects of teaching method at each time of day; the new teaching method has a higher advantage over the old method in the morning than in the afternoon. But Figure 2d is similar to Figure 2a in that the new teaching method leads to better results than the old one at both times of day.

Example 4: Interactions Involving Continuous Variables

Often one or more of the factors of interest are continuous variables, such as test scores, rather than categorical variables. A suboptimal solution that is sometimes adopted is to separate participants into categories on the basis of their scores, perhaps forming two groups using the median as the cutoff. If the scores are 3, 5, 7, 9, 10, 10, 14, and 16, the people with the four lowest scores would constitute the low-score group with the other four in the high-score group. One obvious weakness with this approach is that scores of 3 and 9, which are quite different, are in the same group, whereas similar scores of 9 and 10 are in different groups.

A better solution may be to analyze the data with ANCOVA or multiple linear regression. For ANCOVA, the continuous factors would be treated as covariates. Each covariate must satisfy certain statistical requirements, as detailed in good statistical textbooks. ANCOVA provides information about the effect of each factor, including the covariate, and also analyzes and reports information about all of the possible interactions between and among the factors.

Multiple linear regression is another option that can be used with one or more continuous factors to examine their effect on another continuous variable. To investigate interactions (often termed *moderation*) using multiple linear regression, it is necessary to calculate interaction variables for input to the regression, as in the following example.

The amount of time spent in lessons might be expected to influence students' scores on a later test, so a researcher records how much time each student spends on their mathematics lessons over a week. After a week, students take a test on that topic. The researcher recognizes that lesson time is probably not a good predictor by itself; how well the student spends the time might also have an influence, so a conscientiousness score is included as another factor. Because it is reasonable that conscientiousness might play a greater role when the lessons are longer, it is important to examine the interaction as well, which requires calculating another variable from the two factors. That process is straightforward, if slightly complex, and is briefly outlined here and fully described in good statistical textbooks.

For each continuous factor, a "centered" version must be calculated by subtracting the mean value from the actual value for each individual; the resultant values have a mean of zero and so are termed centered. (It is not necessary to use standardized scores.) The interaction variable is then calculated by multiplying the centered values of one factor by the centered values of the other factor for each individual. Thus, there are three predictors: the centered versions of the two original variables (time and conscientiousness) plus the interaction variable. Standard multiple linear regression can be run and the contributions of each predictor can be evaluated. The β weight and significance for time and conscientiousness describe the main effects of those factors; the β weight and significance of the interaction variable indicates the presence or absence of an interaction. A graph is helpful to see the simple main effects and interpret the interaction; data points for the graph can be calculated by solving the regression equation for high and low, or high, medium, and low values of the predictors. Good textbooks will provide detailed guidance.

Catherine O. Fritz and Peter E. Morris

See also Experimental Designs; Multiple Linear Regression; Standardized Scores; Two-Way Analysis of Variance

Further Readings

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioural sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Harskamp, E. G., Mayer, R. E., & Suhre, C. (2007). Does the modality principle for multimedia learning apply to science classrooms? Learning and Instruction, 17, 465–477.

Keppel, G., & Wickens, T. D. (2004). Design and analysis: A researcher's handbook (6th ed.). Upper Saddle River, NJ: Pearson.

Tabachnick, B. G., & Fidell, L. S. (2007). Experimental designs using ANOVA. Belmont, CA: Thomson Brooks/Cole.

Tabachnick, B. G., & Fidell, L. S. (2014). Using multivariate statistics (6th ed.). Harlow, UK: Pearson.

Richard R Sudweeks Richard R Sudweeks Sudweeks, Richard R

Internal Consistency Internal consistency

853

856

Internal Consistency

Internal consistency is an umbrella-like term that encompasses several different, but related, procedures used to estimate reliability. Although some do not consider internal consistency, strictly speaking, a type of reliability, it is treated in practice as a reliability estimate and it is common to refer to internal consistency as internal reliability. There are several methods for examining internal consistency, but all methods share two common characteristics. First, they all involve the analysis of data obtained from a single test administered once to a group of examinees. Second, they all involve dividing the targeted test into two or more parts, which are then treated as if they were tests themselves.

There are other reliability estimation approaches that do not involve internal consistency, but instead require correlating different administrations of the same test (test–retest reliability) or of different forms of the same test (parallel forms reliability) or scores assigned by different scorers (interrater reliability), but both of these approaches have significant practical limitations. Testing the same group of persons on two different occasions is undesirable from the examinees' point of view and is often impractical to implement. The use of parallel forms is also impractical because of the difficulty of constructing two different forms of the whole test that are equivalent. Similarly, it is often difficult or unnecessary (in the case of objective scoring rules) to arrange for different raters to score the same test.

This entry discusses the estimators of internal reliability, theoretical underpinnings of internal reliability, degrees of parallelism, and computation procedures.

Estimators of Internal Consistency

Instead of comparing scores from replications of the whole test, internal consistency focuses on examining relationships among items within (i.e., internal to) the test. Each item can be treated as a separate test, or the items can be grouped into clusters or part tests. Depending on which estimator is used, the number of parts may be as few as two or as many as the number of individual items in the test. In any case, the part tests or items are treated as if they are interchangeable replicates of the test. The investigator then computes a statistic that summarizes the degree of consistency in the examinees' responses to the various parts.

The estimators in the internal consistency family include, but are not limited to, the following:

- Spearman and Brown's split-half coefficient
- Kuder and Richardson's Formula 20 (KR-20)
- Cronbach's coefficient α
- Hoyt's coefficient
- Kristof's coefficient
- Feldt's coefficient
- Raju's β coefficient
- McDonald's ω coefficient
- Raykov's ρ coefficient

The split-half coefficient proposed independently by Spearman and by Brown in 1910 was the first known estimator of internal consistency. Many of the estimators developed later in this field were attempts to improve on the split-half approach. One advantage of Cronbach's coefficient α is that it is equivalent to the average split-half reliability of all possible half-length tests. The KR-20 estimator is a special case of Cronbach's α that is appropriate when the items are scored dichotomously. More detailed descriptions of these three coefficients are provided under separate entries elsewhere in this volume.

Hoyt's reliability coefficient is computed from the estimated mean squares obtained from the results of a two-way analysis of variance (persons crossed with items) with only one observation per cell. Hoyt's formula is algebraically equivalent to Cronbach's α, so it produces the same result within rounding error. One advantage of Hoyt's approach is that it provides a conceptual link between classical test theory and generalizability theory, which also uses variance components obtained from analysis of variance. The simplest case of generalizability theory is a single-facet design in which persons are the object of measurement and test items are the only facet. That is the same design used by Hoyt. In this simple case, the value of the generalizability coefficient for relative decisions is the same as the value of both coefficient α and Hoyt's coefficient obtained from the same data.

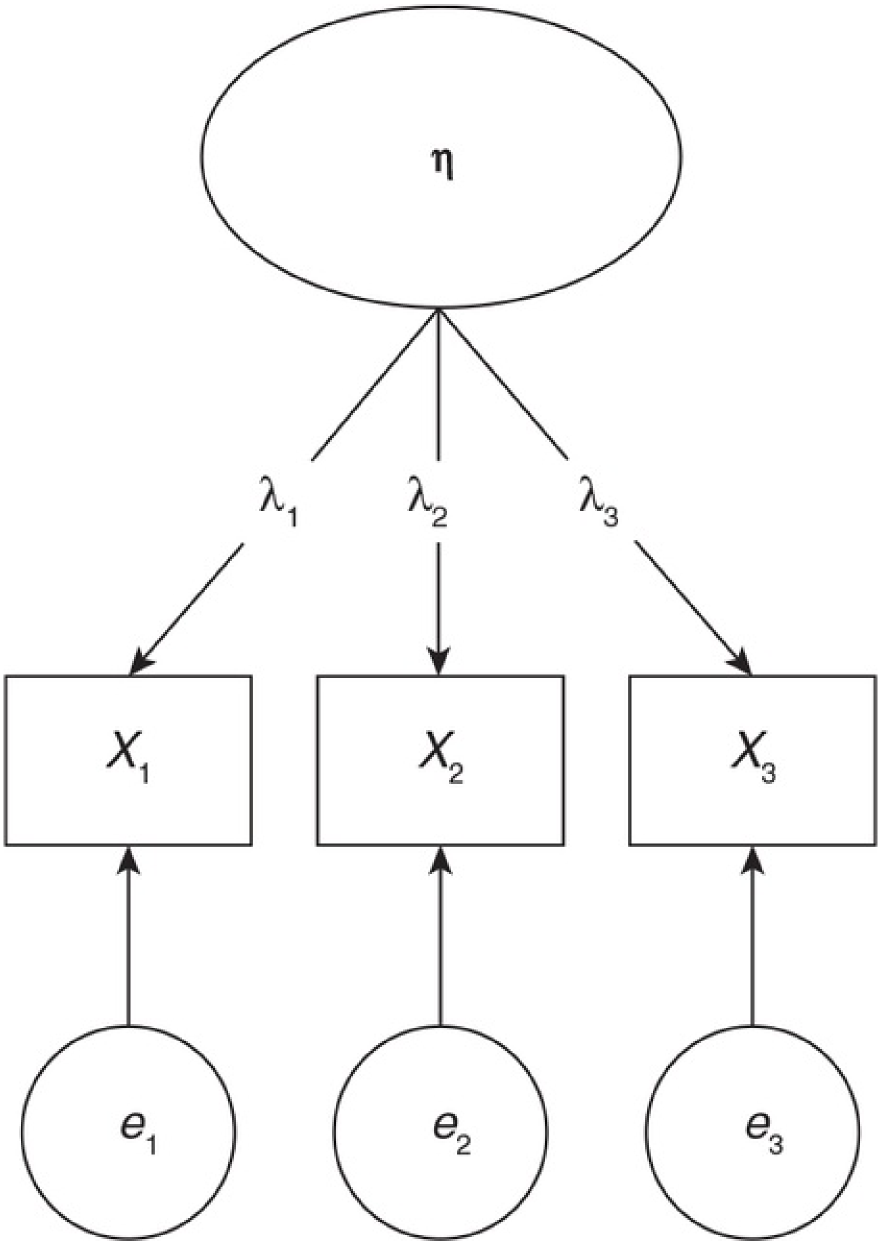
Theoretical Underpinnings

The first seven estimators in the list provided are each based on classical test theory. The last two are model-based estimators based on congeneric test theory, which represents a structural equation modeling approach. In classical test theory, an examinee's score (X_i) on a test is defined as an additive function of the individual's true score (T_i) plus error (E_i). Hence,

$$X_i = T_i + E_i.$$

Congeneric test theory is a special case of classical test theory applied at the item level. The observed score in classical test theory represents an examinee's total score on a test, but in congeneric test theory, the observed score is defined as the examinee's response to a single item. For example, Figure 1 is a path diagram depicting a test consisting of three items (X_1, X_2, and X_3) all intended to measure a single latent variable (η). The E_{i} terms (E_1, E_2, and E_3) represent the error component or uniqueness of each item. These errors are assumed to be uncorrelated. Congeneric theory postulates that an examinee's true score (T_i) on a given item is a linear function of the latent variable (η) being measured. The slope (λ_i) of the linear function for each item is a factor loading (or the discriminating power of the item), whereas the intercept (α_i) represents the relative difficulty or endorsability of that item.

Figure 1 Path diagram depicting a test consisting of three items (X_1, X_2, and X_3) all intended to measure a single-latent variable (η)



For the three-item test illustrated in Figure 1, an examinee’s true score on each item would be modeled as a linear equation as shown with subscripts designating the item. Subscripts representing the examinee have been omitted for simplicity sake.

$$T_1 = \lambda_1 \eta + \delta_1,$$

$$T_2 = \lambda_2 \eta + \delta_2, \text{ and}$$

$$T_3 = \lambda_3 \eta + \delta_3.$$

Substituting each of these formulas for the true score into the classical test theory model applied at the item level produces the following equations, each of which is used to model an examinee’s response to one of the three items.

$$X_1 = \lambda_1 \eta + \delta_1 + E_4,$$

$$X_2 = \lambda_2 \eta + \delta_2 + E_4, \text{ and}$$

$$X_3 = \lambda_3 \eta + \delta_3 + E_4.$$

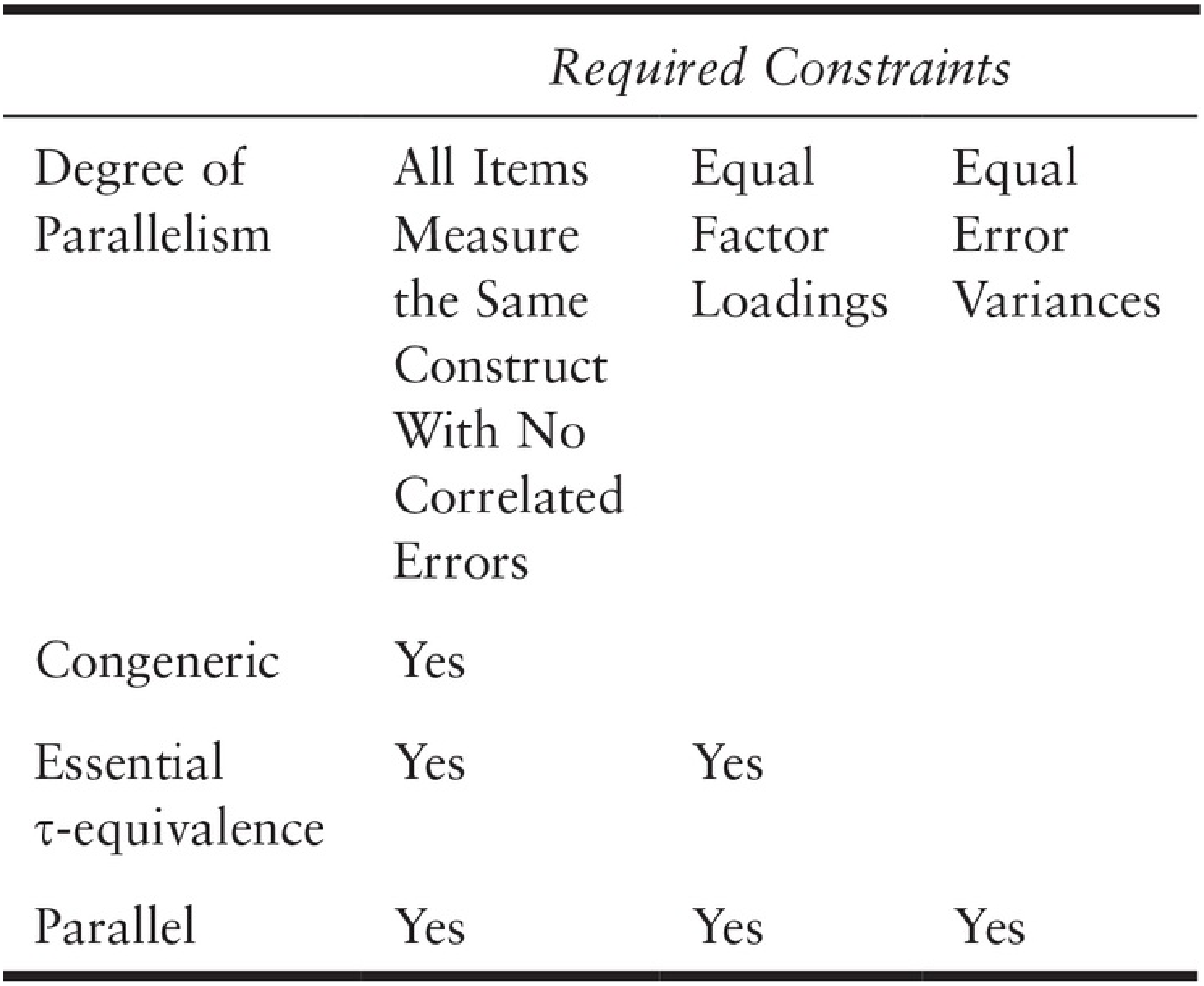
Degrees of Parallelism

In 1967, Novick and Lewis demonstrated that Cronbach’s coefficient α is based on the assumption that the measures in the targeted test are essentially τ equivalent. To understand what essential τ equivalence refers to, one needs to understand two related ideas. The terms *congeneric*, *essentially* τ-*equivalent*, and *parallel* are adjectives used to describe differing degrees of similarity or parallelness among the items or parts of a test. The different possible relationships among the slopes (i.e., factor loadings) and among the error variances associated with the equations for items X_1, X_2, and X_3 can be used to illustrate these three degrees of parallelness.

If the slopes of the items are equal ($\lambda_1 = \lambda_2 = \lambda_3$) and if the error variances are also equal (Var[E_1] = Var[E_2] = Var[E_3]), then the observed variables (X_1, X_2, and X_3) are classified as being parallel measures. If the slopes are equal ($\lambda_1 = \lambda_2$

$= \lambda_3$) but the error variances are not equal (Var[E_1] ≠ Var[E_2] ≠ Var[E_3]), then the observed variables (X_1, X_2, and X_3) are essentially τ-equivalent measures. If the slopes are not equal and if the error variances are not equal, the observed variables are congeneric measures.

As shown in Table 1, these three types of parallelness are cumulative and hierarchically nested. Hence, the three conditions can be empirically tested using structural equation modeling. The analysis begins by testing the congeneric model, which is the least restrictive model. This model specifies that the items all load on the same factor and that there are no correlated errors among the items. If the congeneric model fits the data, then the essentially τ-equivalent model can be tested by imposing equality constraints on factor loadings. Hence, the essentially τ-equivalent model is nested within the congeneric model. Similarly, if the model with equal factor loadings fits the data, then the parallel model can be tested by imposing equality constraints upon the variances. Hence, the parallel model is nested within the τ-equivalent model.



	Required Constraints		
Degree of Parallelism	All Items Measure the Same Construct With No Correlated Errors	Equal Factor Loadings	Equal Error Variances
Congeneric	Yes		
Essential τ-equivalence	Yes	Yes	
Parallel	Yes	Yes	Yes

Computations

One way to compute a test score for each individual examinee is to compute the simple, unweighted sum of each person's responses to each of the items ($Y = X_1 + X_2 + X_3$). In the testing literature, the reliability of the resulting total scores is called composite reliability. McDonald's ω and Raykov's ρ are examples of coefficients proposed to estimate the reliability of an unweighted composite. Each is based on the conceptual definition that reliability is the ratio of true score variance divided by the total variance of the observed scores. Both coefficients can be computed from the results of structural equation modeling. ω can also be computed from the results of an exploratory factor analysis with a single factor extracted. The formula for McDonald's ω is:

$$\omega = \lambda_{i2}\lambda_{i2} + \theta_{ii},$$

where λ_i is the factor loading for the *i*th item and θ_{ii} is the unique variance of the *i*th item. The numerator (the square of the summed factor loadings) provides an estimate of the variance of the true scores. The denominator provides an estimate of the total observed variance. ω is an index of the proportion of the variance in the scale scores explained by the latent variable that is common to all items.

The formula for Raykov's ρ coefficient is

$$\omega = \lambda_{i2}\lambda_{i2} + \theta_{ii} + 2\theta_{ij},$$

where λ_i and θ_{ii} are the same as in ω, and θ_{ij} represents the covariance between any pair of items.

Raykov's coefficient should be used when the model that best fits the data includes correlated error terms. When the model does not include any correlated errors, the extra term in the denominator becomes zero and drops out of the formula. The only difference between Raykov's ρ and McDonald's ω is that ρ explicitly accommodates the possibility that the best fitting model may include correlated errors.

In the 1951 article in which Cronbach described the derivation of coefficient α, he showed that α is a function of the average interitem covariance C_{ij} among the *n* items in a test and the variance of the total scores V_t. His equation 16 cited here describes this functional relationship:

$$\alpha = n_2 C_{ij} V_t.$$

When the items in a test are essentially τ equivalent (which subsumes congenericity, including the assumption that there are no correlated errors), McDonald's ω and Raykov's ρ are algebraically equivalent to Cronbach's equation 16. Therefore, the use of any one of these three coefficients (α, ω, or ρ) is appropriate and defensible when essential τ equivalence holds. When essential τ equivalence does not hold and the data are congeneric, either ω or ρ is more appropriate than α. If the best-fitting model includes correlated errors among two or more pairs of items, then the data are not congeneric and Raykov's ρ is more appropriate.

Instead of computing a composite score by equally weighting all of the items in a congeneric set, some scholars recommend assigning differential weights to the various items and then computing the weighted sum ($Y = w_1X_1 + w_2X_2 + w_3X_3$), where w_1, w_2, and w_3 are the weights. One way to do this is to empirically derive a set of weights that will produce the highest possible internal reliability for that set of items. Since the 1940s, there has been a growing research literature focused on procedures for deriving optimal weights that can be used to obtain maximal reliability for a set of items.

Richard R Sudweeks

See also Coefficient Alpha; Omega; Split-Half Reliability

Further Readings

Bollen, K. A. (1989). Structural equations with latent variables. New York, NY: Wiley.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105–146). New York, NY: American Council on Education.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. Educational and Psychological Measurement, 66, 930–944.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp, 65–110), Westport, CT: Praeger.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

Meyer, J. P. (2010). Reliability. New York, NY: Oxford University Press.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. Psychological Methods, 5, 343–355.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. Behavior Therapy, 35, 299–331.

Wang, T. (1998). Weights that maximize reliability under a congeneric model. Applied Psychological Measurement, 22, 179–187.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach’s α, Revelle’s β, Mcdonald’s, ω_H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70, 123–133.

Brandon W. Youker Brandon W. Youker Youker, Brandon W.

Internal Evaluation

Internal evaluation

856

858

Internal Evaluation

Internal evaluation refers to an evaluation conducted by an organization or program/project staff in which the evaluators are directly accountable to the organization being evaluated. As such, the primary responsibility for the internal evaluation lies with the organization itself. This differs from an external evaluation, which is an evaluation conducted by evaluators from outside of the organization being examined. Internal evaluation (sometimes referred to as in-house evaluation) is essential for organization and program management, as the emphasis is on the concerns and needs of the organization’s administrators, managers, and staff to assist the program managers in better understanding their program theory and improving program processes and outcomes. Furthermore, the internal evaluation process promotes utilization of evaluation findings, reflective practice, and organizational learning.

Michael Scriven points out that the distinction between internal and external evaluation is more of a difference of degree than of kind. For example, if the evaluators hail from the same organization but from a different program or department, they may be considered partially external. Another example of a combined internal–external evaluation is institution or program accreditation in which the internal evaluators conduct a self-assessment, collecting and organizing data for external evaluators appointed by the accrediting organization. The accrediting body then inspects the data collected by the internal evaluators, conducts site visits, and writes an independent evaluation report. In fact, a typical accreditation criterion is that the institution or program seeking accreditation should use the findings from its internal evaluations. This entry examines the history, application, models, history, benefits, and limitations of internal evaluation.

History

During the 1960s and 1970s, evaluation emerged as a profession, yet few organizations had evaluations units and even those that did still relied heavily on contracts with external evaluators. But as professional evaluation grew, there were increasing legislative mandates for the training of evaluators. Likely a result of the perceived importance of evaluation, throughout the 1980s, organizations experienced a notable shift from an emphasis on external evaluation to a focus on internal evaluation; thus, internal evaluation became more commonplace. Strengthening this shift, the passage of the Government Performance and Results Act of 1993 required that all U.S. federal agencies develop strategic plans that described their overall goals and objectives, annual performance plans that included quantifiable measures of progress, and performance reports that described successes. In the 2000s, most federal, state, and local agencies as well as international organizations had internal evaluation units. In fact, evaluation scholars estimate that in North America, approximately half to three quarters of all evaluations are internal.

Application

Internal evaluation is typically used for formative purposes, as the evaluators usually make recommendations and action plans for improving the program; because the evaluators are on staff, they can monitor the implementation of the changes to ensure that the suggestions are converted to action. Nevertheless, internal evaluations can also be summative as is the case for an organization that conducts an evaluation for decision-making purposes, such as, for example, to determine whether or not to continue a program, project, or initiative or for holding its staff accountable to certain standards or practices.

Models

There are two basic models of internal evaluation: what Arnold Love calls the "internal department" and the "embedded internal evaluation." Many large organizations have a separate internal evaluation unit (e.g., the Federal Bureau of Investigation or the World Bank's operations evaluation department). With these internal departments, the evaluators typically report directly to the chief internal departments, the evaluators typically report directly to the chief executive officer or the president of the organization rather than to program administrators or managers. With the embedded internal evaluation, program administrators or managers are often on the evaluation teams as part of their administrative or managerial duties.

Benefits

There are several benefits of conducting an internal evaluation. First, as compared to external evaluation, internal evaluation may reduce evaluation anxiety, as program people often know the evaluators and therefore tend to be able to speak to the program administration and staff more comfortably and candidly. Relatedly, internal evaluators' familiarity with the program, the organizational history and cultural norms, and the political structure consequently lead to avoiding mistakes due to ignorance and often prevent overburdening the organization with irrelevant or overwhelming data collection procedures. Second, internal evaluation costs less money because many organizations have already allocated program resources such as staff time for conducting the evaluation. Third, the information gleaned from the evaluation, as well as the evaluation process, stays in-house, whereas knowledge may vanish with the departure of the external evaluator. For all of these reasons, the internal evaluation may result in an evaluation that is better suited to the information needs of the organization, as well as lead to an increased acceptance of the evaluation findings by the organization's personnel.

Limitations

There are also limitations on internal evaluation. The most significant weakness is evaluator credibility, which causes others to question the evaluator's objectivity and therefore leads to decreased acceptance of the findings by outside parties. This threat to credibility is based on the fact that internal evaluators have an ongoing position within the organization or program being evaluated, and thus the evaluators are dependent on the organization for their employment or career. Additionally, internal evaluators are more at risk of personal retribution and job loss when speaking frankly, especially when reporting negative evaluation findings. The presumption is that administrators, trying to justify decisions or actions and promote their programs, more easily manipulate internal evaluators; hence, the evaluators experience considerable pressure to make the program look good for public relations purposes. As a result, evaluators either dismiss or neglect to report negative findings. The natural conflict of interest causes the internal evaluator to balance providing full disclosure and valid evaluations against advocating for the welfare of the organization or program.

A second limitation of internal evaluations is that they can become a bureaucratic exercise, as the evaluations become so routine that they lose their value and meaning. Ernest House claims that the program people become inured to the evaluation reports and/or the staff become “datawise,” as they learn how to collect data or respond to the evaluators’ questions in a way that avoids sensitive issues or negative findings. A third limitation is that the internal evaluation can become a tool of administration in which the interests of the administrators are confused for the interests of the organization, and thus the evaluators miss the true evaluation needs in favor of data that speak solely to managerial issues. Lastly, it is often the case that the internal evaluators are less knowledgeable about evaluation theory, methodology, and practice than an external evaluator. For instance, historically, internal evaluations have relied heavily on goal-based evaluation methods in which the evaluators gather data and make conclusions exclusively as they pertain to the organization’s or program’s attainment of its stated predetermined goals and objectives, thus failing to investigate unintended or unanticipated outcomes.

One of the best ways to improve internal evaluation is to conduct periodic meta-evaluation. A meta-evaluation is an evaluation of the evaluation. Having an external meta-evaluator provide an independent analysis of the strengths and weaknesses of the internal evaluation can enhance the evaluation process as well as bolster credibility to outside audiences.

Brandon W. Youker

See also Accountability; Conflict of Interest; Data-Driven Decision Making; Goal-Free Evaluation; Impartiality; Objectivity; Program Evaluation; Stakeholders; Summative Evaluation

Further Readings

House, E. R. (1986). Internal evaluation. American Journal of Evaluation, 7(1), 63–64.

Love, A. (1991). Internal evaluation: Building organizations from within. Newbury Park, CA: Sage.

Mathison, S. (2011). Internal evaluation, historically speaking. In B. B. Volkov & M. E. Baron (Eds.), Internal evaluation in the 21st century: New directions for evaluation (pp. 13–23).

Patton, M. Q. (1997). Utilization-focused evaluation: The new century text (3rd ed.). Thousand Oaks, CA: Sage.

Scriven, M. (1991). Evaluation thesaurus (4th ed.). Newbury Park, CA: Sage.

Sonnichsen, R. C. (2000). High impact internal evaluation: A practitioner's guide to evaluating and consulting inside organizations. Thousand Oaks, CA: Sage.

Volkov, B. B., & Baron, M. E. (2011). Issues in internal evaluation: Implications for practice, training, and research. In B. B. Volkov & M. E. Baron (Eds.), Internal evaluation in the 21st century. New directions for evaluation (pp. 101–111).

Andrew Maul Andrew Maul Maul, Andrew

Daniel Katz Daniel Katz Katz, Daniel

Internal Validity

Internal validity

858

861

Internal Validity

In general, the concept of internal validity refers to the degree to which causal inferences are warranted on the basis of a study. Internal validity is thus largely a function of how well a study's design and execution allows researchers to make definitive claims about the causal relationship between one or more independent variables and one or more dependent variables and rule out alternative (i.e., noncausal) explanations for observed associations. Internal validity is often contrasted with external validity, which refers to the extent to which the results of a study can be generalized to situations and people outside the scope of the study itself. It should also be noted that (unlike many uses of the term *validity* in educational and psychological research), the phrase *internal validity* does not refer specifically to testing and measurement but rather to a study taken as a whole. This entry surveys the logic of establishing causal relationships on the basis of studies and discusses some common threats to internal validity.

Establishing a Causal Relationship

Although accounts of the nature of causation differ somewhat between sources, the logic of causal claims in educational and psychological research can generally be understood as invoking counterfactual reasoning of the form "all else being equal, if A had not occurred, B would not have occurred." Although it is obviously not possible to know for certain what would have (not) occurred had events and circumstances been different than they actually are, researchers often work to approximate such counterfactual conditions via carefully controlled studies. Inferring a causal connection between A and B thus requires evidence that, all else being equal, differences in A (e.g., receiving or not receiving an educational intervention) are associated with differences in B (e.g., knowledge as measured by an academic test).

The "all else being equal" clause is important, as causal inference requires researchers to be confident that, other than A, there are not any additional influences on B that could account for observed variation in B. (Such influences, if present, are often referred to as *confounding variables*, and a perceived relationship between A and B that is actually due to a confounding variable is referred to as a *spurious relationship*.) Thus, researchers attempt to experimentally and/or statistically control such potential confounds. Different study designs may achieve this control in different ways; for example, a classic two-group experimental design (sometimes also referred to as a *randomized controlled trial*) involves random assignment of participants into either a treatment group or a control group; the random assignment to groups, in combination with a sample size sufficiently large to control random variation, helps ensure baseline equivalence (i.e., there is nothing systematically different about the populations of participants in each of the groups). If successful, such a design ensures that the only systematic difference between the two groups is the presence or absence of A, and thus, if differences in B are observed between the two groups, the best explanation for those differences is that they were caused by differences in A.

Tight control of all conditions other than variance in A is clearly a strength in terms of isolating potential causal relationships, but it can also be a weakness insofar as it may leave researchers unclear as to the generalizability of the observed relationship to other situations (i.e., the study's external validity). For example, it may be that A causes B only under specific conditions or only for persons with specific characteristics (i.e., there are moderating influences on the causal relationship between A and B); and within a highly controlled experimental context, there may not be sufficient variation in other factors to allow researchers to explore the extent to which causal inferences can be generalized to other settings. For this reason, it is sometimes said that the very strategies that maximize the internal validity of a study may also limit its external validity.

In educational research in particular, rigorously controlled experiments are often not possible for ethical or practical reasons, and thus researchers instead often turn to quasi-experimental or even purely observational study designs, often in turn to quasi-experimental or even purely observational study designs, often in combination with particular statistical techniques that may be used to statistically (rather than experimentally) control potential confounding variables. In the next section, common threats to internal validity and common strategies for mitigating these threats are discussed.

Threats to Internal Validity

In general, internal validity is threatened to the extent that one can find plausible alternative explanations for observed associations between a hypothesized cause and its hypothesized effects.

Some studies involve only a single group, often observed on more than one occasion; for example, a researcher might test students in a single classroom prior to and following an educational program. The first four of the threats described in the following sections (i.e., repeated testing, maturation, history, and regression toward the mean) are particularly relevant to such designs and may largely be mitigated via the inclusion of a control group (i.e., a group receiving no intervention, a placebo, or business as usual).

Repeated Testing

Many study designs involve participants taking a test more than once (e.g., before and after a particular intervention). If the same test is used on both occasions, or even if the test is similar in format or content, participants may perform differently on the second testing occasion purely as a function of increased familiarity with the test (or test format). This threat may be mitigated by the inclusion of a control group, insofar as one may expect the same learning effect to apply to both groups; so if a treatment group exhibits a greater change, it may be used as evidence of the effectiveness of the treatment above and beyond the learning due purely to repeated exposure to the test. This threat may also be mitigated if tests can be designed that measure the same construct and can be linked together but are otherwise as dissimilar as possible (i.e., share no construct-irrelevant variance), though it can be very challenging in practice to design such tests.

In addition to pertaining to traditional tests of knowledge or ability, this threat potentially applies to self-report instruments and other modes of assessment. For example, completing a questionnaire may prime participants to distribute their cognitive resources differently, leading to an observed difference in scores owing solely or mainly to the instrument itself. (In this sense, an instrument may be regarded as an intervention in its own right.)

Maturation

Participants may naturally change (e.g., grow, learn) over time, so observed changes between the two testing occasions may simply be due to natural maturation as opposed to an intervention taking place between the two time points. Again, observing changes in a control group may help mitigate this threat, insofar as the control group may be used to estimate the extent to which such maturation takes place naturally. Alternatively, natural maturation rates for the general population on the variables of interest may already be well established, obviating the need for a control group (for this purpose).

History

In addition to natural maturation, particular historical events (outside of the researcher’s control) may affect the entire target population, making it impossible to determine to what extent observed differences between testing occasions should be uniquely attributed to an intervention. Again, comparing observed differences to those of a control group can help mitigate this effect, if it may be assumed that historical events have influenced the treatment and control groups equally. All else being equal, this threat may also be mitigated by a shorter time period between testing occasions and greater insulation from outside influences, but it may not be possible to guarantee that participants will not be exposed to outside forces.

Regression Toward the Mean

This threat pertains to situations in which participants are selected for a program or intervention on the basis of extreme observed scores (as when, e.g., a remedial academic program is piloted on the lowest scoring students in a school). Because observed scores are often due in part to random variation, scores that are extreme upon initial measurement are often closer to the mean upon second measurement, a phenomenon known as regression toward the mean. In a single-group design with participants selected on the basis of extreme observed scores, it may not be possible to separate true treatment effects from observed scores, it may not be possible to separate true treatment effects from regression toward the mean. This threat may be mitigated by random assignment (so as to ensure the selection of participants with scores from the full range of the distribution rather than only one extreme end). Alternatively, a randomly equivalent control group may be used to estimate the degree to which scores can be expected to naturally regress toward the mean, allowing researchers to estimate the additional effect of the treatment.

Selection

In many real-world settings, people may select themselves into groups (rather than being randomly assigned to a group by a researcher), opening up the possibility that groups are not randomly equivalent at baseline. In this case, it may not be possible to determine whether observed differences between groups are due to the effect of an intervention or to preexisting differences in the participants prior to the intervention (or an interaction between the two). For example, participants who volunteer to participate in a study may (on average) be systematically different from the general population in terms of their attitudes, values, motivation, prior knowledge, or in terms of demographic characteristics.

This threat may, of course, be mitigated via random selection. Failing that, researchers may attempt to control baseline differences using statistical models such as multiple regression or propensity score matching. In general, such techniques may be effective to the extent that relevant baseline differences can be identified and measured a priori. (This is in contrast to random assignment, which, in principle, can be used to control all differences between groups whether or not they are identified and measured.)

Mortality

In addition to differences between groups at baseline due to selection effects, there may be differences between groups at any testing occasion after the first caused by differential patterns of attrition or who drops out of the study. Attrition is not a serious threat to internal validity if it is randomly distributed across groups (though it reduces the effective sample size and therefore reduces statistical power) but may be a serious threat if it is systematically related to features of the study design (e.g., if participants with higher levels of initial knowledge are more inclined to persist through an educational program). Obviously, this threat may be avoided if researchers can take measures to ensure that all participants complete the study. Failing that, researchers may be able to mitigate this threat by closely investigating patterns of attrition and selecting an appropriate analytic strategy (see the entry on *missing data analysis* for more details).

Social Threats

If it is possible for participants in a treatment group and a control group (or different treatment groups) to interact with one another, inferences about the unique effect of the treatment may be compromised. Such interaction may take several forms. For example, members of a treatment group may provide information to members of a control group, with the consequence that a true causal effect of the treatment may not be noticed because it has been applied (to some extent) to both groups. As a different example, if study participants are aware of another group receiving a treatment, they may alter their own behavior out of competitiveness. This threat may be particularly pernicious in natural settings such as schools, in which it may not be possible to prevent students from interacting with one another.

Andrew Maul and Daniel Katz

See also Causal Inference; Ecological Validity; Experimental Designs; External Validity; Quasi-Experimental Designs; Threats to Research Validity

Further Readings

Borg, W. (1984). Dealing with threats to internal validity that randomization does not rule out. Educational Researcher, 13(10), 11–14.

Cook, T. D., Campbell, D. T., & Day, A. (1979). Quasi-experimentation: Design … analysis issues for field settings (Vol. 351). Boston, MA: Houghton Mifflin.

Shadish, W., Cook, T., & Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton

Mifflin.

Stanley, D., & Campbell, J. (1966). Experimental and quasi-experimental designs for research.

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

Interquartile Range Interquartile range

Interquartile Range

Descriptive statistics summarize or describe data sets using measures of central tendency and dispersion. Measures of central tendency (e.g., mean, median) identify the dominant, representative, or typical data value, whereas measures of dispersion (e.g., standard deviation) communicate the spread or variability in the data set. Interquartile range (IQR) is a measure of dispersion that encompasses the middle half of the data by taking the difference between the data values positioned at the 25th and 75th percentiles. The IQR accentuates the central range of the data rather than the maximum and minimum values. This entry explains how to calculate the IQR as well as how IQRs are commonly displayed graphically.

To determine the IQR, the data are first arranged in ascending order and subdivided into four equal portions or quartiles. Each quartile contains 25% of the data observations. Next, the data values associated with the 25th and 75th percentiles are determined. For n observations, the 25th percentile or first quartile (Q1) data value occurs at $(n + 1)/4$ and the 75th percentile or third quartile (Q3) data value occurs at $3(n + 1)/4$; the 50th percentile or second quartile is the median. Often whole integers do not result and interpolation is required. For example, a data set with 16 observations denotes that Q1 occurs at the 4.25 observation, signifying that Q1 is the fourth observation plus 0.25 times the difference between the values of the fourth and fifth observations. Finally, the IQR is found by subtracting the Q1 data value from the Q3 data value. A large (small) IQR indicates a data set with a greater (lesser) central dispersion and more (less) variability.

The following data set ($n = 11$), prearranged in ascending order, is used to illustrate the process for obtaining the IQR:

1 7 9 10 12 13 15 15 16 17 24.

Based on the quartile data position formulas, Q1 occurs at the third observation (9) and Q3 occurs at the ninth observation (16); these values represent the median of the lower and upper portions of the data set, respectively. The IQR is the difference between Q3 and Q1 or 7. In comparison with the data range (i.e., the difference between the maximum and minimum values), the IQR in this example is relatively small and indicates less variability in the central half of the data set than with the data set as a whole.

The IQR is commonly displayed as part of a box plot, a type of graph that shows the position of Q1 and Q3, the median, and the data range. In a box plot, the rectangular box is created from the Q1 and Q3 boundaries, thus highlighting the IQR and the middle half of the observations. The median (second quartile) and sometimes the mean are displayed as single lines within the box. Lines or "whiskers" are drawn extending from the box edges in opposite directions until the maximum and minimum values that are not outlier values are reached. Outliers are defined as data points more (less) than 1.5 IQR above (below) the third (first) quartile; outliers are represented by a single identifier (e.g., asterisk) beyond the whiskers.

Jill S. M. Coleman

See also Box Plot; Descriptive Statistics; Percentile Rank; Quartile

Further Readings

Gravetter, F. J., & Wallnau, L. B. (2012). Statistics for the behavioral sciences (9th ed.). Belmont, CA: Wadsworth Publishing.

Tokunaga, H. T. (2015). Fundamental statistics for the social and behavioral sciences. Thousand Oaks, CA: Sage.

Karen D. Multon Karen D. Multon Multon, Karen D.

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

InterRater Reliability Interrater reliability

862

865

InterRater Reliability

Interrater reliability, which is sometimes referred to as interobserver reliability (these terms can be used interchangeably), is the degree to which different raters or judges make consistent estimates of the same phenomenon. For example, medical diagnoses often require a second or third opinion. Competitions, such as judging of art or a figure skating performance, are based on the ratings provided by two or more raters. Researchers might have raters assigning scores for degree of pathology in an individual or type of verbal response in a study examining communication. In the area of psychometrics and statistics, reliability is the overall trustworthiness of a measure. Common terms to describe reliability include *consistency*, *repeatability*, *dependability*, and *generalizability*. High reliability is achieved if similar results are produced under consistent conditions. For example, measuring an adult's height is often very reliable because the method of measuring height is consistent. This entry reviews the importance and types of reliability, details methods for calculating interrater reliability, and discusses how to choose a method of calculation.

Although there are many different types of reliability methods (including interrater reliability, which is described in detail in this entry), the goal of estimating any type of reliability is to determine how much of the variability in scores (or ratings) is due to errors in measurement and how much is due to the variability in true scores. A true score is the replicable feature of the concept or phenomenon being measured. Errors of measurement are components of the observed score that reflect uncertainty of the true score. Errors of measurement include systematic error as well as random error. This simple equation represents the conceptual breakdown of this relationship:

Observed test score = true score + error score

Values of reliability coefficients are generally reported as correlational indices and range from .00 (all errors) to 1.00 (no error with perfect reliability). Highly reliable scores (i.e., those closer to 1.00) are accurate and reproducible.

The type of reliability often used in various disciplines and professions such as anthropology, education, marketing, medicine, psychology, sports, and even the arts is interrater reliability. Although rating scales can take many forms, they typically require the rater to make a subjective judgment about some characteristic of an object by assigning it to some point on a scale defined in terms of that characteristic. Thus, reliability or the consistency of the rating is of utmost importance because the results should be generalizable and not be the idiosyncratic result of one person’s judgment. Stable characteristics that are clearly defined for the raters are the primary contributors to interrater reliability. Inconsistency of interrater reliability can happen due to many factors. First, the factors might be temporary, such as rater motivation, health issues (e.g., severe headache), or fatigue. Second, the factors contributing to inconsistency may be labeled as specific; examples include not comprehending the task or fluctuations in memory or attention. Third, some aspects of the situation may interfere with careful rating, such as environmental noise or some type of major disruption during the task. Finally, chance (e.g., luck, guessing) may play into the consistency of the ratings.

Thus, the interrater reliability index is the degree to which the scores of different raters are proportional when expressed as deviations from their means. This is not the same as interrater agreement, which requires that the raters make exactly the same decisions about the same phenomenon. That is, when judgments are made on a numerical scale, the raters give identical scores when judging the same object, behavior, or person. Researchers can decide, however, to define agreement as either identical ratings or when the discrepancy is no more than a small number of points (usually one or two points). If the researcher determines that a small discrepancy can be considered agreement, the chi-square value for identical agreement should still be reported. Thus, it is possible to have high interrater reliability but low interrater agreement and vice versa. High interrater reliability and high interrater agreement occur when the scores are identical or nearly identical and there is an adequate amount of variability among the scores. Low interrater agreement and high interrater reliability occur when the raters give different ratings for each instance of the measured phenomenon, but each set of ratings are proportional. For the case of high interrater agreement and low interrater reliability, the variability of the ratings may be very small. Thus, it is possible that whatever is being rated is homogeneous on that characteristic of interest. So the researcher must examine further to see whether that is the case by having the judges rate another set of subjects known to be heterogeneous on the same characteristic. Finally, when both interrater agreement and interrater reliability are low, the ratings are of no value and should not be used for research or other purposes (e.g., to judge a competition).

Calculating InterRater Reliability

Simple classification data or nominal data require at least two raters to produce the categorical score for participants. Many times, the question is how often do the raters agree? To calculate agreement, a contingency table is drawn. For example, two raters make 100 observations and for each observation, the raters check one of three categories. If the two raters check the exact same category in 83 of the 100 instances, then the percentage of agreement is 83 (.83). Thus, the percentage of agreement gives a rough estimate of interrater reliability. Owing to the ease of calculation (which can be done by hand) and the fact it is easy to understand, it is the most popular method of computing interrater agreement or consensus estimate. In addition, it works no matter how many categories are used in each observation. Adequate levels of agreement are typically considered to be at least .70 or 70%.

A better method to use for calculating agreement for nominal data is Cohen's kappa. This statistic ranges from 0 to 1 and represents the proportion of agreement corrected for chance. The calculation for Cohen's kappa is as follows:

$$K = (\rho_a - \rho_c) / (1 - \rho_c),$$

where ρ_a is the proportion of times the raters agree and ρ_c is the proportion of agreement that can be expected by chance. Cohen's kappa is recommended when the same two judges perform the ratings. For this statistic, .50 or 50% is considered acceptable. If the number of judges who are rating each observation is the same, but the observations are rated by different judges, then Fleiss's kappa is the preferred method to use.

If researchers use open-ended replies for questions such as "what do effective teachers do?" the results can be content coded. Typically, three or more raters will agree upon the themed categories and then work independently to apply the will agree upon the themed categories and then work independently to apply the codes to the responses. Themed categories are nominal data. Interrater reliability or consensus about the meaning of the open-ended responses can be calculated with Krippendorff’s α statistic.

In contrast to agreement estimates of interrater reliability that are primarily used with nominal data, consistency estimates are generally used with continuous data. The assumption is that it is not necessary for the raters to have a common interpretation of the rating scale (e.g., Levels 1 to 5) as long as each rater is consistent in assigning scores to the phenomenon being observed. Adequate interrater reliability of consistency estimates are typically .70 or better. The estimates typically used are (a) correlation coefficients (e.g., Spearman, Pearson), (b) Cronbach’s α coefficient, and (c) the intraclass correlation coefficient.

The most widely used statistic for calculating the degree of consistency between independent judges is the Pearson product–moment correlation coefficient. Values for this statistic range from +1 to −1. Those values approaching either end of the range (i.e., +1 or −1) indicate a consistent pattern of rating, whereas those values that are close to zero mean it is almost impossible to predict the score assigned by rater A given the score of rater B. An acceptable level of interrater reliability using the Pearson product–moment correlation coefficient is .70. A major assumption of the Pearson correlation is that the data are normally distributed. Therefore, if the data are not normally distributed, the Spearman rank coefficient should be calculated. For example, if two judges rate candidates for graduate study from strongest to weakest, then a rank order is being used and a Spearman rank coefficient is the appropriate statistic.

Cronbach’s α correlation coefficient is used to compute interrater reliability if there are more than two raters. Again, an acceptable level is considered to be .70; if the coefficient is lower than that, it means that most of the variance of the composite score is error variance and not true score variance.

The most conservative measure of interrater reliability for ordinal and interval data is the intraclass correlation or *R*. It is also considered to be the best statistic available for obtaining interrater reliability. Values close to the upper limit of *R* (1.00) are considered to show a high level of interrater reliability, whereas an *R* approaching 0 means the ratings are extremely unreliable and have no use. The minimal level of an intraclass coefficient correlation considered acceptable is .60. *R* is the proportion of the total variance in the ratings caused by the variance in the phenomena (or persons) rated. Because there is more than one formula available for calculating the intraclass correlation coefficient, the researcher must determine whether (a) the concern is primarily with the average rating of the individual judge or all judges and (b) the mean differences in the ratings of the judges should be considered rater error.

The most popular methods among researchers to determine consistency estimates of interrater reliability are factor analysis and the many-facets Rasch model. However, both methods do require a certain level of expertise to calculate correctly. The primary assumption for both of these measurement estimates is that all information from all raters should be used when calculating a summary score for each respondent or phenomenon. This information from judges must include discrepant ratings. With factor analysis as a method of calculating interrater reliability, the amount of shared variance among the ratings can be determined. The generally accepted minimum level of explained variance is 70%. Once the interrater reliability is calculated, each subject or phenomenon will get a summary score based on the loading on the first principle component underlying the ratings. The many-facets Rasch model determines the ratings between judges empirically. In addition, the difficulty of each item and how severe or lenient each judge is can be directly compared. The many-facets Rasch model can also calculate the degree of intra-rater reliability or how internally consistent that particular judge is in the judge’s ratings. Acceptable rater values for the many-facets Rasch model are greater than .70 but less than 1.30.

Choosing a Method of Calculation

When making a determination about how to calculate interrater reliability, several questions will guide the choice. Questions to be considered include the following: What level of measurement (i.e., nominal, ordinal, or interval) are the data? How many raters are needed to be confident in the results? What is the minimum amount of agreement needed for the raters to achieve? Must the raters agree exactly or is it acceptable for them to differ as long as the differences are systematic? What are the practical considerations to determine interrater reliability such as technical expertise, time needed, and funding?

There is no “best” way to determine interrater reliability. Each way to calculate interrater reliability has its own strengths and weaknesses as well as its own assumptions. For example, chance will very likely impact the percentage of agreement approach, whereas the prevalence of the phenomenon to be rated will affect Cohen’s kappa.

Generally, a simpler statistical method for the calculation of interrater reliability is preferable to a more complicated one. Basic methods can give the results likely to be needed, although more advanced methods can complement those results. Thus, goals of the study, the nature of the data (e.g., level of measurement and the degree of normality), and the resources available (e.g., funding, expertise) will often determine what method of calculation to use. Reliability estimates can be improved by establishing clear guidelines for rating, additional training of raters, and practice. Raters can practice judging the phenomenon, then compare ratings and make the guidelines for rating clearer and more concise in order to become more consistent, and thus, achieve a higher level of interrater reliability.

Karen D. Multon and Jill S. M. Coleman

See also Correlation; Instrumentation; Pearson Correlation Coefficient; Reliability; Spearman Correlation Coefficient

Further Readings

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37–46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378–382.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communications Methods and Measurement, 1, 77–89.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), Best practices in quantitative methods (pp. 29–49). Thousand Oaks, CA: Sage.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology, 22, 358–376.

Jie Chen Jie Chen Chen, Jie

Interstate School Leaders Licensure Consortium Standards Interstate school leaders licensure consortium standards

865

869

Interstate School Leaders Licensure Consortium Standards

The Interstate School Leaders Licensure Consortium (ISLLC) standards are a standards-based framework to reshape principalship and strengthen the core of effective leadership. The ISLLC standards provide state and district leaders guidance on what school leaders should know and do and describe what they can do to fortify organizations, support teachers, guide instruction, and advance student learning. This entry provides an overview of the history of the ISLLC standards; compares the standards in 1996, 2008, and 2015; analyzes the chronical changes of the standards; and discusses the implementation of the standards.

Development of the ISLLC Standards: A Historical Review

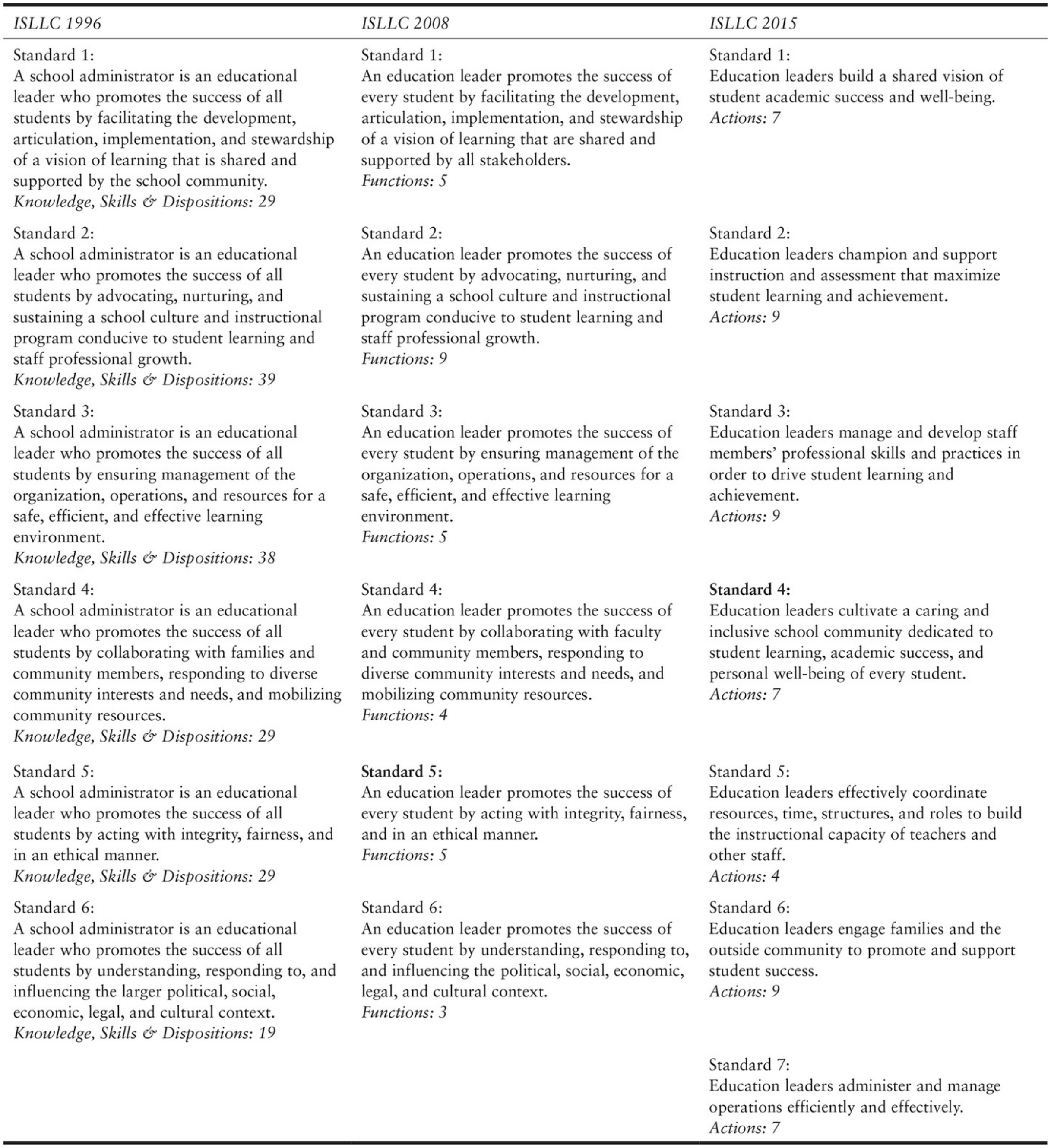
The ISLLC initiative began in August 1994, purporting to develop standards that influence the leadership skills of existing school leaders. The ISLLC standards were first developed in 1996, revised to reflect changes in school leadership expectations in 2008, and refreshed to respond to the changing role and responsibilities of school principals in 2015.

The ISLLC (1996) Standards

In 1996, the ISLLC standards were developed by the Council of Chief State School Officers in collaboration with National Governors Association, the

National Policy Board on Educational Administration, and other organizations. With funding from the Wallace Foundation, the standards were drafted by personnel from 24 state education agencies and representatives from various professional associations to help strengthen preparation programs in school leadership.

The ISLLC 1996 team decided to focus on standards for three reasons: (1) standards could provide appropriate and powerful leverage point for reform; (2) a set of common standards was absent in the area of educational administration; (3) the standards approach could provide the best opportunity to allow different stakeholders to endeavor to improve in various aspects. The 1996 version had six standards (see Table 1). Each standard comprised three elements: the knowledge required for the standard, the dispositions manifest by the accomplishment of the standard, and performances that could be observed by an administrator who is accomplished in the standard.



ISLLC 1996	*ISLLC 2008*	*ISLLC 2015*
Standard 1: A school administrator is an educational leader who promotes the success of all students by facilitating the development, articulation, implementation, and stewardship of a vision of learning that is shared and supported by the school community. *Knowledge, Skills & Dispositions: 29*	Standard 1: An education leader promotes the success of every student by facilitating the development, articulation, implementation, and stewardship of a vision of learning that are shared and supported by all stakeholders. *Functions: 5*	Standard 1: Education leaders build a shared vision of student academic success and well-being. *Actions: 7*
Standard 2: A school administrator is an educational leader who promotes the success of all students by advocating, nurturing, and sustaining a school culture and instructional program conducive to student learning and staff professional growth. *Knowledge, Skills & Dispositions: 39*	Standard 2: An education leader promotes the success of every student by advocating, nurturing, and sustaining a school culture and instructional program conducive to student learning and staff professional growth. *Functions: 9*	Standard 2: Education leaders champion and support instruction and assessment that maximize student learning and achievement. *Actions: 9*
Standard 3: A school administrator is an educational leader who promotes the success of all students by ensuring management of the organization, operations, and resources for a safe, efficient, and effective learning environment. *Knowledge, Skills & Dispositions: 38*	Standard 3: An education leader promotes the success of every student by ensuring management of the organization, operations, and resources for a safe, efficient, and effective learning environment. *Functions: 5*	Standard 3: Education leaders manage and develop staff members' professional skills and practices in order to drive student learning and achievement. *Actions: 9*
Standard 4: A school administrator is an educational leader who promotes the success of all students by collaborating with families and community members, responding to diverse community interests and needs, and mobilizing community resources. *Knowledge, Skills & Dispositions: 29*	Standard 4: An education leader promotes the success of every student by collaborating with faculty and community members, responding to diverse community interests and needs, and mobilizing community resources. *Functions: 4*	**Standard 4:** Education leaders cultivate a caring and inclusive school community dedicated to student learning, academic success, and personal well-being of every student. *Actions: 7*
Standard 5: A school administrator is an educational leader who promotes the success of all students by acting with integrity, fairness, and in an ethical manner. *Knowledge, Skills & Dispositions: 29*	**Standard 5:** An education leader promotes the success of every student by acting with integrity, fairness, and in an ethical manner. *Functions: 5*	Standard 5: Education leaders effectively coordinate resources, time, structures, and roles to build the instructional capacity of teachers and other staff. *Actions: 4*
Standard 6: A school administrator is an educational leader who promotes the success of all students by understanding, responding to, and influencing the larger political, social, economic, legal, and cultural context. *Knowledge, Skills & Dispositions: 19*	Standard 6: An education leader promotes the success of every student by understanding, responding to, and influencing the political, social, economic, legal, and cultural context. *Functions: 3*	Standard 6: Education leaders engage families and the outside community to promote and support student success. *Actions: 9*
		Standard 7: Education leaders administer and manage operations efficiently and effectively. *Actions: 7*

Source: @2008 by CCSSO and member states. The Interstate School Leaders Licensure Consortium (ISLLC) Standards. Available for use with attribution Under a Creative Commons Attribution 4.0 International License (CC BY) which can be found at: https://creativecommons.org/licenses/by/4.0/ the ISLLC Standards are available at: http://www.ccsso.org/documents/2008/educational_leadership_policy_standards_2008.pdf

Note: Changes made from 1996 to 2008 to the text of each standard are underlined in the table. "Knowledge, Skills & Dispositions: 29" means there are in total 29 items in terms of Knowledge, Skills and Dispositions under that standard; "Functions: 5" means there are 5 functions under that standard; "Actions: 7" means there are 7 actions under that standard. ISLLC = Interstate School Leaders Licensure Consortium.

The ISLLC (2008) Standards

In response to the changing nature of the educational environment, especially the changing nature of leadership, the ISLLC (1996) standards were updated in 2008. The 2008 standards retained the structure of the six original ISLLC standards, but for new purposes and audiences, and were policy oriented (see Table 1). “Functions” that define each standard were used to replace the knowledge, skills, and dispositions that defined the 1996 standards.

The ISLLC (2015) Standards

To reflect the changes in school leadership expectations and responsibilities, the 2008 standards were modified in 2015 to clarify the most important work and responsibilities of learning-focused leaders in today’s education context. The 2015 standards provided a framework for state departments of education and districts alike to better prepare, support, and evaluate education leaders in their efforts to help every child succeed. The seven ISLLC (2015) standards specified the essential responsibilities of effective leaders who inspire student learning and achievement (see Table 1). Each standard included a series of actions a leader dedicated to transformational change must take.

Comparison of ISLLC (1996), ISLLC (2008), and ISLLC (2015) Standards

The ISLLC standards were developed and updated to ensure district and school leaders are capable to enhance student achievement and meet new expectations. Table 1 summarizes the ISLLC standards published in 1996, 2008, and 2015. Changes of the standards from 1996 to 2008 are underlined. Numbers of the knowledge, disposition, and performance indicators for the 1996 standards, of the functions for the 2008 standards, and of the actions for the 2015 standards are listed under each standard in the table.

All three versions of the ISLLC standards provided model standards for school leaders, although with different focuses. The updates of the standards signified the ISLLC’s reaction to the dramatic changes in the education policy environment. The ISLLC (1996) standards focused on the topics that form the most important elements of effective leadership.

When the 1996 standards were developed, there was little research or consensus on the characteristics of good school leaders, the principals' role in promoting student achievement, and the best policies and practices for expanding the standards to cultivate effective administrators. To address these limitations, the 2008 standards provided high-level guidance and insights to state policymakers as they work to improve education leadership preparation, licensure, evaluation, and professional development.

Compared to the 2008 standards, the 2015 standards gave more significance to certain leadership domains, such as a school's instructional program, culture, and talent management. While the 2008 standards were policy oriented, the 2015 standards were improvement focused, interdependent, and integrated. The "actions" defining each standard followed a sequence that corresponds to the four stages of the improvement cycle: study, develop, enact, and evaluate.

Implementation of ISLLC Standards

The ISLLC standards were established to strengthen school leadership in a variety of ways, such as creating a framework to better assess candidates for licensure and bringing greater coherence to professional development for school leaders. The standards have been widely used by states, school districts, and other professional associations. The specific areas where the standards can be linked to strengthen school leadership include certification, professional development, preparation program design, administrator selection, licensure, administrator evaluation, preparation program approval, developing or refining standards, and relicensure.

Many states require that administrators qualify for the ISLLC certification. Some states adopt, adapt, or use the ISLLC standards as a model for developing their own standards. With these guiding standards, states have made considerable achievement in addressing school leadership and needs at each stage of an education leader's career. In 2006, 43 states reported using ISLLC standards in some way related to administrator licensure. The ISLLC (2015) standards provide the policy platform for school leadership in 45 states and the District of Columbia.

School districts use the standards in numerous ways—helping develop programs to identify and nurture the development of potential school leaders, evaluating school principals, and designing training programs for school administrators, among others. The most noticeable use of the standards has been in the area of assessment for licensure and relicensure. Districts and school boards can use the standards to assess leadership candidates in terms of their knowledge, skills, and dispositions to collaborate with teachers and motivate student learning. The ISLLC standards effectively specify the responsibilities of all school and district leaders and apply to every phase of leadership.

Jie Chen

See also Accountability; Teachers' Associations

Further Readings

Council of Chief State School Officers. (1996). Interstate school leaders licensure consortium standards for school leaders. Washington, DC.

Council of Chief State School Officers. (2008). Educational leadership policy standards: ISLLC 2008. Washington, DC.

Council of Chief State School Officers. (2015). ISLLC 2015: Model policy standards for educational leaders. Washington, DC.

Murphy, J. (2002, September/October). How the ISLLC standards are reshaping the principalship. Principal, 82, 22–26.

Murphy, J., & Shipman, N. (1999). The interstate school leaders licensure consortium: A standards-based approach to strengthening educational leadership. Journal of Personnel Evaluation in Education, 13(3), 205–224.

Murphy, J., Yff, J., & Shipman, N. (2000). Implementation of the interstate school leaders licensure consortium standards. International Journal of Leadership in Education, 3(1), 17–19.

Richard R Sudweeks Richard R Sudweeks Sudweeks, Richard R

Interval-Level Measurement Interval-level measurement

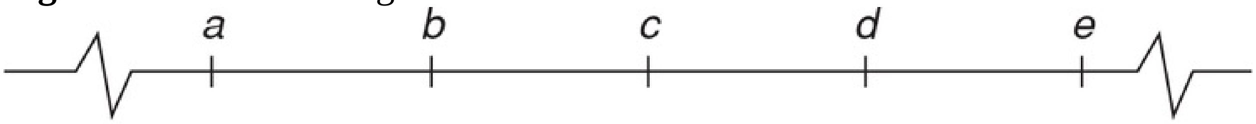
Interval-Level Measurement

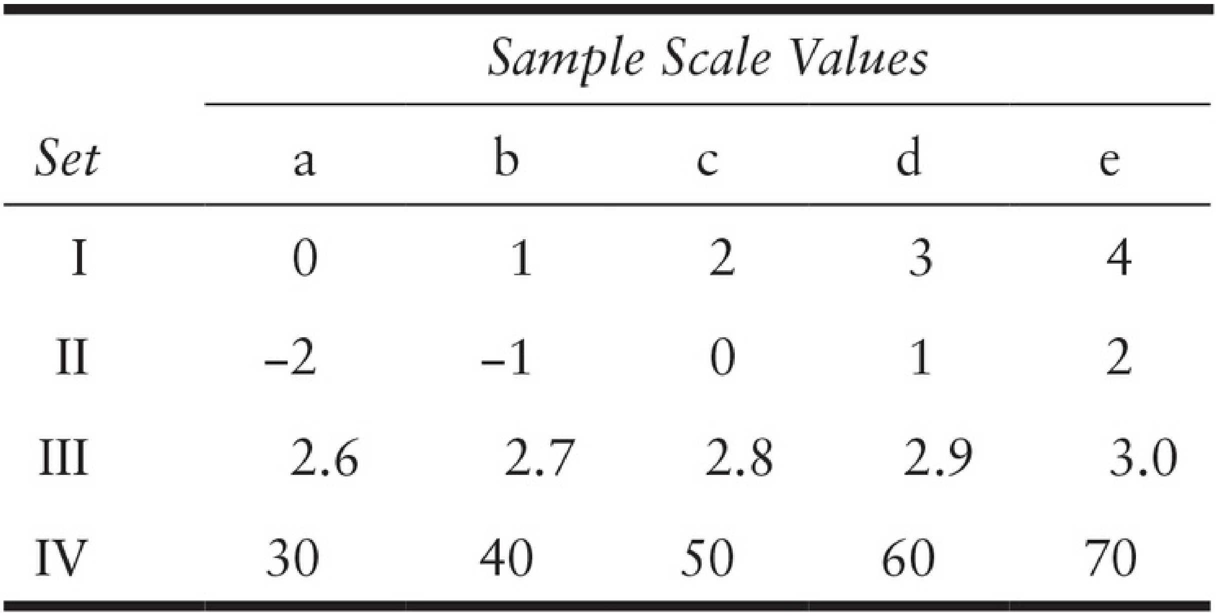
The numerical observations or scores obtained from measuring some definable attribute of a set of objects are at the interval level of measurement if the following three conditions exist:

1. The order of the numbers corresponds to the rank order of the objects with respect to the attribute being measured.
2. The difference between any two consecutive numbers on the measurement scale is the same regardless of which pair of adjacent numbers is considered.
3. The zero point of the number scale used represents an arbitrary origin and does not indicate complete absence of the property being measured.

Scales that satisfy all three conditions are also described as *equal-interval scales* or *equal-unit scores*. The diagram in Figure 1 graphically displays a segment excerpted from an equal-interval scale. The letters *a*, *b*, *c*, *d*, and *e* in the diagram represent an ordered set of five consecutive numbers such as the examples in each row of Table 1.

Figure 1 Schematic Diagram of a Scale





	Sample Scale Values				
Set	a	b	c	d	e
I	0	1	2	3	4
II	–2	–1	0	1	2
III	2.6	2.7	2.8	2.9	3.0
IV	30	40	50	60	70

The first condition specifies that the numbers satisfy the requirements of an ordinal scale including the property of transitivity. This means that the following inequalities hold:

$$e > d > c > b > a.$$

The second condition specifies that all intervals throughout the range of the scale must be the same size. Hence, the following algebraic relationships must be true:

$$(b-a) = (c-b) = (d-c) = (e-d).$$

The third condition specifies that the numeral zero does *not* indicate that an object assigned to that value on the scale completely lacks the measured characteristic. Ratio-level scales have equal intervals and a true zero that indicates a complete absence of the target attribute. In contrast, interval-level scales have equal increments but the zero does not designate a null value.

When an equal-interval scale exists, differences in the numbers convey meaning about the magnitude of the differences between objects with respect to the trait or attribute being measured. Hence, users may be justified in making inferences about how much more or less of the trait certain individuals have. Equal-interval scales are especially useful for measuring within-person growth or change.

Because interval-level measurements do not have a meaningful zero, ratio comparisons of the numbers assigned to individual objects are *not* meaningful. For example, a person who received a score of 60 on an equal-interval scale does not have twice as much of the characteristic measured as an individual with a score of 30. Similarly, an individual who receives a score of 2 does not have half as much of the targeted trait as a person with a score of 4.

However, ratio comparisons of *differences* in interval scale values are meaningful. For example, in Figure 1, the difference ($e - a$) divided by the difference ($d - b$) produces a ratio of 2:1 regardless of which of the four sets of sample numbers are used. This ratio remains the same if a constant (k) is added to each number in the numerator and to each number in the denominator of a ratio of differences. It does not matter what value of k is used as long as the same value is added to each term in the numerator and to each term in the denominator.

$$(e+k)-(a+k)/(d+k)-(b+k)$$

$$=(e+k)-(a-k)/(d+k)-(b-k)$$

$$=(e-a)/(d-b)=2.$$

If a linear transformation is applied to the numbers in an interval scale, the transformed scale will also have equal intervals. For example, the numbers in Sets II, III, and IV shown in Table 1 were derived respectively by applying different linear transformations to the numbers in Set I. A linear transformation may shift the numbers to the left or right along the scale and may also systematically change the size of the intervals, but the resulting intervals will all be shortened or stretched by the same amount. Hence, equal-interval scales are invariant to a linear transformation because even though the metric is changed, the objects maintain their relative standing, and the meaning and interpretation of the resulting numbers are unchanged.

A common procedure in education, psychology, and some other disciplines is to create a summary score for each person either by adding or by averaging each individual’s responses to a series of items in a test or inventory. The resulting scores generally do not have equal intervals because the items differ in difficulty or endorsability. Consequently, such scores only approximate interval-level measures. However, even a cursory inspection of the published literature will show that most analysts treat this as a minor flaw that can be ignored. They simply act as if the resulting scores have equal intervals. Advocates of Rasch scaling readily acknowledge this problem. They assert that the use of the Rasch scaling readily acknowledge this problem. They assert that the use of the Rasch model solves the problem because it produces equal-interval scores.

Beginning in the mid-20th century, a decade-long debate occurred among scholars in education and psychology and related disciplines about whether scores from psychological tests and inventories constitute legitimate interval-level measurements and whether analysts are justified in using parametric statistical tests to analyze quasi-interval measurements. Whole textbooks were written advocating that nonparametric methods should be used when the data are not clearly interval-level measurements. Other writers argued that the scale of measurement underlying the data is not a critical issue when choosing a method of analysis. They asserted that the distributional characteristics of the data are more important than the level of measurement and that the focus should be on deciding what the numbers mean and what conclusions are warranted based on the observed findings.

Richard R Sudweeks

See also Levels of Measurement; Ordinal-Level Measurement

Further Readings

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model (3rd ed.). New York, NY: Routledge.

Gardner, P. L. (1975). Scales and statistics. Review of Educational Research, 45, 43–57.

Hays, W. L. (1994). Statistics (5th ed.). Fort Worth, TX: Harcourt Brace.

Heermann, E. F., & Braskamp, L. A. (Eds.). (1970). Readings in statistics for the behavioral sciences. Englewood Cliffs, NJ: Prentice Hall.

McDonald, R. P. (1999). Test theory. Mahwah, NJ: Erlbaum.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New

York, NY: McGraw-Hill Education.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis. Hillsdale, NJ: Erlbaum.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York, NY: McGraw-Hill Education.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), Handbook of experimental psychology. New York, NY: Wiley.

Jackie Waterfield Jackie Waterfield Waterfield, Jackie

Interviewer Bias Interviewer bias

Interviewer Bias

Interviewer bias relates to aspects of the interviewers and the way in which they ask questions and respond to answers—it is distinct from bias arising from the content or wording of questions. Such bias may stem from perceptions of the interviewer’s identity.

The interviewer’s sex, ethnicity, age, attractiveness, social class, level of education, perceived life experience, or professional background may affect how participants respond to questions, especially where these characteristics seemingly relate to the interview topic. Linked to this is the interviewer’s ability to establish rapport with the interviewee; disclosure, especially on personal or sensitive topics, may relate strongly to the degree of rapport established.

Alternatively, interviewer bias may be due to body language or facial expression or due to paralinguistic aspects of communication such as tone of voice or emphasis. A rising inflection at the end of a question, for example, may suggest that a certain answer is anticipated and prompt the participant to respond accordingly.

Interviewer bias may also arise from expectations or preconceptions on the interviewer’s part. If, for example, a researcher interviewing high school students about bullying assumes that boys and girls will tend to have different perspectives on this issue, this may influence how the interview is conducted, recorded, or analyzed. Moreover, certain questions may be either included or omitted on the basis of such assumptions. For example, the interviewer may assume that girls are more concerned about verbal than about physical bullying and ask questions that reflect this assumption or use prompts during the interview in different ways for boys and girls.

In a survey interview, bias may be thought of as a factor that will cause the interviewee to give an answer that deviates from the "truth"—for example, to underreport the number of cigarettes smoked or to report a favorable opinion on a topic owing to social desirability. In a qualitative interview, however, the idea of a "truthful" response may have less meaning and bias should be interpreted rather differently. Here, interview bias has more to do with the way in which the interviewer's identity and behavior may in some sense influence the nature of the data collected but not necessarily in terms of their truth or falsity.

Normally, bias is something to be avoided. However, an apparent bias in the way that questions are asked may sometimes be a conscious strategy. This approach is often used in the wording of interview questions, but it can also be applied to interviewer behavior. For example, if participants appear reluctant to express a view that may be seen as socially unacceptable, the interviewer may deliberately use tone of voice or body language in such a way as to elicit such views. This does not so much *create* a bias as *counteract* a bias that already exists.

Jackie Waterfield

See also Interviews; Selection Bias; Survey Methods; Surveys

Further Readings

Fowler, F. J., & Mangione, T. W. (1990). Standardized survey interviewing: Minimizing interviewer-related error. Newbury Park, CA: Sage.

Groves, R. M., Fowler, F. J, Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (2nd ed.). Hoboken, NJ: Wiley.

Loosveldt, G. (2008). Face-to-face interviews. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), International handbook of survey methodology (pp. 201–220). New York, NY: Lawrence Erlbaum.

Nicole Mittenfelner Carl Nicole Mittenfelner Carl Carl, Nicole Mittenfelner

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

Interviews

Interviews

872

877

Interviews

Interviews are used prominently in naturalistic and qualitative research because they offer opportunities to collect data that are contextualized and individualized. Although interviews are often considered a hallmark of qualitative methods, it is important to note that they are also used in quantitative and mixed-methods research approaches. Depending on the study's research questions and focus, the goals of interviews vary. Interviews can be used to understand individuals' personal experiences, opinions, and perspectives related to an event or phenomenon. Researchers may examine how individuals' experiences compare and contrast to other participants' perspectives and/or prior research. In this regard, a researcher may use interviews to explore what is shared between participants and what may account for similarities in experiences as well as what is unique and different and what mediates or accounts for the range of experiences.

Interviews, as a form of data collection, are typically used when the goals of the research questions and study aim to understand how participants understand events and phenomena, develop detailed and contextualized descriptions of individuals' perspectives, integrate the perspectives of different participants, and describe participants' experiences and realities holistically. Because interviews help educational researchers understand individuals' lived experiences as well as the range of and variation in individual experiences, opinions, and perspectives within a group or about a phenomenon, they are an important data source in educational research studies. Interview data can help educational scholars and practitioners understand schools and schooling from a variety of perspectives by providing data that are in-depth, individualized, and contextualized. This entry describes the ideal characteristics of interviews, explains the different types of and approaches to conducting interviews, provides an overview of important considerations for constructing interview instruments and conducting interviews, and discusses the processes of recording and transcribing interviews. The entry concludes by describing important considerations for using interview data.

Ideal Characteristics of Interviews

To develop interview data that are contextualized and individualized, researchers should have an understanding of certain characteristics of interviews. At the heart of these characteristics is the recognition that interviewing, like qualitative and naturalistic inquiry, is interpretative. Interpretivist research means that there are multiple realities and not universal truths. In this regard, researchers should acknowledge the relational, subjective, contextualized, and temporal nature of interviews.

Because the researcher is the primary instrument in qualitative research, the interactions between the researcher and participants, from recruitment through data collection and write up, constitute a relationship. Respect for participants, their beliefs, and their experiences should be prioritized. The researcher should also consider and address power dynamics that may manifest in this relationship. As a part of the interpretivist paradigm, interviews should be acknowledged as subjective and not objective truth because the subjective realities of the researcher and the participants impact the questions and data. It is important that researchers understand that multiple contexts shape individuals' experiences and that they try to be as nonevaluative as possible while still recognizing that biases shape and inform all aspects of the research. These biases should be engaged with through reflexive practices in which researchers attempt to systematically assess the impact of their identity and subjectivities on the research. By identifying and reckoning with biases, researchers can try to resist imposing these through the use of evaluative language and nonverbal communication during interviews. However, researchers conducting interviews are not trying to be objective, but, rather, they recognize that the subjectivity of the researcher(s) mediates all aspects of how data are collected, analyzed, and interpreted.

To develop contextualized and individualized data from interviews, questions should be specific to participants' experiences and responses rather than generalized. Part of this involves recognizing the partial and temporal nature of interviews, and that they do not represent the entirety of participants' experiences. To conduct rigorous qualitative interviews that pay close attention to how interviews are relational, subjective, contextualized, and temporal, researchers should reflect individually and with others about how they can best acknowledge and address ways that they may influence the data.

Approaches to Interviewing

Although there are different approaches to interviewing, which are described in this section, the purpose of interviews remains similar in that researchers use interviews to gather focused, individualized information directly from research participants through dialogue. The different types of interviews include structured interviews, semistructured interviews, unstructured interviews, informal interviews, and focus groups. The type of interview varies depending on the methodological approach (i.e., action research study, case study, ethnography, etc.), research questions, and specific goals of a study. Research instruments refer to the tools that researchers develop and use to collect data. Interview instruments, also called protocols, are organized or structured in different ways depending on the type of interview. For example, an interview instrument may have an ordered list of questions or prompts that researchers use to guide the interview. The extent to which this list is detailed or not detailed and the order in which questions are asked depends on the type of interview. Common types of interviews are reviewed in the following sections.

Structured Interviews

Structured interviews follow a survey approach to interviewing in which the exact same question is asked to all participants in the exact same order. Structured interviews are often used in quantitative research, as the responses to structured interviews are often predetermined (also called fixed or close-ended responses) ones from which participants select a response. These types of questions are used in statistical research, as they allow for statistical comparison between groups and subgroups. If the structured interviews contain open-ended responses, they may be used in qualitative research and considered a qualitative questionnaire. However, structured interviews are not typically used in qualitative research, as they do not yield the individualized and contextualized data that qualitative interviews strive to gather.

Semistructured Interviews

In semistructured interviewing, similar questions are asked across study participants, but all participants are not asked the exact same questions in the same order. Thus, in semistructured interviews, the interviewer tends to ask individualized, follow-up questions during the interview. Semistructured interviews tend to use a semistructured interview instrument/protocol in which questions are listed on an interview guide, but the questions are not always asked in the exact same order, and participants’ responses are open-ended, meaning that there are no predetermined answers from which to select. Researchers also often include potential follow-up questions, called probes, on the interview instruments. These questions may or may not be asked depending on participants’ responses. The majority of qualitative interviews tend to follow a semistructured approach.

Unstructured Interviews

Unstructured interviews, which are also called in-depth interviews, adopt a process that allows for interviews to be highly inductive and specific to each participant’s experiences. These interviews do not follow a prespecified list of questions on an instrument, but researchers may note themes or topics that they want to address. Ethnographic studies often use unstructured interview processes.

Informal Interviews

Informal interviews are not typically planned in advance and occur while researchers are conducting research at a research setting (also referred to as conducting field work) with which they are familiar. Informal interviews often occur in ethnographic research that involves prolonged contact with participants and immersion in a setting. Informal interviews, which may resemble casual conversations with participants, are primarily recorded as jottings and then developed into field notes and are an important source data in participant observation studies.

Focus Groups

Focus groups are also referred to as group interviews. Focus groups tend to follow a semistructured interviewing process, and the goals of focus groups are often to gather multiple perspectives about a specific topic. Depending on the amount of time devoted to the focus group and the goals of the group interview, the amount of participants may vary. Some researchers argue for larger focus groups (e.g., eight to 10 participants) so that multiple perspectives can be expressed and larger numbers of participants can be reached. Other researchers argue for smaller focus groups (e.g., four to six participants) so that all participants have adequate time to share their opinions and engage with each other in conversation, which can create more in-depth data.

In addition to the type of interview, researchers should consider a variety of other factors before they conduct interviews that are broadly referred to as sampling considerations. Researchers determine the number and characteristics of participants (often referred to as participant selection) who will be recruited to participate as well as the setting in which the interviews take place. These considerations, which include selection criteria for participants, are guided by the research questions and are important parts of a study’s research design. There are no set answers as to how many participants should be included in an interview study, as representative samples are not typically the goal of qualitative interviewing.

Constructing and Conducting Interviews

There are a variety of questions that researchers might include on an interview instrument. The types of questions and how they are organized are primarily determined by the guiding research questions and the study’s goals. Questions might be related to participants’ experiences and behaviors, including what participants do or have done; opinions and values about a specific topic, event, or phenomenon; emotional experiences and feelings; knowledge about an event, topic, or phenomenon; sensory experiences related to what individuals see, hear, touch, taste, and smell; and demographics and backgrounds.

Especially for educational research, the intended participants should be carefully considered as each data collection instrument is developed. For example, interviewing high school seniors necessitates a different set of questions and language than when interviewing school principals. In addition, considering who will be conducting the interviews is especially important. Considering the will be conducting the interviews is especially important. Considering the interviewer’s positionality, or relationship to the research setting and participants, is necessary when developing the interview instrument. For example, a researcher might consider how and if the interviewer is an insider, outsider, or somewhere in between as well as how this positionality might influence the data collected as well as the relationship between the interviewer and the participant. All interview instruments should be tested, rehearsed, and revised prior to conducting the interview. Interviewing is a skill that researchers develop, and becoming a good interviewer takes time and practice. Most importantly, the goal is to listen carefully to participants. Once data collection has formally begun, instruments may be adjusted based on formative, or on-going, data analysis.

When conducting an interview, interviewers should be as nonjudgmental as possible. Even when researchers share similar perspectives with participants, bias can be reflected in nonverbal communication and affirmation. Depending on the research methodology employed, researchers may be more acquainted with participants, for example, in ethnographic and practitioner research studies. In these situations, the opinions and viewpoints of researchers might be expressed. In other situations, such as a one-time interview with a variety of participants, the interviewer’s perspectives may not be relevant. As previously described, this does not mean that the researcher or the data are neutral; this refers to how the interviewer engages with participants during interviews; and this varies depending on the type and goals of the interview, the study’s research questions, and the overall methodological approach. In all interview situations, good interviewers respect participants by listening closely to what they are saying and create a safe and nonjudgmental space.

Recording and Transcribing Interviews

Aside from informal interviews, interviews are ideally audio recorded, with participants’ consent. Prior to conducting an interview, researchers should engage participants in a process, often referred to as informed consent, by which they explain the research study and goals, describe the purpose and process of the interview (including all time requirements and expectations of participants), and give participants opportunities to ask questions. During this time, interviewers also ask if the participant consents to having the interview recorded. Researchers should determine in advance how they will record an interview and make sure that the recording device is working properly prior to the interview. Recording the interview makes transcription possible. Transcription is the

Recording the interview makes transcription possible. Transcription is the process of turning audio or visual data into textual form. When data are represented in a textual form, they become easier to sort, label, and annotate, which are important data analysis processes often referred to as coding.

It is important to note that transcribing interview data is not simply a technical, neutral process. Choices regarding whether to include or exclude pauses, restarts, filler words, and so on, are important, and the rationales behind these choices should be documented in final reports. Transcription can be a timely process if researchers transcribe the interviews themselves, and it can be an expensive process if researchers employ transcription services. How interviews will be transcribed and by whom are important decisions that researchers should consider during the research design process. An additional consideration is ensuring that individuals' privacy wishes are upheld and that data are kept secure. Having a plan for how transcripts and all other data will be organized and stored is also something that researchers should take into account. If researchers are unable to record an interview because of participants' wishes or other factors, notes that are taken during the interview should be fully developed as close to the time of the interview as possible.

Using Interview Data

Interview data should be collected and analyzed rigorously and systematically. The processes for analyzing data depend on the research questions and methodological approach (i.e., action research, ethnography, case study research, etc.); however, most qualitative interview studies follow inductive analytical processes in that researchers develop codes, concepts, themes, and theories from data that are contextualized and emerge from the data and engagement with participants. Inductive analysis can be combined with deductive methods in which data are analyzed by looking for ideas or themes that come from preexisting theories or prior research.

When analyzing and then writing up interview data, illustrative quotes are often used as a way to centralize participants' experiences and share ideas in their own words and ways of speaking. These quotes and examples are not anecdotal or randomly selected; they reflect important themes determined by researchers after careful and systematic analysis of all of the interviews (and related data) that aim to holistically reflect individuals' experiences. Thus, instances in which some participants' statements differ from other participants, as well as statements in which participants' views align, may be included. The examples and illustrations used are guided by the study's focus, goals, and research questions. Interview data are often combined with other forms of data, both qualitative and quantitative, and all data are systematically analyzed in relation to each other in ways that reflect the study goals and overall approach to data collection and analysis. The systematic analysis of different data sources and methods can enhance a study's validity by comparing and challenging findings and interpretations, and these processes are broadly referred to as analytical triangulation.

It is important that researchers transparently describe interview processes in final write-ups. This means describing the number of participants involved, the number of interviews conducted, who conducted them, how they were conducted (i.e., procedures, techniques, structure, length, etc.), how they were transcribed, and important contextual information about the participant and the interview setting(s). Because the researcher is considered the primary instrument in qualitative research, it is necessary that researchers reflexively examine how aspects of their identities and relationships with participants and/or a setting impact a study from the questions they ask to the participants selected. Qualitative interviews do not attempt to constitute objective data; they represent the subjective realities of participants, and the subjectivities of the researcher also shape the data. However, for qualitative interview studies to be rigorous, researchers must constantly monitor and address how they themselves influence the research process.

Findings from qualitative interview data are not generalizable to broader populations; generalizability is not the goal of qualitative interview studies and naturalistic research in general. By presenting readers with in-depth and contextualized data and interpretation, interview studies are applicable, or transferable, to other contexts. This is part of why context and transparency are so important in interviewing, as they help readers of such studies to appropriately apply and understand the findings. Conducting a rigorous interview study entails that researchers have a research design that uses the appropriate methods to answer the research questions while still allowing for flexibility, prioritizes participants' experiences, acknowledges issues of power inherent in the research relationship, attempts to conduct research that holistically describes the participants and/or phenomenon, and transparently describes the processes, challenges, and limitations of the data and study.

Nicole Mittenfelner Carl and Sharon M. Ravitch

See also Ethical Issues in Educational Research; Field Notes; Focus Groups; Interviewer Bias; Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Brinkmann, S., & Kvale, S. (2015). Interviews: Learning the craft of qualitative research interviewing (3rd ed.). Los Angeles, CA: Sage.

Josselson, R. (2013). Interviewing for qualitative inquiry: A relational approach. New York, NY: Guilford.

Kvale, S. (2007). Doing interviews: The Sage qualitative research kit. Los Angeles, CA: Sage.

Ravitch, S. M., & Carl, N. M. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: Sage.

Rubin, H. J., & Rubin, I. S. (2012). Qualitative interviewing: The art of hearing data (3rd ed.). Thousand Oaks, CA: Sage.

Sally M. Reis Sally M. Reis Reis, Sally M.

Nicholas W. Gelbar Nicholas W. Gelbar Gelbar, Nicholas W.

Iowa Tests of Basic Skills

Iowa tests of basic skills

877

880

Iowa Tests of Basic Skills

Designed in 1935 as a service project by researchers at the University of Iowa to enhance and improve educational instruction in the state, the Iowa Tests of Basic Skills (ITBS) is a widely known and used standardized achievement test. The Iowa Tests, recently renamed the Iowa Assessments, were originally designed to measure students' content knowledge of academic subjects. The ITBS measures academic achievement in 15 areas for students in kindergarten through Grade 8. The Iowa Tests of Educational Development measure academic achievement in nine separate content areas across Grades 9–12. The Next Generation of Iowa Assessments, released in 2016–2017, focuses on using assessment to inform instructional decisions by teachers, alignment with Common Core Standards, and an emphasis on student growth. The Iowa Tests are designed to measure the educational achievement of all students, as they serve as a fundamental assessment tool used to measure student content knowledge and skills across all areas of the curriculum.

This entry reviews various uses of the ITBS, its reliability and validity, and the forms and individual tests available for assessment purposes. The entry also highlights the revised ITBS, known as the Next Generation Iowa Assessments, and finally discusses advantages and disadvantages of using standardized tests.

Uses of the Iowa Standardized Assessments

The Iowa Tests (referred to as the ITBS in this entry) is a group-administered norm-referenced achievement test, designed to compare individual student achievement scores to those of a representative norm group of peers. The design of norm-referenced tests enables educational leaders, parents, and state policy makers to compare students with other students who are in the same norm group. For example, the available norms for the ITBS include districts of similar sizes, regions of the country, socioeconomic status, ethnicity, and type of school, in addition to a representation of students nationally. By comparing to a normative sample, the Iowa Assessments, like most standardized assessment tests, allow districts to gather data to improve, differentiate, and personalize instruction for all students.

The ITBS subtests include vocabulary, reading comprehension, mathematics, social studies, science, and other sources of information such as maps and diagrams. The Iowa Tests Assessments meet most states' requirements for an annual, nationally normed standardized test for assessing student academic progress in various content areas. The ITBS, like most other respected standardized achievement tests, meet certain psychometric markers and standards for reliability, validity, and analysis of the absence of bias, explained in the next section.

Validity and Reliability of the Iowa Tests

The ITBS have been proven to be technically sound, with many years of research on reliability and validity having been conducted by researchers at the University of Iowa as well as by other scholars who have used the ITBS to measure academic progress in various content areas. The ITBS include a broad variety of item types and are designed based on research regarding their validity and reliability.

Reliability refers to the amount of random variability in the scores produced by an assessment. Could a student take it again and achieve a similar or same score? Validity means that the test measures accurately what it is intended to measure. In addition, valid tests must be unbiased, meaning students must not be disadvantaged by where they live or by their individual or group characteristics. Valid assessment of achievement using the ITBS for a particular school is one that matches the school's education standards and learning outcomes. That is, the skills and content knowledge that contribute to success in the ITBS should be similar to the skills and knowledge that are taught in the school or district that has decided to use the Iowa Assessments. Whether the match is appropriate is something that can only be determined by a careful analysis of the test items early in the decision-making process.

The ITBS determine validity using a careful consideration of typical course coverage, instructional approaches, and recommendations of national curriculum and standards. The developers of the ITBS indicate that the content is carefully selected to represent nationally identified curriculum and current standards as well as to be inclusive for diverse populations. The researchers who developed the ITBS caution that the validity of the assessment process depends upon how the results of the tests are used. Ultimately, the validity of the ITBS depends on how the information is used to improve instruction and learning in the schools that use it. Like other lengthy standardized tests, the ITBS have been found to be highly reliable.

Forms and Individual Iowa Tests

Three forms of the ITBS are available. The Complete Battery offers educators the broadest range of testing available, and educators within a school or district can choose to administer as much or specific parts of the Complete Battery as they choose. The Core Battery offers the same level of diagnostic but focuses on basic achievement in the critical content areas of reading, language arts, and mathematics. The Survey Battery is a quick screening instrument, typically used when time is a concern. The current ITBS include Levels 5–14 for students in Grades 2–8. All early-level tests are read aloud to students. These early Level 5–8 tests are administered to students from kindergarten through second grade (K–2), while Levels 9–14 are administered to students from 3rd grade through 12th grade. The individual tests briefly explained in the following sections are included in the Complete Battery of the ITBS, but not all age levels include all tests.

Vocabulary

The vocabulary test assesses students’ breadth of understanding of general vocabulary and is a useful indicator of overall verbal ability. At the early levels, students hear a word, sometimes used in a sentence, and choose one of three pictures. At more advanced levels, students select the answer that they believe has the same meaning as the word. The word analysis section assesses students’ phonological understanding of word parts. Word analysis skills increase to more complex word-building tasks in higher level tasks.

Listening

The listening section includes short scenarios followed by comprehension questions presented orally to assess literal understanding, how well students follow directions, and students' ability to make inferences, understand concepts and sequences, and predict outcomes.

Reading Comprehension

The reading comprehension section assesses students' abilities to read words in isolation and to use context and picture cues for word identification. Students also answer questions about a picture that tells a story and in higher levels demonstrate their comprehension of sentences and stories. As levels increase, students read various types of passages of increasing length and difficulty and in different narratives, such as poems, fiction, and nonfiction in the science and social sciences, and answer questions to assess comprehension.

Language

The language tests measure students' understanding of how language is used to express ideas, examining skills such as the use of prepositions, comparatives and superlatives, and singular–plural distinctions, as well as spelling, capitalization, punctuation, or usage. Lower level tests emphasize oral language, and written language is assessed in higher levels.

Mathematics

The mathematics tests, based on the standards of the National Council of Teachers of Mathematics, emphasize the ability to use quantitative reasoning and to think mathematically in a wide variety of contexts. Early tests assess students' knowledge of beginning math concepts, focusing on numeration, geometry, measurement, and problem solving using addition and subtraction. Higher level math concepts and estimation are a focus for older students as is comprehension of number properties and operations, geometry, measurement, algebra, probability and statistics, and estimation skills. Increasingly challenging assessments include multistep word problems with selection of appropriate methods to solve real-world math problems, some using data displays such as tables and graphs.

Social Studies

Social studies tests measure knowledge content based on thematic strands identified by the National Council for the Social Studies in areas such as history, geography, economics, and government and society. At higher levels, some questions focus on understanding of political cartoons, graphs, or charts in areas such as time lines or excerpts from historical texts.

Science

Science tests assess students' knowledge of scientific principles and information but also the methods and processes of scientific inquiry in accordance with the standards from the National Science Teachers Association. Assessments address areas such as scientific inquiry, life science, earth and space science, and physical science.

Sources of Information

Sources of information tests measure students' abilities to use and assess the helpfulness of sources of information such as maps, diagrams, tables, and charts. For younger students, these are tested using skills in alphabetizing and in using picture dictionaries, tables of contents, and maps. At higher levels, the ability to use maps, diagrams, and reference materials is tested.

The Next Generation of Iowa Assessments

According to information on the Iowa Assessment website, the Next Generation of the ITBS, known as the Iowa Assessments, was released in 2016–2017 and includes the ability to measure growth in achievement, using norming methods to estimate national performance and achievement trends and provide a measurement tool that will assess both strengths and weaknesses of both individuals and groups. Together, these techniques should be able to produce reliable scores that satisfy the demands of users as well as meet professional test standards.

The Advantages and Disadvantages of Using Standardized Tests

Both positive and negative aspects exist in the use of standardized achievement tests, such as the pressure that many teachers and administrators feel to ensure that students score well. Some teachers have felt pressure to teach to the tests used in their districts and states, especially during the No Child Left Behind era. Some teachers have reported spending more time teaching to the test than in trying to have students engage in enjoyable learning experiences. During the No Child Left Behind time, an unprecedented number of state and federal agencies used tests to measure student learning and impose sanctions based on these scores alone, including enabling parents to transfer their students to other schools. Accordingly, standardized tests, used poorly, can cause stress for both students and teachers. Standardized achievement tests like the ITBS have been used to make decisions about whether to promote or retain students in school and to identify learning disabilities, developmental delays, or other disabilities. The positive use of standardized tests in a much more narrow and reasonable way is to aid teachers in determining what their students know and don’t know, and how much students in a school or district are learning, when compared to other similar cohorts of students in other schools or districts. However, standardized tests should never be the only assessment used for student learning.

Standardized tests should be only one part of a comprehensive assessment system used to assess student achievement. Assessment based on student performance on real learning tasks should also be used to measure learning, with multiple criteria such as projects, essays, portfolios, content area assignments and assessments, observations of student work, criteria reference tests that are developed at the school or district level, portfolios, and even assessment of students’ creativity and interests. Standardized tests can reveal some things about students but not everything about students, particularly with regard to their creativity, ambitions, interests, talents in areas outside of academic performance such as the arts, and leaderships. Even when core academic learning is measured, standardized tests only measure what students know about the information that is included in the tests, but there is certainly much more information that they should or will want to learn in and out of school.

The ITBS are often used in research settings. They are group administered

The ITBS are often used in research settings. They are group administered, which make them less resource intensive than individually administered assessments. As previously stated, the ITBS have also been developed using the gold standard approaches for instrument development as well as documenting reliability and validity. As the ITBS are group administered, at the upper elementary grade level, all students must read the items, which may be a concern for students who are not strong readers, regardless of whether they have been formally identified with a learning disability. It may be important for researchers to use screening measures for reading, so that students with reading difficulties can have the test items read to them. Another concern about the use of the Iowa Assessments for research is the need for alignment between researchers' definition of academic achievement and the content assessed by the Iowa Assessments. In other words, it is important to ensure that the Iowa Assessments are measuring the content needed by the researchers. If researchers are looking for a global measure of academic achievement, the Iowa Assessments are a good choice.

Sally M. Reis and Nicholas W. Gelbar

See also Achievement Tests; Reliability; Standardized Tests; Validity

Further Readings

Dunbar, S. B., Welch, C. J., Hoover, H. D., & Frisbie, D. A. (2011). Iowa Assessments, Form E. Rolling Meadows, IL: Houghton Mifflin Harcourt Riverside.

Iowa Assessments. (n.d.). Retrieved September 11, 2016, from https://itp.education.uiowa.edu/ia/default.aspx

Iowa Testing Programs—College of Education—The University of Iowa. (n.d.). ITBS research guide. Retrieved September 12, 2016, from https://itp.education.uiowa.edu/ia/ITBSResearchGuide.aspx

Welch, C. W., & Dunbar, S. (n.d.). Measuring growth with the Iowa Assessments™ (Rep.). Retrieved http://www.hmhco.com/~/media/sites/home/hmh-assessments/assessments/iowa-

assessments/pdf/iowa_measuring_growth_with_iowa_assessments.pdf

David Torres Irribarra David Torres Irribarra Irribarra, David Torres

Ipsative Scales

Ipsative scales

880

882

Ipsative Scales

Ipsative Scales are person-centered scales designed to assess two or more attributes simultaneously through comparisons that produce an intraindividual profile of the relative strengths of those attributes. Generally speaking, under Ipsative Scales, respondents “distribute points” across the properties that are being assessed, such that obtaining a high score in one or more of the assessed properties necessarily means lower scores in other attributes. In this way, the scores that individuals obtain in a given attribute are dependent on the rest of their scores in all the other attributes that are being simultaneously considered. In 1944, Raymond Cattell coined the term *ipsative* from the Latin *ipse*, meaning “himself,” to refer to scales in which individuals’ scores on an attribute were assessed relative to their scores on other attributes. This entry reveals the characteristics of Ipsative Scales and then describes its advantages and disadvantages.

Characteristics of Ipsative Scales

Formally, a scale is considered fully or purely ipsative when the sum of the scores of all the assessed attributes equals a constant for all the respondents. As a consequence, all respondents have the same mean score across all the attributes, and the scores on each measured attribute can be understood as deviations from that mean.

Ipsative measures are contrasted in the literature with normative measures that attempt to assess each of the attributes of interest separately, allowing in principle the possibility that a respondent would score high or low in all of them.

The person-centered focus of Ipsative Scales makes them suitable only for intraindividual comparisons as opposed to interindividual comparisons.

Ipsative Scales can be constructed through the use of specific item formats, such as forced-response items or ranking items. Ipsative forced-response items can assess the relative preference of the respondents by, for instance, presenting them with multiple options—each associated with one of the attributes being assessed—and asking the respondents to pick among them the one that represents them the most. Ranking items can produce ipsative scores by asking respondents to assign a ranking to all the options presented to them in a question, prompting them to weigh the relative strengths of the attributes represented by each option. Additionally, ipsative scores can be obtained through the use of traditional item formats, such as Likert-type scales, by centering a set of Likert-type items that assess multiple attributes on the mean score for each person (i.e., subtracting the person mean from each item), generating ipsatized data from items originally conceived to produce normative measurements.

Advantages and Disadvantages of Using Ipsative Scales

Ipsative Scales are most often used for the purpose of reducing or controlling self-report bias due to social desirability or halo effects, among others. The idea is that Ipsative Scales can be more robust to these kinds of biases by, for instance, forcing respondents to choose among options with similar levels of social desirability or by asking respondents to rank among multiple options, therefore eliminating the possibility of assigning high scores to all options. In sum, the dependency among the scores of multiple attributes is considered an advantage from this perspective, as it can be used to counter or eliminate certain response biases.

Notwithstanding the potential advantages associated with the reduction of biases that affect self-report assessments, the use of Ipsative Scales is associated with some disadvantages. A first issue to consider is related to the context of application in which it makes sense to use Ipsative Scales. The fact that they are designed for intraindividual comparisons among attributes makes them ill-suited for usage in contexts in which interindividual comparisons are the focus, such as personnel selection processes. A second issue associated with the use of Ipsative Scales touches on the potential restrictions on the statistical analysis that can be conducted on data collected with them due to the artificially induced dependency conducted on data collected with them due to the artificially induced dependency among the scores in the assessed attributes. This dependency violates the assumptions of most traditional statistical methods used to analyze questionnaire data, which has led to debates over the conditions under which it is or is not reasonable to use methods such as multiple regression or factor analysis on data from Ipsative Scales or even whether it is possible to use those methods. It is worth noting that, since the 2010s, a number of efforts have been made to develop and apply more advanced statistical models to account for the characteristics of ipsative data, such as the use of a Thurstonian item response model, a Rasch ipsative model for multidimensional pairwise-comparison items, and the use of methods such as the closed geometric mean, nonparametric bootstrap test, and permutation test.

David Torres Irribarra

See also Attitude Scaling; Likert Scaling; Rating Scales

Further Readings

Baron, H. (1996). Strengths and limitations of ipsative measurement. Journal of Occupational and Organizational Psychology, 69(1), 49–56.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. Psychological Methods, 18, 36–52.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. Psychological Review, 51, 292–303.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. Psychological Bulletin, 74(3), 167.

Horst, P. (1965). Factor analysis of data matrices. New York, NY: Holt, Rinehart and Winston.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser:

The use of ipsative personality tests. Journal of Occupational Psychology, 61(2), 153–162.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. Journal of Occupational and Organizational Psychology, 77, 531–552.

van Eijnatten, F. M., van der Ark, L. A., & Holloway, S. S. (2015). Ipsative measurement and the analysis of organizational values: An alternative approach for data analysis. Quality … Quantity, 49(2), 559–579.

Wang, W. C., Qiu, X., Chen, C. W., & Ro, S. (2016). Item response theory models for multidimensional ranking items. In Quantitative psychology research (pp. 49–65). Springer International Publishing.

IQ

See Intelligence Quotient

IRB

See Institutional Review Boards

IRT

See Item Response Theory

Insu Paek Insu Paek Paek, Insu

IRTPRO

IRTPRO

882

885

IRTPRO

Item response theory (IRT) is used for scoring test takers, test score equating, test development, and computer-adaptive testing, to name a few of its purposes. For such applications, estimation of IRT models is necessary, and IRT model estimation in general requires complex estimation procedures. IRTPRO is a computer program that estimates the parameters of many popular IRT models and user-specified versions of IRT models that are suitable for categorical ordinal and nominal item response data (i.e., dichotomous, polytomous, and mixed responses). IRTPRO for Windows was developed in 2011. Compared to traditional IRT programs, IRTPRO provides an easy-to-use window user-interface for the implementation of popular IRT model estimation and is suitable not only for unidimensional but also for multidimensional IRT modeling with more recent estimation methods and model-data fit statistics.

In this entry, specific features of IRTPRO are reviewed, including IRT models, estimation of IRT models, person latent score estimation, assumptions of IRT models, and model-data fit indices. The user interface is also described, and information on obtaining IRTPRO is provided.

IRT Models in IRTPRO

As in conventional IRT programs, IRTPRO provides popular unidimensional models that are suitable for a test measuring a single construct. The item response function (IRF) in IRTPRO, which describes the relation between latent trait scores and an expected item score, has the cumulative logistic function form rather than the normal ogive function that has been popular in the past.

For unidimensional dichotomously scored item responses, the Rasch, the two-parameter logistic (2PL), and the three-parameter logistic (3PL) models are included in IRTPRO. For unidimensional polytomously scored categorical item responses, the graded response (for ordinal responses), the generalized partial credit (for ordinal), and the nominal (for nominal responses) models are available in addition to the Rasch family polytomous ordinal item response models, such as the rating scale and the partial credit models.

When a test measures more than a single construct, multidimensional IRT models can be used. Commonly observed multidimensionality includes a simple structure (or between-item multidimensionality), in which there are multiple sets of items and each item set measures a single construct such that an item loads only onto a single factor (e.g., a test battery that consists of item sets measuring correlated multiple constructs), and a bifactor structure in which an item loads always on both one primary (or general) factor and one specific factor (e.g., a reading test in which there are reading passages and a set of items are associated with each reading passage). Multidimensional versions of the aforementioned unidimensional models are available in IRTPRO. Also, a multidimensional model with a complex dimensionality (e.g., within-item multidimensionality in which some items load onto a single dimension and other items load onto a few dimensions) that does not follow the previously mentioned dimensional structures or a user-specified unidimensional and multidimensional models can be estimated through imposing model parameter constraints (equality and fixing as an arbitrary value).

Estimation of IRT Models in IRTPRO

A traditional estimation method for popular IRT models has been the marginalized maximum likelihood estimation with the expectation-maximization algorithm (MMLE-EM). IRTPRO is equipped with MMLE-EM. In addition, IRTPRO has a fully Bayesian Markov chain Monte Carlo and two other non-Bayesian model estimation methods, which are Metropolis-Hastings Robbins-Monro (MH-RM) and an adaptive quadrature with MMLE-EM. The latter two are non-Bayesian high-dimensional model estimation methods. MH-RM can handle high-dimensional models more efficiently than the adaptive quadrature method as the number of dimension increases. The fully Bayesian Markov chain Monte Carlo method can be used for unidimensional and multidimensional model estimations. To take full advantage of MH-RM or the MCME method, especially for high-dimensional models, it is recommended to fine-tune their especially for high-dimensional models, it is recommended to fine-tune their option values (rather than simply using the program’s default values) that are best suited to the data under analysis.

Another interesting feature in the IRTPRO estimation procedure is that a dimension reduction technique, which provides a more efficient estimation than the regular MMLE-EM, is available for a two-tier multidimensional modeling such as a bifactor model.

When MMLE-EM is used, IRTPRO provides four types of standard error estimation methods: supplemented-EM (S-EM), cross-product method, sandwich method, and M-step method. S-EM makes up for the weakness of the EM algorithm in that it uses the EM history and produces the standard errors. Cross-product method is easier to use and works more efficiently than S-EM, while S-EM could produce potentially more accurate standard errors when its options for the implementation are carefully chosen for the data at hand. The MH-RM model estimation method and the standard error estimation method of S-EM are not found in the past conventional IRT software. (A recent IRT software flexMIRT and an R IRT program “mirt” have these estimation options.) As usual in IRT and other latent variable modeling, person latent trait score or latent ability is assumed to follow a normal distribution. To facilitate the computational efficiency and convergence in estimating IRT models, IRTPRO provides options for item parameter prior distributions when non-Bayesian estimation is used. For the item slope parameter, it supplies a log-normal and a normal distribution. For the item intercept parameter, a normal distribution can be used as a prior distribution. When the 3PL model is chosen, one can choose either a β distribution for the pseudo-guessing (or lower asymptote) parameter or a normal distribution for the logit of the pseudo-guessing parameter as a prior distribution. Employing a prior distribution for the pseudo-guessing parameter is especially useful in the estimation of the 3PL model because the presence of the pseudo-guessing parameter makes the estimation of the 3PL model very challenging for small sample sizes. Adding the slope parameter prior to the 3PL model estimation in addition to the pseudo-guessing parameter prior is not uncommon in practice in the applications of the 3PL model for large-scale assessment programs.

Besides the IRT model parameter estimation, IRTPRO can be used for classical test theory item analysis (item easiness, *p* value, and item-total correlation) and for calculating reliability such as coefficient α.

Person Latent Score Estimation in IRTPRO

In a non-Bayesian approach for IRT model estimation as in IRTRO, person ability or latent score estimation is done with what is called the “divide and conquer” strategy. Item and person ability population parameters are estimated first and then person latent scores are estimated using the item and the population parameter estimates from the first stage, which are treated fixed for the estimation of person latent scores. Popular person trait score estimators are maximum likelihood, expected a posteriori, and maximum a posteriori, which is also called Bayes’s modal estimator. One can select either expected a posteriori or maximum a posteriori estimator in IRTPRO.

When an IRT model uses a (response) pattern scoring (e.g., in those non-Rasch family models), the one-to-one correspondence between the observed summed score and latent score is not observed. The one-to-one correspondence might be desired when using a non-Rasch family model such as the 2PL or 3PL model. For such an occasion, IRTPRO offers an option to calculate the summed score expected a posteriori that has the one-to-one correspondence between the latent score and summed score.

Assumptions of the IRT Models in IRTPRO

Often, IRT models are used in a confirmatory approach rather than an exploratory approach (e.g., in the simple structure dimensionality, one should specify which item measures that construct). If the researchers use a unidimensional model, they assume that a single construct is being measured in the test. Choosing a multidimensional model implies that more than one constructs or factors or latent traits are measured in the test. The first assumption in using an IRT model in IRTPRO is that a user should make sure of the number of dimensions that underlie data from the test.

Second, what is known as local independence must be assumed. Local independence means that items or any sets of items are not associated with each other after controlling for the specified (single or multiple) latent trait(s). In this sense, local independence can be understood as conditional independence of item responses on the assumed dimensions.

Third, the chosen IRF should be a good description of observed response behavior (i.e., appropriateness of the IRF). If data from a multiple-choice item behavior (i.e., appropriateness of the IRF). If data from a multiple-choice item test behave more like the 3PL model, using a simpler model would be a less accurate description of the data. The IRFs used for all models in IRTPRO are parametric, and they imply monotonicity in the latent traits (i.e., increasing expected item score as latent trait scores increases). When using IRTPRO, due to its parametric modeling, more than the monotonicity (i.e., the appropriateness of the specified functional form of the IRF) is required.

Model-Data Fit Indices in IRTPRO

IRTPRO provides an item-level fit statistic, S-X^2, and several overall model-data fit statistics, which are LD-X^2, M^2, and traditional information criteria such as Akaike information criterion and Bayesian information criterion. S-X^2 and M^2 are statistical significance tests for an individual item fit and an overall model-data fit. LD-X^2, Akaike information criterion, and Bayesian information criterion are descriptive model-data fit indices. LD-X^2 is specifically designed for diagnosing the extent to which the local independence assumption is violated. Akaike information criterion and Bayesian information criterion are useful for comparing nested or nonnested models. When a simpler model is nested within another more general model (i.e., imposing the constraints on some of the more general model leads to the simpler model), one can also conduct a traditional model comparison test such as the likelihood ratio test using the −2 log-likelihood values computed by IRTPRO. If the Bayesian Markov chain Monte Carlo estimation is employed, one can use the estimation results to conduct Bayesian style model-data fit investigations (e.g., posterior predictive checking).

User-Interface in IRTPRO

Compared to the past and present IRT software and from a (IRT) consumer point of view, a main benefit of IRTPRO is its ease in implementing IRT model estimation. Some well-known IRT software has user-interface capability, but the window user-interface in IRTPRO is more user-friendly and complete. Estimating popular IRT models such as 2PL, 3PL, and graded response is done only with the mouse point-and-click through a window user-interface. The experience of using IRTPRO for estimating popular IRT models is very similar to that of a well-known window user-interface statistical package, SPSS. Even for more advanced options such as item prior distribution selection, imposing model parameter constraints, starting value specification, or output file control, IRTPRO provides an intuitive window user-interface, with which users can select and specify what is intended with ease. This window user-interface in IRTPRO guides how to use the software much more efficiently for new users than if they were to learn a syntax-based program.

How to Obtain IRTPRO

IRTPRO (Scientific Software International) is a commercial software. The IRTPRO full version is available for free for a 15-day trial period, and its student version is freely downloadable, which is limited up to a test of 25 items, a maximum of 1,000 test takers, and three-dimensional model analyses.

Insu Paek

See also BILOG-MG; FlexMIRT; Item Response Theory

Further Readings

Han, K. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. Applied Psychological Measurement, 38, 486–498. doi:10.1177/0146621614536770

Paek, I., & Han, K. T. (2012). IRTPRO 2.1 for Windows (item response theory for patient-reported outcomes). Applied Psychological Measurement, 37, 242–252. doi:10.1177/0146621612468223

Scientific Software International. (2015). IRTPRO User’s Guide. Skokie, IL: Author.

Websites

Scientific Software International:
http://www.ssicentral.comsales@ssicentral.com

ISLLC

See Interstate School Leaders Licensure Consortium Standards

Lilli Japec Lilli Japec Japec, Lilli

Lars Lyberg Lars Lyberg Lyberg, Lars

ISO 20252

ISO 20252

885

886

ISO 20252

Educational research is based on surveys and other data sources. It is imperative that data are of high quality to best serve the research purposes. High quality can be achieved through various measures such as well-trained research service providers using gold standard methodology and quality control procedures. A prerequisite for high quality of products such as survey estimates and research analyses is that the processes leading to the final products are well designed and stable in terms of variability. One way to achieve stability is to use standards.

The International Standards Organization (ISO) develops and publishes global standards in all kinds of businesses and industries. ISO 20252 is a process standard for market, opinion, and social research that is relevant to educational research. The intention is that its implementation should stimulate continuous improvement of the research products and make it easier to compare the findings obtained by different research service providers. It can also replace the various existing national standards.

This entry gives the raison d’être for ISO 20252, provides a brief overview of its requirements, and concludes with a description of the certification and accreditation procedure.

Raison d’Être for ISO 20252

Market, opinion, and social research has increasingly become a global industry.

The number of projects aiming at comparisons across countries, regions, and cultures is increasing and so are the demands for high-quality data that can accomplish this. Standardization of certain processes can help in this endeavor if properly implemented.

This expanding industry needs various elements of control to make sure that service providers in different countries work in similar ways and that they perform certain activities and adhere to core values that will generate high quality. ISO 20252 is also supposed to help users and clients to assess the service provider.

ISO 20252 is developed so that it should sustain a number of years. New methods and techniques are applied on a continuing basis and therefore the standard is in principle free from guidance on methods and technologies. The standard should be applicable no matter what infrastructures are in place at various service providers.

The standard was developed by a worldwide group of professionals working within market and social research as well as official statistics. Standards have been available in many different organizations and fields such as the Code of Practice within the European Statistical System, European Society for Opinion and Marketing Research’s guidelines, and the U.S. Office of Management and Budget’s survey guidelines. ISO 20252 is the first global standard for survey and market research. All these standards overlap to some extent.

The Requirements

There are around 400 requirements in the standard, and the focus is on the statistics production process. The requirements cover all steps that are known to have a large impact on data quality or cost, such as sample design, questionnaire design, data collection, competence and training, monitoring of interviewers, the use of validation methods, and presentation of research results. The standard specifies *what* the organization should do but not *how*. For example, one requirement is that if the client or the service provider deems it necessary to test a questionnaire, then a pretest of the questionnaire should be carried out. The standard does not, however, specify how to test the questionnaire or which method to use.

The standard also applies to subcontractors unless the choice of subcontractors is beyond the control of the research service provider. The research service beyond the control of the research service provider. The research service provider should control the quality of the service provided by the subcontractor.

The standard has a client focus, and transparency and traceability in methods are important requirements. Checklists and templates are tools that are used to reduce unnecessary variation in an organization and therefore important when implementing a standard. The ISO 20252 also requires a research service provider to carry out regular internal audits in order to ensure compliance with the standard. Finally, the research service provider needs to have procedures in place for continuous improvement, and the senior management is responsible for reviewing and improving the process management system. For instance, all problems and complaints should be documented and rectified to prevent future occurrences.

An issue of controversy has been the fixed validation levels for interviewer monitoring and coding control. The main argument for fixed levels is the concern that some service providers would otherwise exert too little control or no control at all. The main argument against fixed levels is that resources might not be used in an optimal way.

Certification and Accreditation

A research service provider can use the ISO 20252 as a quality guideline to improve its processes. For some organizations, it is also important to have an ISO 20252 certificate in order to demonstrate to their customers that their products meet the requirements in the standard. To get certified according to ISO 20252, an external certification body must perform an audit. These external audits can vary in terms of number of audit days depending on factors such as whether the organization is already certified according to ISO 9001.

The external certification body that performs the audit can be accredited. This means that it has a formal recognition from an independent body that it operates according to international standards related to the certification process. Accreditation is not compulsory, so a research service provider can become certified by an external certification body that is not accredited.

Lilli Japec and Lars Lyberg

See also Accreditation; Auditing; Certification; Market Research; Survey

Methods; Surveys

Further Readings

Biemer, P., & Lyberg, L. (2003). Introduction to survey quality. Wiley.

European Society for Opinion and Marketing Research. (2007). International code on market and social research.

Organisation for Economic Co-operation and Development. (2011). Quality framework and guidelines for OECD statistical activities. Version 2011/1.

Allison Jennifer Ames Allison Jennifer Ames Ames, Allison Jennifer

Item Analysis

Item analysis

886

891

Item Analysis

Approaches to determine how well multiple-choice or true-false items are performing on educational or psychological assessments make up a suite of tools referred to as *item analysis*. Item analysis is an examination of item-level properties, and the tools are intended to aid in building assessments with an adequate level of reliability, contributing to the validity of inferences made from an assessment. The material presented here provides a basic overview of item analysis. Topics addressed include the purpose of item analysis, users of item analysis, the techniques involved, and interpretation of item analysis results.

What Is Item Analysis?

Educational and psychological assessments are used to measure a broad range of knowledge, skills, abilities (KSA), and other constructs. These assessments are used for both high-stakes and low-stakes decisions. Examples include professional certifications, university admissions, psychological testing, and formative feedback, among others. The delivery of these assessments is just as broad. Common modes include paper-and-pencil delivery, electronic scoring, online or computer delivery, and computer-adaptive formats.

Regardless of use or delivery, making inferences about an individual’s KSA and other constructs involves examining an overall summed score, or percentage correct, and assessing the validity of those scores. Making valid inferences from assessments means that the use and interpretation of the assessment lead to the proper conclusions about the individual’s KSA and other constructs. Part of making valid inferences is related to how well the items are performing related to the item’s intended purpose. This is because an individual’s item-level responses serve as more refined pieces of information than a single, overall summed score. Subsequently, if the items are measuring the KSA and constructs as anticipated, then decisions related to the assessment scores can be made with a certain confidence. If the items are not performing as intended, then those decisions cannot be made with the same confidence level.

Investigating the performance of items, and whether the items are measuring the information in which they were intended to measure, is the purpose of item analysis. Results from these analyses are used to determine the items that are able to contribute to, in an acceptable manner, the measurement of the individual’s KSA and other constructs. Additionally, item analysis determines that items might need to be revised or removed from the assessment. There are two general frameworks that can aid in an item analysis: item response theory and classical test theory. The focus here will be on item analysis under the classical test theory framework.

Why Is Item Analysis Done?

Item analysis serves several purposes. The first, already briefly introduced, relates to identifying items that contribute to the making of valid inferences from educational and psychological assessments. This will be called the *inferential stage*. Part of making valid inferences is related to the *reliability* of the assessment. Reliability is a measure of how stable an individual’s assessment scores are. Another way to think of reliability is *reproducibility*. If an individual were to take the assessment again, how similar the individual’s new scores would be to the first administration is the degree of reproducibility. If they are identical, the scores are perfectly reproducible and reliable. Such scores are free from measurement error. If assessment scores are not highly reproducible, they are not considered reliable and the inferences made will not be valid because they provide a measure that contains a large amount of measurement error. Reliability is therefore a necessary, but not sufficient, condition for validity. The result is that many components of an item analysis are designed to assess the degree of reliability of the scores.

Another purpose of item analysis relates to assessment construction, in which the objective is to provide an assessment, at a minimum length, which will yield useful information about the individuals. Often, this is accomplished through piloting a large number of items and selecting those with the most desirable measurement properties for the objectives of the assessment. This will be called the *developmental stage*, occurring prior to the inferential stage. Item analysis needs to be conducted at both stages. To see why, consider items that are developed and piloted with one population in mind. However, if that same assessment is used to make inferences about another population, the same measurement properties might not hold. Thus, it is a best practice to conduce item analysis at both stages.

Who Uses Item Analysis?

Two types of practitioners interested in item analysis have emerged from the previous overview. One group comprises those in the developmental stage, seeking to build an assessment or a battery of assessments. This group includes testing companies employing assessment developers, content experts, and item writers. Also included are smaller, in-house assessment developers and surveyors as well as educators creating assessment instruments. If an item is piloted in this developmental stage and found to be a poor contributor to the overall reliability of the assessment, it is often revised or removed from the potential item bank altogether. Another group interested in item analysis is one that seeks to use the assessments, already developed by the former group, to make inferences about a set of individuals. This group could include survey administrators, teachers, assessment directors, school psychologists, and a wide range of other individuals.

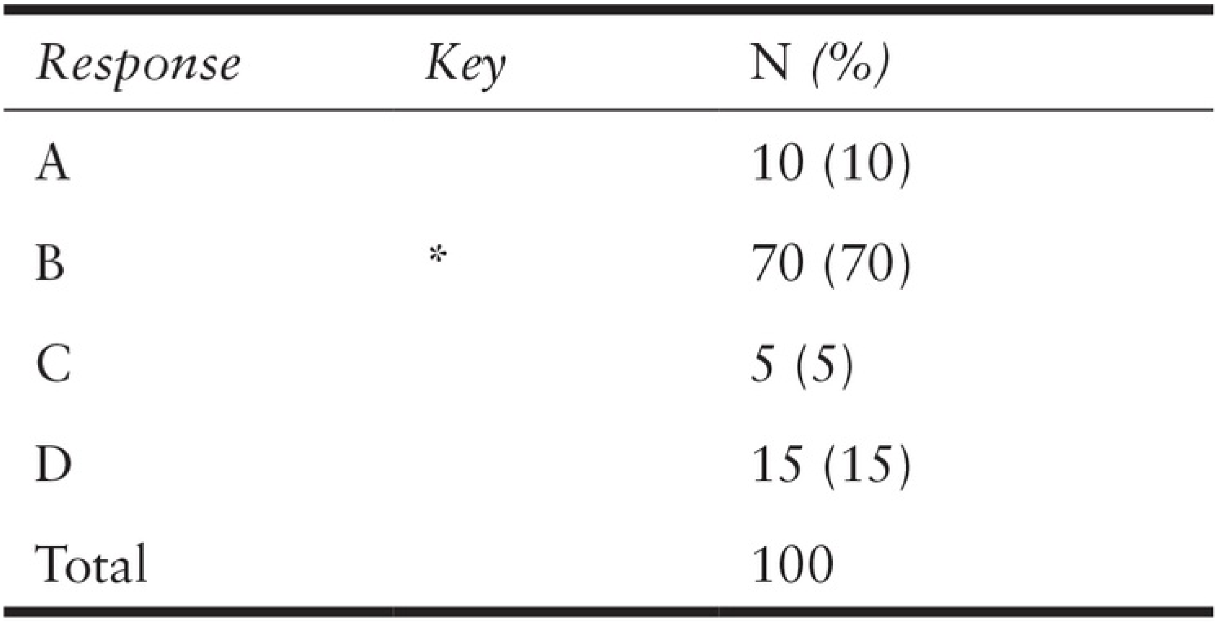
Item analysis plays a crucial role for both sets of practitioners. The techniques used by both of these groups overlap, and there is not one set of tools for the developmental stage that does not also serve a purpose to the inferential stage. For this reason, the following techniques cover a broad range of methods for use by practitioners in a wide variety of positions.

What Are the Techniques Used in Item Analysis?

To become familiar with the methods used in item analysis, it will be useful to introduce some terminology about the makeup of a multiple-choice item. These item types are comprised of a *stem*, which is the part of the item in which the question or problem is posed; a correct response, often referred to as the *keyed response*; and two or more incorrect responses called *distractors* or *foils*. Data in the *raw form* still shows the exact choice of an individual, such as “A” or “C.” In contrast, response data that have been *scored* have been transformed so that correct responses are assigned a 1, and all incorrect responses are assigned a 0. A *summed score* is then the sum of the scored items indicating the number of correct responses.

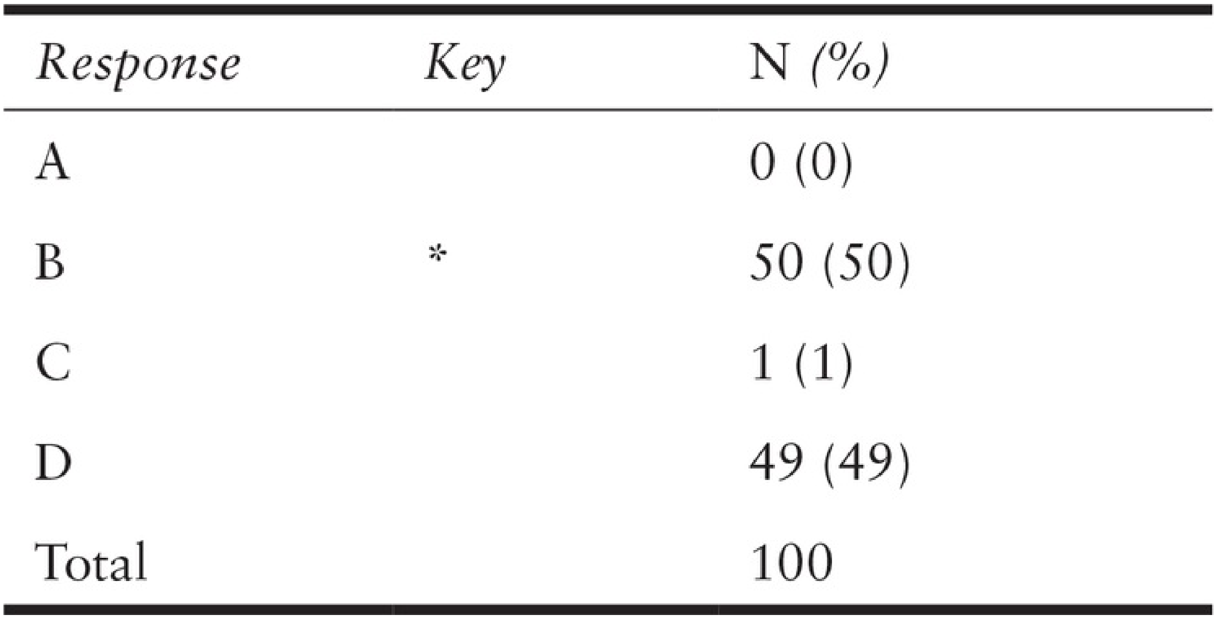
Distractor Analysis

A good first step in analyzing a multiple-choice assessment is to examine frequencies (counts) for each of the item choices, including all options, both correct and incorrect. This *distractor analysis* technique is performed before the assessment has been scored and responses are still in their raw form. For example, consider an example of a distractor analysis found in Table 1. The correct response is “B,” as indicated by the asterisk in the “key” column. The final column on the right indicates the number, and percentage, of individuals choosing each response option. For this item, about 70% of individuals have chosen the correct response and an approximately equal number of individuals have chosen the incorrect responses.



Response	*Key*	N *(%)*
A		10 (10)
B	*	70 (70)
C		5 (5)
D		15 (15)
Total		100

Now consider Table 2. The correct response is still “B,” but for this item, 49% of individuals chose distractor “D,” 1% chose “C,” and none chose “A.” Distractor analysis can be informative in analyzing the effectiveness of each response option. In Table 2, potentially ineffective distractors are found in response options “A” and “C” because almost no individuals selected these response options. When distractors are ineffective, there is a greater possibility that individuals will be able to choose the correct response simply by guessing, which impedes the making of inferences from the assessment, as the item-level information is not indicative of an individual’s actual KSA. If this occurs, the ineffective distractors can be rewritten or replaced. Practitioners will examine the stem and wording of the distractors to investigate why almost no individuals selected options “A” and “C” as well as why so many individuals chose option “D.” If the item is revised, it needs to go through the entire item analysis procedure again at the developmental stage before being used in an inferential stage.



Response	*Key*	N *(%)*
A		0 (0)
B	*	50 (50)
C		1 (1)
D		49 (49)
Total		100

The remaining item analysis techniques refer to methods applied after an item has been scored. That is, the responses are no longer in their raw form but consist of *dichotomous data*. Dichotomous is a term referring to two levels of a variable; in this situation, the data are either correct (score of 1) or incorrect (score of 0). The techniques that follow also assume an assessment has been scored correctly such that there are no *miskeys*, in which a 1 is assigned to an incorrect response or a 0 to a correct response. As such, a logical next step in item analysis is to ensure the key is correct and no items have been miskeyed.

Item Difficulty

After scoring an item, the simple average of the item can be used as an indicator of *item difficulty*. This measure is the proportion of individuals responding to an item correctly. Because it is a proportion, item difficulty is often termed the *p* value. The item difficulty value for item *i* can be computed via:

$$p\ \text{value}_i = \frac{(N_{\text{correct},i})}{(N)},$$

where N is the total number of individuals responding to the item and $N_{\text{correct},i}$ is the number of correct responses for the item.

Scores for *p* values range from 0 to 1, with values of 0 indicating no individuals responded correctly and *p* values of 1 indicating all individuals responded correctly. Smaller *p* values indicate more difficult items and larger *p* values indicate less difficult items, an interpretation that may seem counterintuitive at first. To see why this is so, return to the item in Table 1, in which 70 out of 100 individuals responded correctly. The item *p* value is 70/100 = 0.70. In contrast, examine the item in Table 2, in which 50 out of 100 individuals responded correctly for a *p* value of 50/100 = 0.50. Even though the Table 2 item has a lower *p* value, it is more difficult because a lower proportion of individuals responded correctly.

In the developmental stage, item difficulty is one indicator used to flag items for potential revision or removal from the assessment. Often, thresholds are set to flag items that are too easy or too difficult. For example, an item could be considered too easy if 95% or more of the individuals respond correctly, resulting in items with *p* values greater than .95 considered for removal or revision. These types of items are targeted because the amount of information provided, in regard to the KSA being measured, is quite low. For example, a very easy item (e.g., *p* value of .98) indicates that practically all individuals are responding correctly, irrespective of their level of KSA or other construct. That is, the item is not able to *discriminate*, or distinguish, between those possessing low and high levels of the construct being measured. The same holds true for very difficult items. With very hard items (e.g., *p* value of .02), nearly all individuals are responding incorrectly, and this item cannot discriminate between low-and high-level individuals.

Item Discrimination

In general, making inferences from assessments assumes that individuals scoring high on the summed score have a higher probability of responding to any given item correctly. If the opposite is observed, such that lower scoring individuals have a higher probability of responding to an item correctly, then the item is not discriminating well and could be measuring a KSA besides the intended one. Thus, the assessment needs comprise items that are able to discriminate between individuals in order for summed scores to serve their desired purpose.

One measure of item discrimination is that of the *item-total correlation* (ITC). The ITC is the Pearson product–moment correlation, ranging from −1 to +1, between the responses for the reported item and the individuals' total assessment scores. Values of the ITC near +1 indicate that the item has adequate discrimination. Negative values of the ITC imply the opposite—that lower scoring individuals are more likely to get the item correct. Values near zero indicate, regardless of their summed score, individuals are equally likely to respond correctly or incorrectly. Thus, ITC values nearer to +1 are most desirable, and near-zero or negative ITC values flag an item for revision or removal. Keeping in mind that item analysis practitioners are often tasked with selecting a group of items in order to maximize overall assessment reliability, choosing items that are most discriminating will work toward accomplishing this goal. Computation of the ITC for item *i* is accomplished via:

$$\begin{aligned}\mathrm{ITC}_i = {} & [(M_{\text{correct}} - M_{\text{total}}) \,/\, (\text{score variance})] \\ & \times [(p\,\text{value}_i)/(1 - p\,\text{value}_i)],\end{aligned}$$

where M_{correct} is the average assessment score for individuals who responded correctly to the item, M_{total} is the average assessment score for all individuals, and *score variance* is the variance of the summed score for all individuals. The more discrepant the M_{correct} and M_{total}, the higher the ITC. This indicates that the further apart these scores are, the better the item is at distinguishing between high and low performers.

Another measure of item discrimination is the *corrected* ITC (CITC). This is sometimes called the *item-remainder* or *item-rest* value. CITC values indicate something very similar to the ITC. However, the CITC is the Pearson correlation between the item and the summed score using all items except the item under consideration. The interpretation of the CITC is the same as the ITC. The difference is in the computation, as the CITC excludes the considered item's contribution to the total score, preventing a bias in the computation. However, when the assessment is sufficiently long, at least 25 or more items, including or removing the item in computation of item discrimination is not a major concern.

Item Variance

Item variance reflects the variability of the scored responses for each item and is directly tied to an item difficulty value. Smaller item variances indicate that individuals tended to respond similarly on the item, and larger item variances indicate that individuals tended to respond differently to the item. The following formula illustrates that the variances for item *i* is the product of the item's *p* value and one less the item's *p* value.

$$\text{Item variance}_i = (p\,\text{value}_i) \times (1 - p\,\text{value}_i).$$

Applying this formula to the item in Table 1, with a *p* value of .70, the item variance is (0.70) × (1 − 0.70) = 0.70 × 0.30 = 0.21. The item variance of the item in Table 2 is (0.50) × (1 − 0.50) = 0.50 × 0.50 = 0.25. More individuals responded similarly to the first item than for the second item, which is reflected in the item variance for the first item (0.21) being smaller than for the second item (0.25).

Item Reliability

Closely related to item variance and item discrimination is item reliability. By choosing those items with higher item reliability values, assessments can be constructed with a higher overall reliability. The item reliability for item *i* is computed via

$$\text{Item reliability}_i = \sqrt{(\text{item variance}_i) \times \text{ITC}_i}\,.$$

The item variance acts as a weight for the ITC in computation of the item reliability. If, for example, 2 items have identical ITC values, then they are equally discriminating. The item with the higher variance will be given a greater statistical weight in contributing to the overall measure of reliability of the assessment. However, as long as items with nonextreme difficulty values are selected (i.e., they would not be flagged for being too difficult or not difficult enough), the ITC and item reliability index provide similar information. Little is gained from using the item reliability over the ITC or CITC values unless items with difficulty values that are extreme are used.

Reliability

As previously noted, item analysis is primarily concerned with determining the items that contribute to strengthening an assessment's reliability. Reliability is a measure of how stable an individual's assessment scores are, with the idea that stable scores have less measurement error. Assessments should be constructed of items that contribute a minimal amount of measurement error to assessment scores, contributing to the overall reliability of the assessment.

Items with higher discrimination and smaller variances tend to contribute to higher overall reliability. Examining one commonly used measure of reliability (*coefficient* α) reveals why this is so. This class of measures includes Chronbach's α, Kuder Richardson 20 (KR-20), and Hoyt's analysis of variance. Values of coefficient α closer to 1 are desirable, with a commonly accepted heuristic of 0.80 and above designated as desirable. However, each assessment serves a different purpose, and some situations may require higher values of coefficient α, whereas others could find a lower value acceptable.

From the following formula, it can be seen that the smaller the item variances, in general, the higher the coefficient α.

$$\text{coefficient } \alpha = [I\,/\,(I-1)]\times \left[1-\left(\sum_{i=1}^{I}(\text{item variance}_i)\right) / (\text{total assessment variance})\right],$$

where *I* is the number of items on the assessment and indicates the sum of all *I* item variances. Coefficient α can also be computed using only item-level statistics. This version of the formula also shows that higher item discriminations yield higher values of coefficient α, in general.

$$\text{coefficient } \alpha = [I / (I-1)] \times \left[1 - \left(\sum_{i=1}^{I} (\text{item variance}_i)\right) / \left(\sum_{i=1}^{I} (\text{item reliability}_i)^2\right)\right].$$

To summarize the item analysis tools discussed here, desirable item properties include nonextreme item difficulty values, positive item discrimination, and higher item reliabilities. These properties will help build, at the developmental stage, assessments with higher overall reliability. This higher reliability can contribute to the body of evidence to support the making of valid inferences at the inferential stage.

What Computing and Software Resources Will Be Needed?

There are multiple computing and software resources available for practitioners to perform an item analysis. Several general purpose programs can perform an item analysis. These include, among others, SAS, SPSS, Stata, R, and Excel, each with multiple guides for how to perform item analysis in these programs. There are other software programs that are used specifically for assessment data analysis, including item analysis.

Allison Jennifer Ames

See also Classical Test Theory; Classroom Assessment; Coefficient α; Item Development; Reliability; Validity

Further Readings

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt Brace Jovanovich.

Osterlind, S. (1989). Constructing test items. Norwell, MA: Kluwer Academic

Publishers.

Penfield, R. D. (2013). Item analysis. In K. Geisinger (Ed.), APA handbook of testing and assessment in psychology. Washington, DC: American Psychological Association.

Smith, J., & Cody, R. (2014). Test scoring and analysis using SAS. Cary, NC: SAS Institute.

Wilson, M. (2005). Constructing measures: An item response modeling approach. New York, NY: Psychology Press.

TickMeng Lim TickMeng Lim Lim, TickMeng

Item Banking

Item banking

891

894

Item Banking

Traditionally, the task of preparing and maintaining assessment items for any institution of learning has been tedious and laborious. This is particularly true for institutions with a large student population. When the concept of *item banking* was first introduced, it generally referred to a database that facilitates the storage of a large number of assessment items to enable easy selection and retrieval. Little emphasis was placed on the process of test automation and item quality control. With the rapid advancement of computer technology, the notion of item banking has evolved from a test item storage system to a more advanced test management system. Besides the capability to store a large quantity of assessment items and their associated information systematically, an item bank also serves as a system that deals with assessment item entry, selection, quality management, print-ready examination paper generation, and item analysis. The type of assessment items that can be stored in an item bank may include essay-type items, multiple-choice items, and assignment task items. This entry highlights benefits and key features of item banking as well as the process of developing an item bank.

Benefits of Item Banking

It is a daunting and time-consuming task for schools to create new tests every year and, at the same time, to maintain the quality and consistency of the test standard. This same problem is faced in institutions of higher learning. As institutions of higher learning keep growing in size, the task of conducting formal assessments on students has become even more difficult and cost intensive. With the large number of students and numerous courses offered, the processes of preparing, conducting, and reporting with respect to both formative and summative assessments are major issues of concern. This is even more so for open and distance learning institutions in which learners are widely distributed and formal assessments such as examinations are not conducted in just one or two places but in many different learning environments.

A well-designed item bank system helps greatly in reducing the laborious tasks associated with test paper preparation and test item management. It enables the systematic classification and storage of test items and their associated data according to the predetermined criteria. It also facilitates the retrieval of items and generation of test papers based on the required criteria. Effective and efficient item banking should consist of the following functionalities:

- storage and organization of items according to subjects, topics, and cognitive levels of difficulty;
- autogeneration of a test specification table with reference to the required criteria;
- random selection of test items and autogeneration of print-ready test papers based on the test specification table generated;
- capturing of test scores and associated metadata for performing the necessary data analysis;
- autogeneration of a data analysis report on items used in a test or an examination;
- identification of poor items or faulty items based on the analysis for the purpose of improvement; and
- facilitating the reuse of test items in a controlled manner.

The structure of multiple-choice questions (MCQs) favors their storage and administration with an item bank. With more and more assessment examinations conducted in the form of MCQ item tests, it is worthwhile investing in a good item bank in which items can be deposited, added, used, enhanced, and reused while maintaining the reliability and validity of the test or examination paper.

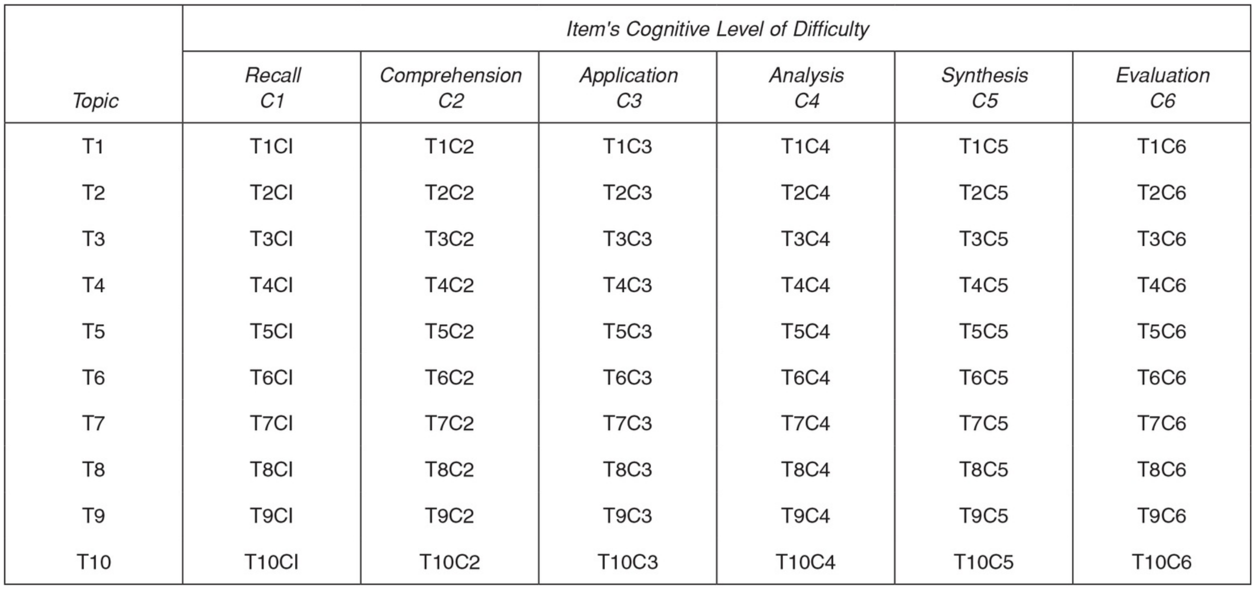
Important Features of Item Banking

The following are some of the important features of item banking:

Classification of Items

Normally, test items in an item bank are classified according to subjects, topics, and cognitive levels of difficulty. Figure 1 illustrates a common structure of classification. The figure in the form of a table shows that the items for a typical course are classified into 10 topics and six cognitive levels based on Bloom's taxonomy. In other words, we may imagine that there are 60 storage cells (10 topics × 6 levels of cognition). The test items that have undergone the thorough development, review, and moderation processes may be deposited into the respective storage cells.

Figure 1 Classification of Items



	Item's Cognitive Level of Difficulty					
Topic	*Recall* *C1*	*Comprehension* *C2*	*Application* *C3*	*Analysis* *C4*	*Synthesis* *C5*	*Evaluation* *C6*
T1	T1CI	T1C2	T1C3	T1C4	T1C5	T1C6
T2	T2CI	T2C2	T2C3	T2C4	T2C5	T2C6
T3	T3CI	T3C2	T3C3	T3C4	T3C5	T3C6
T4	T4CI	T4C2	T4C3	T4C4	T4C5	T4C6
T5	T5CI	T5C2	T5C3	T5C4	T5C5	T5C6
T6	T6CI	T6C2	T6C3	T6C4	T6C5	T6C6
T7	T7CI	T7C2	T7C3	T7C4	T7C5	T7C6
T8	T8CI	T8C2	T8C3	T8C4	T8C5	T8C6
T9	T9CI	T9C2	T9C3	T9C4	T9C5	T9C6
T10	T10CI	T10C2	T10C3	T10C4	T10C5	T10C6

Item Analysis

Item analysis is an important feature of an item bank. It is particularly useful for monitoring the quality of MCQ items. It helps to ascertain the quality of the items stored in the system. Item analysis also helps in the calibration of items for the purpose of reuse. There are basically two kinds of analysis: the classical model and the latent trait model. The basic classical model of analysis seeks to determine the difficulty level of any item and the ability of the items to discriminate better performing students from the weak students. The latent traits model analyses are based on the item response theory and the Rasch model, which are used to determine the psychometric properties of items and scales. Information received from item analysis reports enable the system user to identify items that need to be reviewed and improved or discarded.

Item History

It is important for an item bank to maintain a record of each and every item that has been used and administered. The item history should also include metadata associated with each item, such as date and frequency of use, and an item analyses record. Such information will serve as a control factor when there is a need to generate a valid and reliable test paper from the item bank.

Item Bank Development Process

In developing an item bank, it is crucial to have very thorough and careful planning to ensure that the item bank serves its purpose of managing assessment items effectively and efficiently.

There are a number of important steps in the design and development of an item bank: (a) system requirement analysis, (b) system design and development, and (c) user acceptance test.

System Requirement Analysis

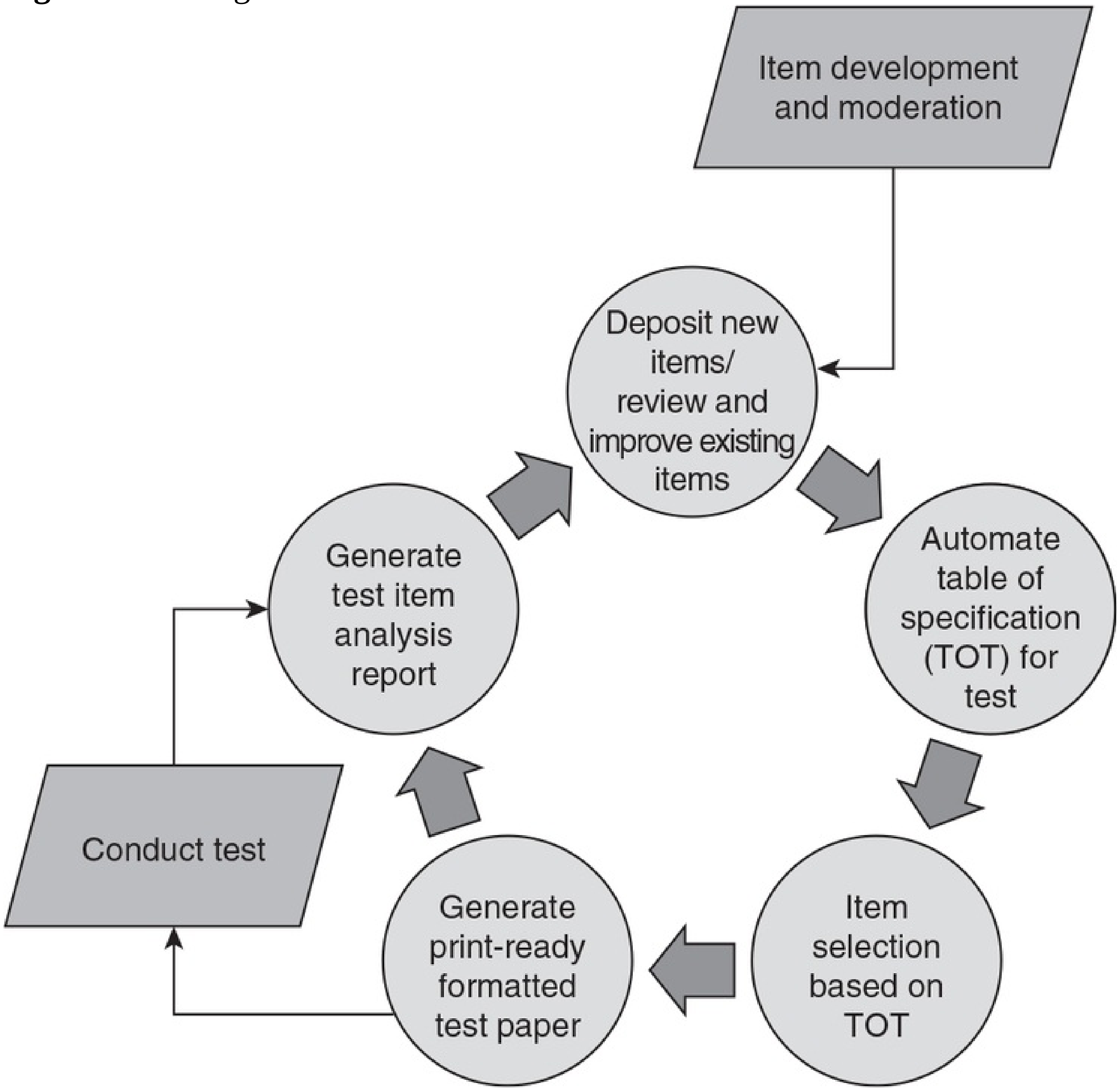
This is the process of determining the requirements expected of the system. In simple terms, it means to list out clearly what the system can do. The list of the expected requirements of the item bank system by the institution of learning serves as a contract between the institution and the developer of the system. In analyzing the requirements, the following key questions need to be addressed:

- How many assessment items can the system store (essay type, MCQ items, assessment items, etc.)?
- Can the system store an unlimited number of items for each type of item?
- What are the aspects of security that need to be considered?
- Should the system be a web-based system or otherwise?
- What kind of item analysis should be included?
- How should the items within each type be categorized?
- What are the different types of reports that can be generated from the item bank system?

System Design and Development

To ensure the usability and user-friendliness of the item bank system, the system developers need to work closely with the educators who are familiar with assessment processes. Figure 2 shows the important processes associated with the use of an item bank.

Figure 2 Banking Processes


Item development and moderation
Deposit new items/ review and improve existing items
Automate table of specification (TOT) for test
Item selection based on TOT
Generate print-ready formatted test paper
Conduct test
Generate test item analysis report

The course planners, course coordinators, and subject matter experts should be involved in the system design process and work closely with the system developers to ensure that the system is designed and developed to the needs of developers to ensure that the system is designed and developed to the needs of the users.

User Acceptance Test

Like any other system development process, an item bank that is developed needs to undergo a very stringent user acceptance testing process to ensure that the system works well and is able to manage the required tasks that it is intended to perform. As the name implies, the users and the stakeholders should be the ones involved in the testing process. Issues or errors identified should be reported immediately to the developers so they can be fixed.

Future Considerations

Currently, the use of an item bank may not appear to be a necessity to some institutions. But item banking is certainly an important step toward online testing and testing on demand if learning institutions are considering moving toward offering flexible entry and exit for their future potential learners.

TickMeng Lim

See also Item Analysis; Item Development

Further Readings

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. Quality of Life Research, 16(1), 133–141.

Downing, S. M., & Haladyna, T. M. (2006). Handbook of test development. Erlbaum.

Allison Jennifer Ames Allison Jennifer Ames Ames, Allison Jennifer

Richard M. Luecht Richard M. Luecht Luecht, Richard M.

Item Development

Item development

894

898

Item Development

Creating fair, valid, and reliable assessments, at a minimally sufficient length to yield useful information about the individuals, requires the use of high-quality items. What constitutes appropriately high quality depends on the content being measured, assessment stakeholders, and the purpose of the assessment. All these considerations must be defined and clearly articulated before item writing can take place. The process of creating these items with desirable properties is termed *item development*. This entry provides the basic concepts necessary to understand the item development process. Particular attention is given to identifying item and assessment purpose, item types, response space, assessment blueprints, and item specifications.

Validity and Reliability for Item Development

An end goal of assessment is to make valid and reliable conclusions about individuals and groups in order to provide formative or summative feedback and to drive learning improvement. Because the principle of validity and reliability provides the foundation for the purpose of item development, it is important that item developers become familiar with these concepts. Valid inferences about an individual’s knowledge, skills, abilities (KSA), and other constructs require that the use and interpretation of the assessment scores lead to the proper conclusions about the individual’s KSA. These valid inferences are related to how well the items are performing with regard to the item’s intended purpose. If the items are measuring the KSA and constructs as anticipated, then decisions related to the assessment scores can be made with a certain confidence. However, if the items were not designed, developed, or performing as intended, then those decisions about an individual's KSA cannot be made with much confidence. For example, a third-grade multiplication item purporting to measure a third grader's multiplication ability should not also require nonmultiplication vocabulary for a correct response to be given.

An item that required both multiplication and extraneous vocabulary would likely have considerable measurement error. More measurement error would imply the group of items are not measuring the KSA and other constructs as planned. One method to quantify measurement error is reliability, a measure of how stable an individual's assessment scores are. Stable scores have less measurement error. However, assessment reliability does not automatically guarantee the scores from an assessment will yield valid inferences.

Keeping the end goal in mind of making valid, reliable inferences, the question then becomes, how are items developed so that the inferences made can have these desirable properties? That is the primary purpose of this entry—to provide a general overview of item development, with emphasis placed not just on item writing tips but also on the end goal of valid inferences being made from the assessment.

Assessment Purpose

The first step in the item development process is to determine the assessment's purpose. The purpose guides item development by addressing questions such as: How will the assessment scores be used? Which stakeholders will receive the score reports? is the assessment intended for certification, licensure, accreditation, and so on? Will the emphasis be formative or summative? What consequences exist for the examinees, that is, is the assessment high stakes or low stakes? Are there time constraints? What other restraints, such as cost and other resources, are present? Who will be scoring the assessment, how will scoring be done, and what is the desired turn-around for providing feedback?

Answering these questions will have implications for designing the items and the entire assessment. Assessment length, general difficulty of the assessment items, testing conditions, and the type of information to be provided as feedback to examinees must all be decided upon prior to item design and development.

Without these key pieces of information, item writers have very little guiding them in the item-development process. To illustrate, consider certification and licensure assessments in which a *cut score* is set to determine those examinees who earn certification and those who do not. Psychometric principles require that many items should be targeted at, or near, the cut score to provide a high level of measurement information about examinees. Item developers working on such assessments can build items with a targeted level of difficulty in mind. Without being very clear on the assessment purpose, developers would lack the necessary guidance to produce the sufficient number of items at the intended target.

Content and Assessment Blueprint

Each item type must be chosen to align with the test purpose as well as the content to be covered by the assessment. Together, these aspects—purpose and content—guide the *fidelity* of the item. Fidelity refers to how well the item aligns with the KSA and other constructs the assessment aims to measure. For example, if the assessment is being used as an end-of-section grade for memorization of multiplication tables, use of a word problem may introduce an extraneous dimension (i.e., reading comprehension skills) that interferes with measuring the primary KSA. This type of item would exhibit low fidelity. This is often a compromise, however. Some item types are costly in terms of time to administer and score the item and may require other resources not readily available.

Once the test purpose is determined, more fine-grained aspects can be established. These include the content areas in the assessment blueprint. The blueprint is a list of specifications making up the KSA and other constructs that the assessment will cover and reflects the content and cognitive processes the assessment is purporting to measure. Specifically, the blueprint provides the number, or proportion, of items in each content area to make up the assessment. The proportions reflect the relative importance of each content area, as well as level of cognitive process, to the purpose of the assessment. The blueprint is then used to guide item development.

Item Types and Response Space

Purpose and content are critical in determining the type of item to be developed. If near-instantaneous, formative feedback is the intended purpose of an assessment, an essay item is inappropriate for use, as assigning a score to such an item, and providing adequate feedback, requires considerable time and effort. A multiple-choice item may be more useful, as examinees could even score the item themselves with proper instruction. Two general item categories—selected response and constructed response—do not have to be mutually exclusive on an assessment. However, careful consideration must be given to which type of item, or group of item types, is most appropriate for the intended use of the assessment scores. With assessment purpose and content in mind, item types can then be chosen for the development process.

Multiple-choice items are *selected response*-type items. They have a *stem*, which is the part of the item in which the question or problem is posed, a *keyed* (correct) *response*, and two or more incorrect responses termed *distractors* or *foils*. Individuals must then select from the *response space*, which includes the keyed and distractor responses, and choose the correct option. True/false-type items are another form of selected response, as the stem provides a statement, and individuals must select from the response space whether this statement is true or false. Other types of multiple-choice items include matching, complex options, and item sets, among others. For selected response items, data in the raw form still show the exact choice of an individual, such as “A” or “True.” In contrast, scored response data have been transformed, so that *correct responses* are assigned a 1 and *all incorrect responses* are assigned a 0 to produce dichotomous data. A summed score is then the sum of the scored items, so that the summed score is the number of correct responses.

Items can also have *constructed* or *created response* options that the individual generates rather than selecting from a set of provided responses. Examples of constructed response options include anecdotal evidence, demonstrations, discussions, essays, exhibitions, experiments, fill in the blanks, interviews, observations, oral examinations, performance pieces, projects, research papers, short answer items, and many others. These items are not typically scored with a traditional system in which 0 is given for an *incorrect response* and 1 for a *correct response*. Instead, a rating scale is generated (e.g., a rubric), and each item is scored along several dimensions to arrive at an overall score.

The *response space* for these constructed response items is much larger than for selected response items. More specifically, the response space represents the response to an item that will be used as the basis of inference. The response space for an item also includes how examinee responses will be scored.

Generally, it is the rubric creator or scale generator who implements the outcomes space, and considerable effort is put into defining the desirable outcome space as well as validating the rubric (or scale). Mark Wilson provides some guidelines for articulating a response space. It should be well-defined, finite and exhaustive, ordered, context specific, and research based.

There are both benefits and drawbacks to selected and constructed response items. Selected response items generally have a more objective scoring procedure in place. Either a student has selected the correct response or he or she has not. These item types are also less taxing in terms of the resources needed to score the items, and feedback can be provided to individuals more rapidly than with constructed response items. However, the selected response allows for the correct response purely by chance alone. In fact, with a true/false item, an individual has a 50% chance of a correct response just by guessing. In this instance, the level of measurement information is rather low because of the very small response space. This is one of the benefits of using constructed response items—that chance guessing alone is not likely to produce a correct response in a sufficiently large response space. However, the resources and subjectivity involved in scoring the item are much greater. The possibility for more measurement information to be obtained, though, is also greater, a considerable benefit of these item types.

Item Specifications

Similar to an assessment blueprint, item specifications provide detailed requirements for each item to be included on the assessment. These specifications include the item type, content, and cognitive areas covered, as well as related items (i.e., item sets), and other information to guide item writers. Item specifications are used to write multiple items that can be used interchangeably to ensure consistency across items. Item specifications can also increase the efficiency of item development. The number of items written depends upon whether the assessment will be used 1 time, how many individuals will take the assessment, and test security concerns. If a large number of items are written so that multiple forms can be constructed, the *item bank* will house multiple items of similar specifications.

One component of the information included in the item specification should be about the level of the KSA and other constructs coupled with the inference required to make conclusions about the KSA. Thomas Haladyna delineates five distinct behavior domain items: declarative knowledge, higher inference cognitive KSA, higher inference psychomotor KSA, lower inference cognitive KSA, and lower inference psychomotor KSA. Take, for example, a math word problem:

Mallory is making party baskets for birthday party favors. She has invited 60 people to the party. Cards are sold in bags of 20, hats are sold in bags of 10, and there are five candy bars in a package. How many of each should she buy so there are an equal number of balloons, horns, and candy bars in each basket and at least one card, hat, and candy bar in each basket?

Lower inference cognitive KSA, representing simple, easily observable outcomes, reflects a higher level of ability such as arithmetic in a problem-solving process. The math word problem example represents such a behavior: The problem-solving process leads to an easily observable result (a correct or incorrect response). For such a behavior domain, selected or constructed response items are useful and are able to be objectively scored.

Using the same word problem, consider that the behavior of interest is now related to the efficiency of the problem-solving process, a higher inference cognitive KSA. An individual could attempt several trial and error methods, an inefficient route, before eventually arriving at the correct answer. A multiple-choice item, or other selected response item, may not capture the efficiency of the problem-solving process, whereas a constructed response option will show the steps the individual took to solve the problem, and efficiency can then be scored using a rating scale.

More on Item Development

Once assessment purpose, content, and item specifications have been clearly articulated, item developers can begin to write the items. Editorial style, the consistent use of punctuation, formatting, citation, and grammar, must also be considered at this stage. This can be particularly difficult if multiple item writers are independently generating items. Clearly defined guidelines must be provided and adhered to, and it is a best practice to have an editor review all the items to unify the format and grammatical style. This holds true for the general assessment instructions, individual or group item instructions, and the items themselves.

Multiple-choice items have two main components: the stem and response options. Both should be brief, avoiding repetition. Other recommendations for multiple-choice items include having a clear and distinct answer, with distractors that are not unnecessarily tricky or vague. These distractors should also be reasonable, so as not to be eliminated too easily. Wording should be concise and unambiguous, usually in third person unless a first-person narrative specifically adds to the item's ability to measure the intended KSA.

When writing items, stereotypes and culturally specific references should be avoided. Similarly, slang and superfluous details should be kept to a minimum. These writing tips help ensure that the item is clear and doesn't introduce additional measurement error. Grade-level appropriate language should always be used, and archaic language should be replaced if not the direct object of inference. For example, "knave" is used in Shakespeare's *Romeo and Juliet*. Using knave in an item unrelated to the play may be inappropriate, but entirely reasonable in the context of an assessment on *Romeo and Juliet*. Many of the same principles for writing selected response item types also apply when writing the prompt for constructed response item formats.

Item Analysis for Item Development

Once items have been written, it is best to obtain an independent item review by content experts and other item writers. Field testing the item, often referred to as *pilot testing*, is recommended as a best practice step. Field testing is a useful tool to study an item's performance, allowing for analysis and focus groups in the developmental phases. If an item is found to perform poorly—at any stage of this process—the item can be revised (including the distractors, stem, correct response, or prompt) and reviewed again. It will often be the case that the item must be removed from the bank and new items written.

A good first step in item analysis of multiple-choice items is to examine frequencies (counts) for each of the item response options, including the correct and all incorrect options. This *distractor analysis* technique is performed before the assessment has been scored and responses are still in their raw form. This technique allows writers to gauge whether certain distractors are ineffective and whether the correct response is clear and distinct from the incorrect responses.

After scoring an item, the simple average of the item can be used as an indicator of item difficulty. This measure is the proportion of individuals responding to an item correctly. Item difficulty is one indicator used to flag items for potential revision or removal from the assessment due to an overly easy, or difficult, item. For example, an item could be considered too easy if 98% or more of the individuals respond correctly. These types of items are targeted for removal because the amount of information provided, in regard to the KSA being measured, is quite low. That is, a very easy item indicates that practically all individuals are responding correctly, no matter their level of KSA or other construct. Thus, the item is not able to discriminate, or distinguish, between those possessing low and high levels of the construct being measured. The same holds true for very difficult items. With very hard items, nearly all individuals are responding incorrectly, and this item cannot discriminate between low-and high-level individuals.

Item variance reflects the variability of the scored responses for each item and is directly tied to an item difficulty value. Smaller item variances indicate that individuals tend to respond similarly on the item and larger item variances indicate that individuals tend to respond differently to the item. Items with higher discrimination and smaller variances tend to contribute to higher overall reliability, a desirable property for an assessment.

Allison Jennifer Ames and Richard M. Luecht

See also Classroom Assessment; Item Analysis; Item Banking; Reliability; Validity

Further Readings

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt Brace Jovanovich.

Haladyna, T. (1997). Writing test items to evaluate higher order thinking. Needham Heights, MA: Pearson.

Osterlind, S. (1989). Constructing test items. Norwell, MA: Kluwer Academic Publishers.

Wilson, M. (2005). Constructing measures: An item response modeling approach. New York, NY: Psychology Press.

Christine E. DeMars Christine E. DeMars DeMars, Christine E.

Item Information Function Item information function

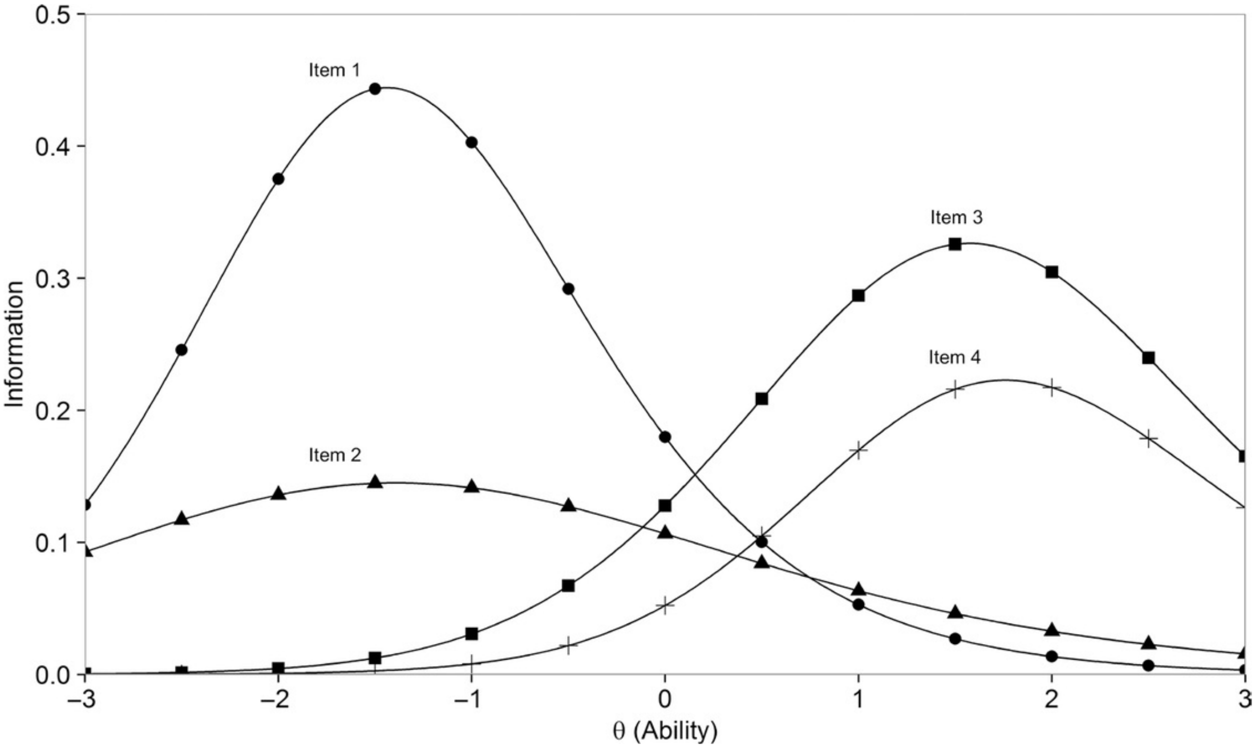
898

903

Item Information Function

In item response theory, the item information function is a measure of how much statistical information a test item provides. Item information is a function of θ, the ability, proficiency, skill, or trait measured by the examinee's responses to the test items. Figure 1 shows the information functions for four items. Items 1 and 2 are more informative at low values of θ than high values, and Items 3 and 4 are most informative at higher values of θ.

Figure 1 Information for four items. Items 1 and 2 have the same difficulty but Item 1 is more discriminating. Items 3 and 4 are more difficult than Items 1 and 2. Item 4 has the same parameters as Item 3, except that the lower asymptote is higher for Item 4.



The amount of information provided by an item, relative to θ, depends on the item’s parameters. The items in Figure 1 are dichotomous items (scored 0 or 1), so they have up to three item parameters (plus an examinee parameter) describing the probability of correct response:

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}},$$

where $P_i(\theta_j)$ indicates the probability of a correct response to item i from examinee j with ability θ_j, a_i is the item discrimination, b_i is the item difficulty, and c_i is the lower asymptote. D is an optional scaling parameter which either equals 1 (in which case it can be dropped) or 1.7; if D = 1.7 the a parameters will be approximately on the normal (probit) metric. The lower asymptote is the probability of correct response for examinees with very low ability. For example, if low-ability examinees perceive all of the distractors to be about as likely as the correct answer, the c parameter might be equal to random guessing. The subscripts are often omitted.

For the *two-parameter logistic* model, all lower asymptotes are zero, so the c parameter can be dropped from the model. For the *one-parameter logistic* (1PL) model, all item discrimination parameters are equal. The Rasch model is mathematically equivalent to the 1PL model; the *a* parameter is dropped from the model (all *a* = 1), and the item discrimination is displaced onto the variance of θ: Tests composed of more discriminating items have greater θ variance.

The peak of the item information function occurs at or just above the item difficulty, depending on the model. In Figure 1, Items 3 and 4 are more difficult than Items 1 and 2. The information function steepens as the *a* parameter increases. Comparing two items with the same *b* and *c* parameters but different *a* parameters, the item with the higher *a* parameter will have more information but in a narrower range of θ. In Figure 1, for example, *b* = −1.5 and *c* = .05 for items 1 and 2, but *a* = 1.4 for Item 1 compared to 0.8 for Item 2. Thus, Item 1 has more information for low θ values, but slightly less information than Item 2 for very high θ values. The *c* parameter dampens the information function, especially for hard items. Items 3 and 4 have the same *a* and *b* parameters (*a* = 1.2, b = 1.5), but the *c* parameter = .05 for Item 3 compared to .25 for Item 4. Thus, Item 4 provides less information.

Although the term *information* has a statistical, mathematical definition in this context, conceptually it corresponds to the everyday use of the term. If an item has lots of information near an examinee's θ value, than the examinee's response to that item helps estimate θ better. So if an item has a high *a* parameter and a *b* parameter near the examinee's θ, it is useful for estimating θ. But if the item is much too hard or much too easy, or simply not very discriminating, the response to that item does not provide much information about what the examinee knows or can do.

Mathematically, the information function is the negative of the second derivative of the log-likelihood (natural log of *P*(θ)) of the observed response (0 or 1 if the item is dichotomous). In item response theory, analysts typically use Fisher's information, which is the expected value of the information so it does not depend on the examinee's response. Typically, the term *information*, without any further modifier, indicates Fisher's information, not observed information, and for the 1PL and two-parameter logistic models, observed and expected information are equal. Although equivalent to the expected value of the 2nd derivative of the log of *P*(θ), Fisher's information can be calculated from *P*(θ) and its first derivative, without using the second derivative at all. For dichotomous items, one computational formula for Fisher's information is:

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)(1 - P_i(\theta))},$$

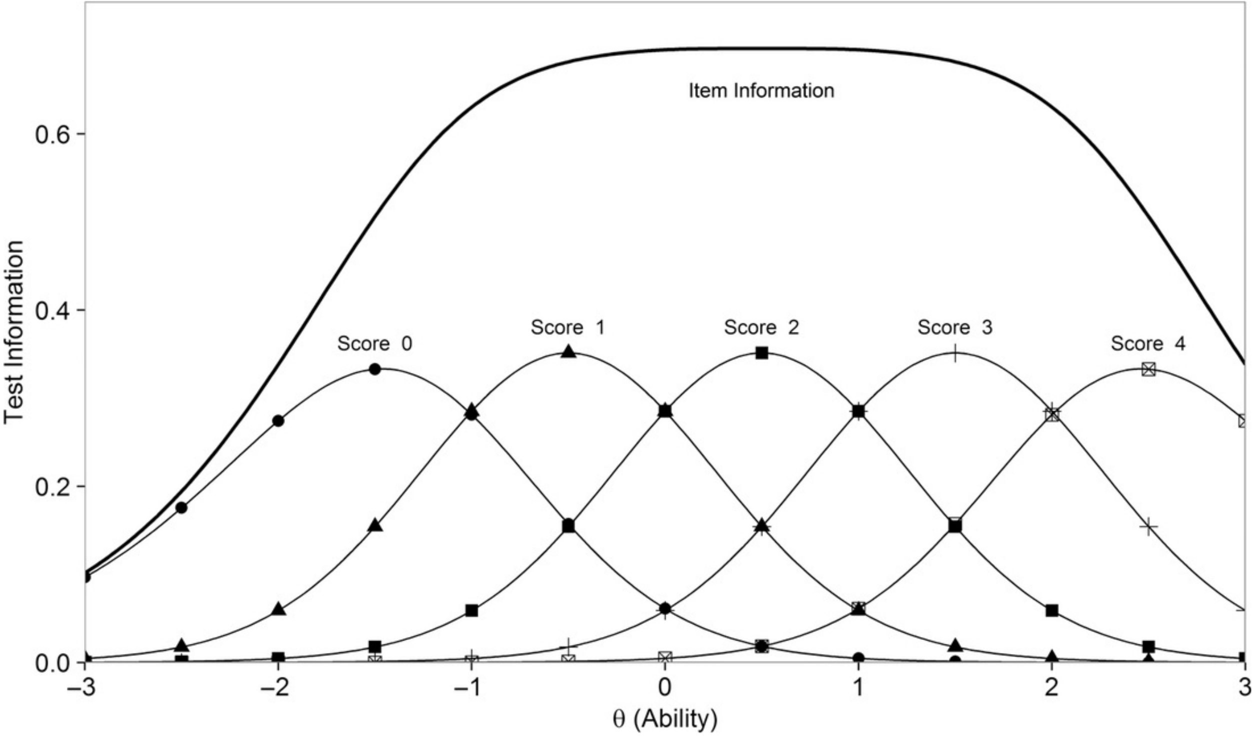
where $I_i(\theta)$ is the Fisher’s information for item *i*, $P_i(\theta)$ is the probability of correct response, conditional on θ and the item parameters, and is the first derivative of $P_i(\theta)$. For data that follow the three-parameter logistic, . The ability at which the item information function reaches its maximum is When $c = 0$, the latter term equals 0, so for two-parameter logistic or 1PL/Rasch data maximum information is obtained at $\theta = b_i$. Referring back to Figure 1, maximum information is reached at $\theta = -1.44$ for Item 1, −1.39 for Item 2, 1.57 for Item 3, and 1.76 for Item 4. Notice how the information peaks at a θ value somewhat above *b*, and the higher *c* parameter for Item 4 leads to a greater difference between *b* and the point of maximum information.

Equation 2 for dichotomous items is a special case of the equation for polytomous items (items with more than two score categories). For polytomous items,

$$I_i(\theta) = \sum_{k=1}^{K} \left[\frac{\left[P_{ik}'(\theta)\right]^2}{P_{ik}(\theta)} \right],$$

where *k* is the *k*th score category for item *k* (0 may be the first score category, etc.), *K* is the total number of categories, $P_{ik}(\theta)$ is the probability of scoring in category *k* of Item *i*, given θ, and is the first derivative of $P_{ik}(\theta)$. $P_{ik}(\theta)$ may be based on any polytomous item response theory model, including the graded response model, the partial credit model, or the nominal response model. Each category contributes a share to the information. The information share for category *k* is $P_{ik}(\theta)I_i(\theta)$, equivalent to , where $P''(\theta)_{ik}$ is the second derivative of $P_{ik}(\theta)$. Note that $P''(\theta)_{ik}$ does not appear in Equation 3 because the second derivatives sum to zero across categories, so $P''(\theta)_{ik}$ can be ignored when calculating the overall item information. If one treated as the category information share, the sum would still be accurate, but the individual category shares would be slightly incorrect. Figure 2 illustrates the information shares and item information for an item that follows the graded response model. Similar graphics could be drawn for other polytomous models. The information shares are depicted for each category, plus the item information is shown with a solid dark line. This item had 5 categories, with thresholds = (−1, 0, 1, 2). Thus, the information share for score = 0 peaked below the first threshold, the information share for score = 1 peaked between the first and second threshold, and so on.

Figure 2 Information for a polytomous item. The item information (dark line) is the sum of the category information shares.



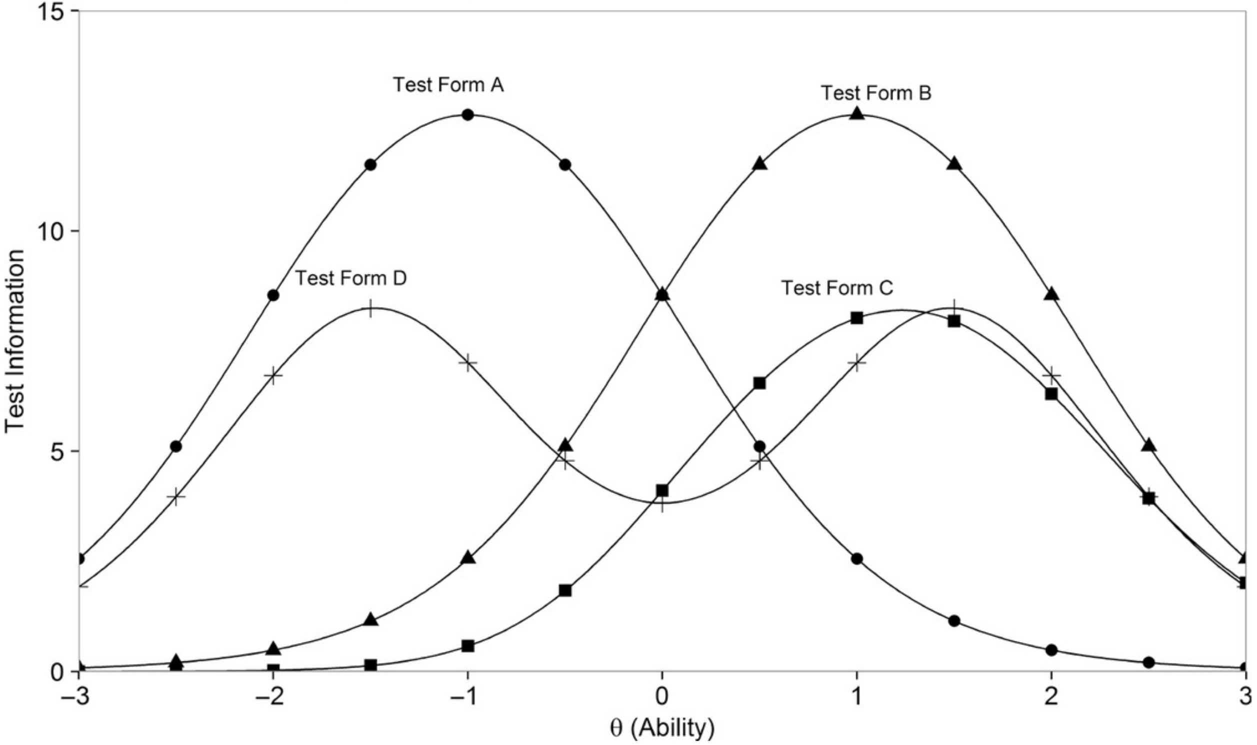
Item information is useful for computer adaptive testing. One of the most common ways of selecting the next item to administer is maximum information item selection. After the examinee has responded to several initial items, θ can be estimated and the next item is the item that provides the most information at , where is the estimate of θ (although in many cases the test developers also take into account item utilization and content balance). Another common item selection method, especially early in the test, is global item information. One measure of global item information is the mean information within a range around , or perhaps a weighted mean, with the information weighted by the examinee's likelihood function. Another index of global information uses Kullback–Leibler information instead of Fisher's information. The Kullback–Leibler information at θ_1 is a measure of how well the item differentiates θ_1 from

. This information is then averaged (integrated) over a range centered around .

Of course, item information is not the only criterion in selecting items in computer adaptive testing. Typically, content balance is taken into account and algorithms are used to avoid overexposing items to too many examinees.

The most common application of the item information function is to calculate the test information function, from which the standard error of θ can be estimated. The test information is literally the sum of the item information, so items may be selected to match a target test information function. Figure 3 shows the test information function for four different 20-item tests. For simplicity, all *a* parameters = 1.8 (the *a* was scaled with *D* = 1, so *a* would be just over 1 in the normal metric). Test A was an easy test, with item difficulties uniformly spaced from −2 to 0. Tests B and C were harder, with item difficulties uniformly spaced from 0 to 2. Test D had 10 items with *b* = −1 and 10 items with *b* = 1. The lower asymptote was 0 for all items on Tests A, B, and D, but 0.2 for Test C.

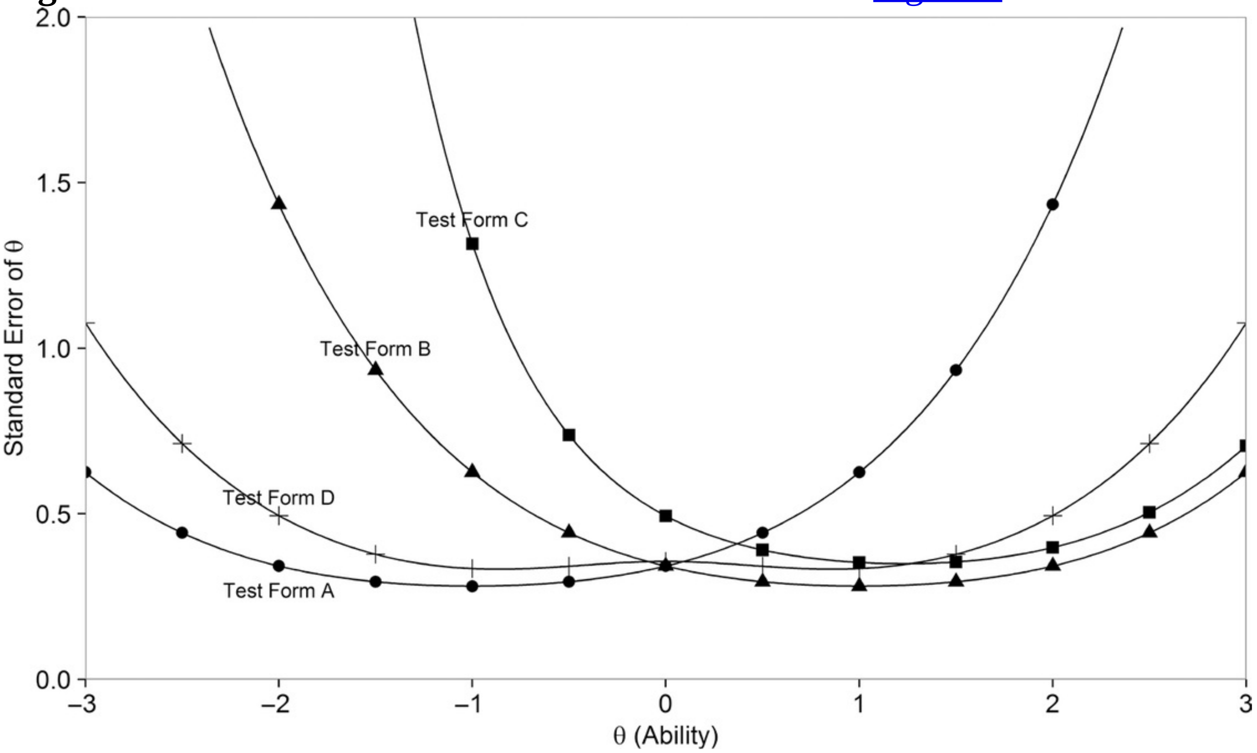
Figure 3 Test information for an easy test (Form A), a hard test (Form B), a hard test with higher lower asymptotes (Form C), and a test with item difficulty limited to two values (Form D).



First, notice that Tests A and B (which differed only in the difficulty parameters) had similarly shaped test information, but the peak information occurred at a lower θ value for the easier test, Test A. Test C, with the same item difficulties and discrimination parameters as Test B, had lower test information due to the nonzero *c* parameters. Test D had more uniform information. Even though it had no middle difficulty items, both the easy and hard items provided some information in the middle so the dip in the middle is only a slight dip.

The test information function is useful because it has an inverse relationship with the standard error of θ. The estimate of the standard error of θ is . Figure 4 shows the standard error functions corresponding to the test information functions in Figure 3. This standard error function represents the theoretical standard error of θ where is estimated through maximum likelihood.

Figure 4 Standard error of θ for the test forms shown in Figure 3.



Christine E. DeMars

See also Item Response Theory; Maximum Likelihood Estimation; Standard Error of Measurement

Further Readings

Bradlow, E. T. (1996). Negative information and the three-parameter logistic model. Journal of Educational and Behavioral Statistics, 21, 179–185. doi:10.2307/1165216

Eliason, S. R. (1993). Section 3: An introduction of basic estimation techniques. In Maximum likelihood estimation: Logic and practice (pp. 39–45). Newbury Park, CA: Sage.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Chapter 6: Item and test information and efficiency functions. In Fundamentals of item response theory (pp. 91–98). Newbury Park, CA: Sage.

Yen, W. M., Burket, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. Psychometrika, 56, 39–54. doi:10.1007/BF02294584

Anne Corinne Huggins-Manley Anne Corinne Huggins-Manley Huggins-Manley, Anne Corinne

Item Response Theory

Item response theory

903

908

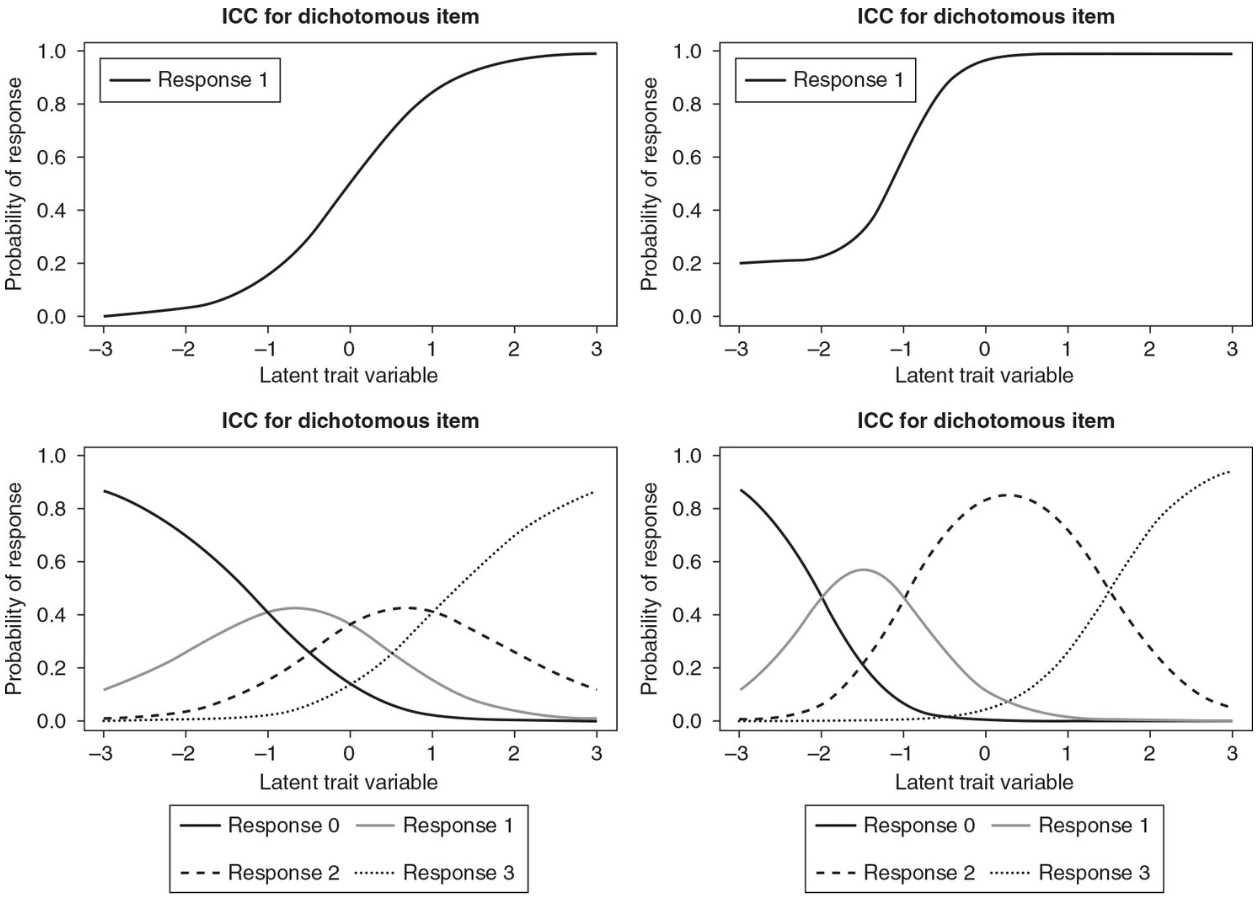
Item Response Theory

Item response theory (IRT) is a measurement framework for the development of tests and the scoring of item responses on tests. Key aspects of the IRT framework include the focus on items as the units of observed measurement, the fitting of parametric statistical models to categorical item response data, the estimation of a latent trait variable, and the conditional nature of reliability and the standard error of measurement. IRT is relevant to the field of educational measurement in that it is widely used by measurement practitioners for developing and scoring standardized tests, such as the SAT, ACT, and statewide K–12 achievement tests in mathematics, reading, and other academic domains. IRT is also widely used for the scoring of scale data, such as attitude scales consisting of a series of Likert-type items. In educational research, IRT is most often used for developing measurement tools, evaluating reliability of test data, and estimating latent abilities that can then be used as variables in research studies. The remainder of this entry provides details for the conceptual understanding of some selected statistical models in the IRT family, reliability and error under the IRT framework, and the relationship of IRT to other measurement theories.

Statistical Models in the Item Response Theory Family

IRT encompasses a family of parametric statistical models that are fit to item response data to estimate scores on a latent trait variable. Many models within the family fall into one of the two types: those that are built for dichotomous item responses (i.e., the data for each item can take on only two values) and those that are built for polytomous item responses (i.e., the data for each item can take on three or more values). Fitting an IRT model to a data set of item responses produces, among other things, a series of item characteristic curves (ICCs) that relate the underlying latent trait variable to the probability of scoring in each of the item response categories. Figure 1 shows examples of ICCs for several dichotomously and polytomously scored items.

Figure 1 Item Characteristic Curves



Conceptually, the ICC is the statistically estimated answer to the measurement question: “What is the probability of an examinee providing a particular response to this test item, given that the examinee has a particular trait level?” The ICC reflects the nature of all IRT model specifications in that the latent trait variable is treated as an independent variable that has a functional relationship to the multiple dependent variables of observed item responses. For dichotomously scored items, the ICC is displayed as an *S*-shaped curve representing the conditional probability of scoring a 1 on the item as opposed to a 0. The curve representing the conditional probability of scoring a 0 is not shown because it is redundant information (i.e., the probability of a 0 response is one minus the probability of a 1 response). For polytomously scored items, the ICC is displayed as a series of conditional probability curves, one for each response category on the item. At any point on the latent trait variable, the sum of the probabilities of scoring in each of the possible response categories is 1.

Selected Models for Dichotomous Item Responses

Three commonly used IRT models for dichotomous item responses are the one-

Three commonly used IRT models for dichotomous item responses are the one-parameter logistic model (1PL), two-parameter logistic model (2PL), and three-parameter logistic model (3PL). These models are distinguished by the types of item parameters that are used to estimate the ICC, otherwise stated as the types of item parameters that are used to define the relationship between the latent trait variable and the probability of scoring a 1 on the item. The 3PL defines the ICCs according to three item parameters that can vary across items, the 2PL defines the ICCs according to two item parameters that can vary across items, and the 1PL defines the ICCs according to one item parameter that can vary across items.

The 3PL model assumes that the responses to each item on the test are driven by the examinee’s trait level and three item properties: the difficulty of the item (i.e., how hard it is to score a 1 as opposed to a 0), the discrimination of the item (i.e., how strong the relationship is between the latent trait variable and the responses on the item), and the lower asymptote of the item’s ICC (i.e., how likely it is that an examinee of infinitely low trait could score a 1 on the item). Item difficulty parameters are often referred to as *b* parameters, and they define the location of the ICC along the latent trait variable scale. Item discrimination parameters are often referred to as *a* parameters, and they define the slope of the ICC at the point of inflection of the *S*-shaped curve. Item lower asymptote parameters are often referred to as *c* parameters, and they define the lower probability boundary of the ICC. This last parameter is sometimes called a pseudo-guessing parameter, as it is used to statistically account for the fact that examinees can guess a correct answer on a multiple-choice ability test item regardless of trait level.

The 2PL model is equivalent to the 3PL model except it forces all items to have a lower asymptote of zero. In doing this, the responses to each item on the test are conceptualized and analyzed under the assumption that they are driven by the examinee’s trait level and only two item properties: item difficulty and item discrimination. The 2PL model is therefore a special case of the 3PL model in which the *c* parameter is fixed to zero.

Similarly, the 1PL model is a special case of the 2PL and 3PL models. In the 1PL model, item responses are assumed to be a function of the examinee’s trait level and item difficulty only. This is accomplished by fixing the item discrimination parameters in the 2PL model to one across all items. Some versions of the 1PL model allow the discrimination parameter to be different from one but continue to force it to be equal across all items.

The relative differences in the number of constraints on item parameters across the three models leads to at least three general implications for practice. First, because the estimation of more item parameters requires more data, the minimum sample size requirement for calibrating data under the 1PL model is less than that of the 2PL model, which is less than that of the 3PL model. Second, parameter constraints result in less flexible models, so the 1PL model cannot fit better to a set of data than the 2PL model, which cannot fit better to a set of data than the 3PL model. Third, the allowance of *c* parameters greater than zero is associated with some difficulties in model estimation. However, the 3PL is still used in many modeling applications, where the belief is that examinee guessing most likely occurs and, therefore, needs to be modeled.

Selected Models for Polytomous Item Responses

Polytomous IRT models can be classified into groups in several ways, one way of which is based on the manner in which response categories within items are modeled. There are models that estimate cumulative response category parameters, models that estimate adjacent response category parameters, and models that estimate independent response category parameters. The latter two categories of models are nested and both are considered “divide-by-total” models, but they are treated separately here to account for their differential assumptions about the scale of measurement of the observed item responses (i.e., models for estimating adjacent response category parameters assume the observed item data are ordinal and models for estimating independent response category parameters assume the observed item data are nominal).

An example of an IRT model that estimates cumulative response category parameters is the graded response model (GRM). It is a polytomous extension of the 2PL model. The GRM estimates multiple item difficulty parameters and one item discrimination parameter for each polytomous item on a test. The item discrimination parameter defines the steepness of the curves in the ICC. For example, the item in the lower left quadrant in Figure 1 has a smaller discrimination parameter than the item in the lower right quadrant. The cumulative nature of the item difficulty parameters can be conceptually understood in relation to the ICCs for polytomous items in Figure 1. The first item difficulty parameter of the item is located at the point on the latent trait variable for which there is an equal probability of responding in category 0 or responding in any of the categories greater than 0. The second item difficulty parameter is located at the point on the latent trait variable for which the probability of responding in categories equal to or less than 1 is equal to the probability of scoring in categories greater than 1. Similarly, the third item difficulty parameter is located at the point on the latent trait variable for which there is an equal probability of responding in categories less than 3 or in Category 3.

An example of an IRT model that estimates adjacent response category parameters is the generalized partial credit model (GPCM). The model was originally developed as an extension of the partial credit model, which is part of the Rasch family of latent measurement models. Like the GRM, the GPCM estimates three item difficulty parameters and one item discrimination parameter for each of the polytomous items shown in Figure 1. The discrimination parameter in the GPCM operates in an analogous manner to the GRM. However, the item difficulty parameters are estimated through an adjacent categories approach. In relation to the ICCs for polytomous items in Figure 1, the first difficulty parameter under the GPCM is located at the point on the latent trait variable for which there is an equal probability of scoring in Category 0 or Category 1. The second difficulty parameter is located at the point on the latent trait variable for which there is an equal probability of scoring in Category 1 or Category 2. The third difficulty parameter is located at the point on the latent trait variable for which there is an equal probability of scoring in Category 2 or Category 3. Notice that these interpretations of difficulty parameters are each respective to adjacent response categories, ignoring other categories. This is the distinguishing factor between polytomous IRT models that estimate cumulative response category parameters and polytomous IRT models that estimate adjacent response category parameters.

The nominal response model (NRM) is an example of an IRT model that estimates independent response category parameters. Because the item response data are not assumed to have an underlying order, a separate slope and intercept parameter are estimated for each response category. An ICC from the NRM would have multiple curves just as in the polytomous ICCs shown in Figure 1, yet the curves would be free to take on a wider variety of forms than would be permitted under polytomous IRT models that assume ordered categories. The slope for each response category of an item is analogous to a discrimination parameter in that it defines the strength of the relationship between the latent trait variable and the nominal response category. The intercept parameter for each response category of an item is analogous to a difficulty parameter in that it defines the location of the nominal response category curve along the latent trait variable scale.

Additional Information on Item Response Theory Models

IRT models rely on strong, explicit assumptions. Specifically, the observed data across all items are assumed to be unidimensional. Once the effect of the unidimensional latent trait is removed from the item data, it is assumed that there are no other relationships among the data (i.e., locally independent data). On a conceptual level, the models assume that all items measure the same underlying trait and no other traits in common with other items. On a statistical level, the conditional likelihood functions are utilized to estimate items, and examinee parameters are built from products of probabilities that are only appropriate to calculate if those probabilities are locally independent.

A variety of parameter estimation techniques are available for IRT models. When all examinee and item parameters are unknown, joint maximum likelihood and marginal maximum likelihood are common approaches to the simultaneous estimation of all parameters. When prior distributions of examinee parameters are known, Bayesian methods are commonly incorporated into the joint or marginal maximum likelihood estimation of posterior distributions of the examinee latent trait variable. Two Bayesian options include maximum a posteriori scoring for locating the mode of the posterior distribution as the estimate of the latent trait score and expected a posteriori scoring for locating the mean of the posterior distribution as the estimate of the latent trait score.

Conditional Reliability and Measurement Error in Item Response Theory

A feature of the IRT framework that has both theoretical and practical implications is the definition of reliability and measurement error as conditional on the latent trait variable. Conditional reliability and error incorporate the intuitive notion that tests may produce more reliable data for some examinees as compared to others, depending on their trait levels. For example, an algebra test for high school students may be able to produce reliable data for students who are already taking algebra and have some understanding of algebraic concepts. But that same test may not be able to produce reliable data for high school

But that same test may not be able to produce reliable data for high school students who have had little exposure to algebra and are scoring very low on the test. In IRT, the amount of error in test scores is free to vary across different points along the latent trait variable.

Reliability of latent trait scores is formalized in IRT by a test information function. For any given item and IRT model that was fit to the item data, the amount of conditional information that the model can provide about the latent trait from each item can be calculated with prespecified formulas. For example, in the 2PL model, conditional information is defined as:

$$I(\theta) = 2.89a^2 (P)(1 - P),$$

where *a* is the item’s discrimination parameter, θ is the latent trait variable, and *P* is the conditional probability of obtaining a 1 on the item. The conditional item information is then summed across all the items on the test to obtain the conditional test information function. The higher the information, the higher the reliability of latent trait scores at that particular level of the latent trait variable. The conditional standard error of measurement can then be calculated as:

$$\mathrm{SEM}(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

The higher the information, the lower the standard error of measurement of latent trait scores at that particular level of the latent trait variable.

Relationships Between Item Response Theory and Other Measurement Frameworks

IRT has many connections to, and distinctions from, other measurement theories and statistical approaches to measurement. Following is a comparison of IRT to Rasch measurement, classical test theory (CTT), and factor analysis.

IRT and Rasch measurement models both fall into the broader family of multivariate (multinomial) logistic regression models, and several Rasch models are nested within IRT models (i.e., are special cases of IRT models). However, the two measurement frameworks are separated on theoretical grounds. Rasch models conform to a theory of objective measurement, which translates to a restriction of item parameters in the models that forces equal discrimination power across items. IRT is more flexible in the type of item parameters that are free to vary in its family of models.

IRT is often contrasted to CTT. The contrasts include IRT's emphasis on items as the units of observed measurement and CTT's emphasis on total test scores as the unit of observed measurement, IRT's use of falsifiable statistical models and CTT's use of a true score model that cannot be directly tested or rejected with a set of test data, IRT's conditional definition of reliability and the standard error of measurement and CTT's unconditional definitions, and the sample and test independence of IRT parameters as compared to the sample and test dependent outputs of CTT-based test scoring. However, some measurement experts have demonstrated that IRT is better seen as an extension of CTT rather than a competing theory.

IRT models have much overlap and, sometimes, equivalency to confirmatory factor analysis (CFA) models. For example, fitting a CFA model to dichotomously scored item data results in factor loading and item intercept estimates that are perfectly correlated with 2PL item discrimination and item difficulty parameter estimates, respectively. That same CFA model fit to polytomously scored item data can produce equivalency in model fit and parameter estimation of the GRM. Overall, many IRT models can be estimated within a CFA framework, but there are some IRT models that have parameters that are distinct from the literature and practice of CFA.

Anne Corinne Huggins-Manley

See also *a* Parameter; *b* Parameter; *c* Parameter; Conditional Standard Error of Measurement; Item Information Function; Local Independence; Multidimensional Item Response Theory; Psychometrics

Further Readings

Baker, F. B., & Kim, S. (2004). Item response theory: Parameter estimation techniques (2nd ed.). Boca Raton, FL: Taylor … Francis Group.

de Ayala, R. J. (2009). The theory and practice of item response theory. New York, NY: The Guilford Press.

DeMars, C. (2010). Item response theory: Understanding statistics measurement. New York, NY: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Norwell, MA: Kluwer Academic.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Mahwah, NJ: Erlbaum.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. Oxford, UK: Addison-Wesley.

Ostini, R., & Nering, M. L. (2006). Polytomous item response models. Thousand Oaks, CA: Sage.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2010). Handbook of modern item response theory. New York, NY: Springer.

J

Thomas G. Ryan Thomas G. Ryan Ryan, Thomas G.

Johari Window

Johari window

909

909

Johari Window

The Johari window is an investigative tool developed by Joseph Luft and Harrington Ingham to illuminate and detail, in a somewhat measured way, interpersonal awareness. (The “Johari” in the name is a combination of the researchers’ first names.) The window is made up of four quadrants conceptualizing how one sees oneself and is seen by others. The Johari window can be used to help a person understand his or her behavior and how it affects other people in ways that the person previously did not recognize. It is made up of four areas: the open area, the blind area, the hidden area, and the unknown area.

Most often, the users of the Johari window have a goal to expand their open area. The open area represents what the users know, and what others know, about themselves. This could include information such as height, weight, ethnicity, skin color, mood, and manner of dress. In addition, information the user may have shared, such as preferred food, movie, pastime, and/or religion, may be known to others. Information in the open area includes information the user wants people to know in order to establish common bonds and friendships. As a person ages, more information may become visible (known) in this area, as more is disclosed, shown, and observable.

The blind area represents what other people know about a person that is unknown to the person himself or herself. This could include information about the person’s past that was never revealed to the person, or feelings that the person has trouble facing, but are evident to others. By getting feedback from others, a person can reduce this area in relation to the other areas and increase the person’s self-awareness.

the person’s self-awareness.

The hidden area represents what a person knows, but others do not know, about himself or herself. This could include a fear of success or feeling uncomfortable when speaking in public. This area can be reduced in size when a person establishes trusted relationships with others so the person can safely disclose upsetting and embarrassing moments from the past, moving them to the open area.

The unknown area involves what neither the person nor others know about that person. This could include the person’s abilities to save someone in an emergency, which until the emergency arises is unknown to that person or others.

Thomas G. Ryan

See also Qualitative Research Methods

Further Readings

Luft, J. (1961). The Johari window: A graphic model of awareness in interpersonal relations. NTL Human Relations Training News, 5, 6–7.

Luft, J. (1984). Group processes: An introduction to group dynamics. Palo Alto, CA: Mayfield.

Luft, J., & Ingham, H. (1955). The Johari window: A graphic model of interpersonal awareness. Proceedings of the Western Training Laboratory in Group Development. Los Angeles: University of California.

Copelan Gammon Copelan Gammon Gammon, Copelan

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

John Henry Effect

John henry effect

John Henry Effect

The John Henry effect refers to the bias introduced to an experiment when members of the control group are aware that they are being compared to the experimental group and behave differently than they typically would to compensate for their perceived disadvantage. This alteration renders the control group ineffective as a measure of baseline performance and skews the results of the experiment. The John Henry effect is also associated with resistance to change and innovation, which may be perceived as disruptive or threatening to the present status of members of the control group. This entry describes the John Henry effect's origins and potential implications for research into innovations in education.

History

The John Henry effect is named after legendary American folk hero John Henry, a steel driver in the 1870s, who, when faced with replacement by the steam drill, worked so hard to outperform the machine that he died in the process. In this scenario, John Henry represents the status quo (or control condition) that is threatened to be overturned by the innovation of the steam drill (or experimental condition), prompting him to work harder than he otherwise would have and producing atypical results.

Research Implications

The term *John Henry effect* was coined by Robert Heinich in 1970 in response to the failure of innovative education techniques to produce significantly better results than traditional classroom teaching. In examining the results, Heinich suggested that the classroom teachers performed well above average when they were aware that their work was being compared to alternative education methods. This prompted further inquiry by Gary Saretsky who noted the John Henry effect as a confounding influence on experimental evaluation of innovation.

Saretsky observed this effect in a 1972 analysis of the Office of Economic Opportunity’s experiment in performance contracting. The classroom teachers instructing the control group actively worked to outperform the “outsider” performance contractors who worked with the experimental group, leading to atypical performance from the control group students. This change invalidated results that may have been used to promote reform in educational practices and underlined the need to control for similar confounding factors in future experiments.

The John Henry effect has frequently been compared with the Hawthorne effect, also known as the observer effect. It refers to the phenomenon of actors modifying their behavior in response to the knowledge that they are being observed. Although related, the John Henry effect differs in that it focuses on consequences or perceived threat, whereas the Hawthorne effect focuses on the awareness and interaction between actor and observer. Additionally, the John Henry effect is typically associated with the control group and the Hawthorne effect with the experimental group.

Copelan Gammon and Marc H. Bornstein

See also Hawthorne Effect

Further Readings

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology, 69(2), 334–345.

McCracken, R. A. (1968). An observation of Hawthorne effect in an experiment in the teaching of reading in first grade a hypothesis. Washington, DC: ERIC Clearinghouse.

Saretsky, G. (1972). The OEO P.C. experiment and the John Henry effect. The Phi Delta Kappan, 53(9), 579–581.

Saretsky, G. (1975, April 2). The John Henry effect: Potential confounder of experimental vs. control group approaches to the evaluation of educational innovations. Paper presented at the American Educational Research Association’s Annual Meeting, Washington, DC.

Zdep, S. M., & Irvine, S. H. (1970). A reverse Hawthorne effect in educational evaluation. Journal of School Psychology, 8, 89–95.

Miles Allen McNall Miles Allen McNall McNall, Miles Allen

Joint Committee on Standards for Educational Evaluation Joint committee on standards for educational evaluation

Joint Committee on Standards for Educational Evaluation

The Joint Committee on Standards for Educational Evaluation is a standards developing organization. Standards developing organizations are created for the purpose of developing, promulgating, amending, and reissuing technical standards that apply to a particular industry or field of professional practice. Established in 1975, the Joint Committee is a private nonprofit organization accredited by the American National Standards Institute. This entry discusses the makeup of the Joint Committee and its standards for evaluation.

The Joint Committee was created for the explicit purpose of developing standards of quality in educational evaluation. It is composed of a coalition of 14 professional associations with an interest in the quality of evaluation. Sponsoring organizations include the American Educational Research Association, the National Council on Measurement in Education, the American Psychological Association, the National Education Association, the American Evaluation Association, and the Canadian Evaluation Society.

The Joint Committee has published three sets of standards for evaluation: the *Personnel Evaluation Standards*, the *Program Evaluation Standards*, and the *Classroom Assessment Standards for PreK–12 Teachers*. The focus of this article is on the oldest of these standards, the *Program Evaluation Standards*.

Early on, the Joint Committee identified four general attributes of quality in evaluation: accuracy, utility, propriety, and feasibility. In 1976, a panel of 36 individuals with expertise in evaluation was commissioned to elaborate these standards. In December 1976, the Joint Committee and project staff met to critique and revise the initial draft standards. In July 1977, the Joint Committee sent a revised draft of the standards for educational evaluation to a national panel of 50 evaluation experts for review. Based on the panel's review, a final set of standards was completed in 1979 and published in 1981. The *Program Evaluation Standards, Third Edition* (2011) consist of 30 program evaluation standards organized into five categories: utility, feasibility, propriety, accuracy, and accountability.

The eight utility standards are designed to ensure that evaluation stakeholders find genuine value in the processes and products of evaluations. The utility standards include standards on attention to stakeholders, or taking into consideration the interests of the full range of individuals with a stake in the evaluation; explicit values, or attending to the individual and cultural values underlying judgments of merit, worth, or value; relevant information, or ensuring that evaluative information serves the identified and emergent needs of stakeholders; and timely and appropriate communication and reporting, indicating that evaluations should attend to the information needs of different audiences, when those audiences need information and in what format.

The four feasibility standards are designed to improve evaluation effectiveness and efficiency and include standards dealing with project management, or the use of effective project management strategies; contextual viability, or the need to recognize and balance the cultural and political interests of individuals and groups with a stake in the evaluation; and resource use, or the effective and efficient use of evaluation resources.

The seven propriety standards are designed to support what is proper, legal, and just in evaluations and include standards on responsive and inclusive orientation, or the need for evaluations to be responsive to the evaluation context and the needs of various stakeholders; human rights and respect, or the need to conduct evaluations in a manner that protects the rights and maintains the dignity of evaluation participants; transparency and disclosure, or the need to provide complete descriptions of evaluation findings and limitations to all stakeholders whenever legally and ethically possible; and conflicts of interest, or the need to openly and honestly address real or perceived conflicts of interest.

The eight accuracy standards are designed to promote the dependability and trustworthiness of evaluation findings and include standards covering valid information, or the need for evaluative information to support valid conclusions;

reliable information, or the requirement for evaluations to produce reliable and dependable information; and sound designs and analyses, or the use of evaluation designs and analytic procedures that are both technically sound and appropriate to the purposes and context of the evaluation.

Finally, the three accountability standards are designed to support full documentation of evaluations and to encourage the practice of meta-evaluation (i.e., the evaluation of evaluations). These standards cover evaluation documentation, or the full documentation of the purpose, design, procedures, and products of evaluations; internal meta-evaluation, or evaluators' use of the *Program Evaluation Standards* to periodically assess the quality of their own work; and external meta-evaluation, which calls for the routine performance of external meta-evaluations to judge the quality of evaluations by the *Program Evaluation Standards*.

Since their publication in 1981, the *Program Evaluation Standards* have been adopted and adapted worldwide, including by the Canadian Evaluation Society and the African Evaluation Society. However, the *Program Evaluation Standards* are not the only recognized standards for quality in evaluation. The Organisation for Economic Co-operation and Development has published the *Quality Standards for Development Evaluation*. In addition, professional evaluation associations such as the American Evaluation Association and the Canadian Evaluation Society have promulgated practice guidelines for their members, some of which overlap with the *Program Evaluation Standards*. For example, one of the key principles of the American Evaluation Association's *Guiding Principles for Evaluators* is systematic inquiry, which calls for evaluators to conduct systematic, data-based inquiries that meet the highest technical standards. This principle overlaps substantially with the *Program Evaluation Standards*' accuracy standards.

The routine and systematic application of the *Program Evaluation Standards* supplemented by the evaluation guidelines promulgated by national evaluation associations when designing evaluations, conducting evaluations, and evaluating the quality of evaluations after the fact holds the promise of substantially increasing the quality and impact of evaluation practice and enhancing the professionalism of the evaluation field.

Miles Allen McNall

See Also American Educational Research Association; American Evaluation Association; American Psychological Association; *Guiding Principles for Evaluators*; National Council on Measurement in Education; *Standards for Educational and Psychological Testing*

Further Readings

American Educational Research Association. (1977, September 12). Joint Committee on Standards for Educational Evaluation Update—September 1977. AERA Division H Newsletter. Retrieved February 4, 2016, from http://www.jcsee.org/about-jcsee/more-background-info

American Evaluation Association. (2004, July). American Evaluation Association guiding principles for evaluators. Retrieved February 4, 2016, from http://www.eval.org/p/cm/ld/fid=51

Organisation for Economic Co-operation and Development. (2010). Quality standards for development evaluation. Author.

Sanders, J. (1999, April). The development of standards for evaluations of students. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada. Retrieved February 4, 2016 from http://www.jcsee.org/wp-content/uploads/2009/09/JCGeneralBackground.pdf

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). The program evaluation standards: A guide for evaluators and evaluation users. (3rd ed.). Thousand Oaks, CA: Sage.

Jana Craig-Hare Jana Craig-Hare Craig-Hare, Jana

Journal Articles Journal articles

Journal Articles

Journal articles introduce scholarly ideas and research. Articles present the authors' original research as reports of empirical studies, literature reviews, case studies, or theoretical or methodological statements. Beginning as unpublished manuscripts of research not previously published, these articles typically undergo a peer-review process by one or more academic referees before being accepted or rejected for publication within a journal. The primary goal of a published journal article in an educational research journal is for the author(s) to communicate ideas, concepts, and research findings to the educational community.

There are many different types of journal articles. Empirical studies represent original research and consist of specific sections that reflect the research process. Literature reviews embody critical evaluations of previously published material through a research synthesis or meta-analysis. Case studies report materials obtained while working with an individual or groups of individuals within a community or an organization. Theoretical articles use existing research literature to advance theory, refine theoretical constructs, present new theory, or challenge existing theory. Methodological articles focus on methodological approaches, allowing the reader to access and compare proposed methods with current use and application.

Journal articles begin with an abstract summarizing the contents of the article. Although the elements of a manuscript may be determined by the type of journal article being written, standard reporting sections include (1) *introduction*, to identify the problem being addressed and purpose of the study; (2) *method*, describing the procedures used to conduct the study as well as participant characteristics and demographics; (3) *results*, reporting findings and analysis of data; and (4) *discussion*, summarizing the study findings and discussing implications for future research. References are included at the end of the article and provide documentation about the literature identified in the article. Appendixes and supplemental materials may also be included at the end of the article. As a general rule, educational research publications in the behavioral and social sciences utilize American Psychological Association formatting guidelines and publishing standards.

There are varied requirements for submitting articles to journals. It is advised to carefully review the requirements of the targeted journal. Once the journal editor receives the manuscript, the blind peer-review process begins. Experts in the field constitute the review board. The review board is charged with evaluating each manuscript to determine its publication potential. Reviewers may identify strengths and criticisms of the article as part of their critique. A general framework and rubric guide their review.

The journal editor presents the final recommendation regarding the article, typically choosing from one of four options: reject, revise and resubmit, accept with major revisions, and accept with minor revisions. If a journal article is rejected, the author may choose to resubmit it for consideration in a different journal. It is important to remember, however, that a manuscript can only be submitted for article consideration in one journal at a time.

Jana Craig-Hare

See also Abstracts; APA Format; Authorship; Demographics; Literature Review; Methods Section; Results Section

Further Readings

Becker, H. S. (2007). Writing for social scientists: How to start and finish your thesis, book, or article (2nd ed.). Chicago, IL: University of Chicago Press.

Belcher, W. L. (2009). Writing your journal article in 12 weeks: A guide to academic publishing success. Thousand Oaks, CA: Sage.

Huff, A. S. (2009). Designing research for publication. Thousand Oaks, CA: Sage.

Andrew Maul Andrew Maul Maul, Andrew

Judgment Sampling Judgment sampling

Judgment Sampling

Judgment sampling (a type of purposive sampling) occurs when units are selected for inclusion in a study based on the professional judgment of the researcher. This is in contrast to probability sampling techniques in which units are drawn with some probability (e.g., randomly) from the population of interest. This entry describes the common forms of judgment sampling and discusses their advantages and limitations.

Judgment sampling may be used for a variety of reasons. In general, the goal of judgment sampling is to deliberately select units (e.g., individual people, events, objects) that are best suited to enable researchers to address their research questions. This is often done when the population of interest is very small, or desired characteristics of units are very rare, making probabilistic sampling infeasible. Judgment sampling is often associated with qualitative and mixed-method study designs.

In some cases, it may be possible to sample the entire population of interest; this is referred to as “total population sampling.” For example, college professors may be interested in the perspectives of their own graduate students or researchers may be interested in the perspectives of current four-star Army generals (of which there are only 12).

Another common case, also referred to as “maximum variation sampling,” involves deliberately sampling subjects so as to maximize the range of one or more attributes of interest. This may be especially valuable when resources permit collecting data from only a small number of subjects, and random sampling would likely fail to capture the desired range of variation. For example, a researcher developing a survey of political attitudes may wish to conduct in-depth interviews with a small number of people to ensure that the survey is depth interviews with a small number of people to ensure that the survey is capable of capturing diverse viewpoints and may deliberately seek out people with extreme positions to do so.

A close variant of the maximum variation sampling approach, also called “extreme case sampling,” involves deliberately seeking out unusual or deviant cases (which might be missed in a simple random sample); for example, a researcher may wish to study the practices of the most highly competent or creative members of a profession (by whatever criteria). Conversely, a researcher may wish to focus only on typical cases or on units that share some set of characteristics, in which case extreme cases may be deliberately avoided.

Advantages and Drawbacks of Judgment Sampling

As illustrated by the examples given previously, judgment sampling may be a more efficient means of acquiring information from desired types of units than probabilistic sampling, especially when the population is small or the desired characteristics of units are rare or resources for data collection are limited. However, due to the necessarily subjective nature of human judgment, judgment samples can be prone to researcher bias. Further, depending on the nature of the sampling procedure, it may be difficult to frame a judgment sample as forming a representative sample from a population, limiting the extent to which one can confidently generalize (i.e., make inferences) from the sample.

Andrew Maul

See also Convenience Sampling; Random Assignment; Simple Random Sampling

Further Readings

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. Psychometrika, 66(2), 209–227.

Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In Utility, probability, and human decision making (pp. 141–162). Dordrecht, The Netherlands: Springer.

K

Robert L. Johnson Robert L. Johnson Johnson, Robert L.

Kelvin Terrell Pompey Kelvin Terrell Pompey Pompey, Kelvin Terrell

Kappa Coefficient of Agreement Kappa coefficient of agreement

915

919

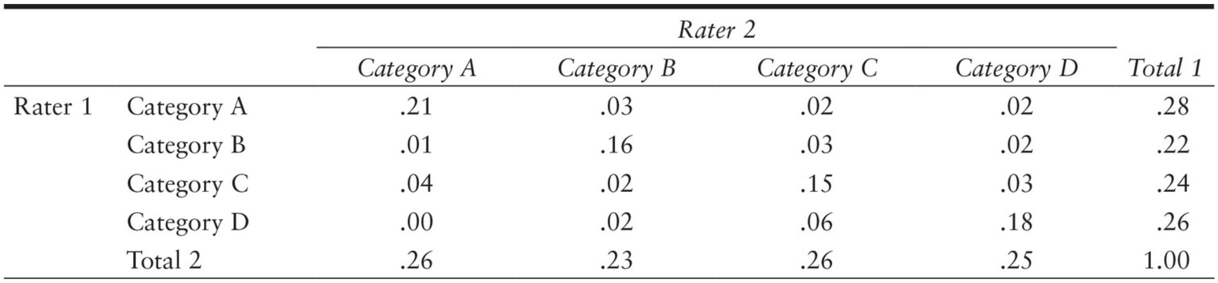
Kappa Coefficient of Agreement

Kappa, one of several coefficients used to estimate inter-rater and similar types of reliability, was developed in 1960 by Jacob Cohen. In its original conception, kappa, denoted κ, was an index used to measure the level of consistency between two raters who use rubrics or other instruments to place subjects (i.e., people) into one of κ nominal categories. For example, two evaluators might use a rubric to classify the instructional strategies used in a classroom as one of two nominal categories, such as effective or ineffective, or two psychologists might use an instrument to identify a person’s depression as major depression, bipolar disorder, persistent depression, or psychotic depression. In both of these cases, subjectivity might lead to disagreements about the category assigned and raise concerns about the use and interpretation of the instrument. For this reason, coefficients of agreement, such as κ, have been developed.

Since its inception, κ has been one of the most widely utilized coefficients of agreement and has been used in such fields as education, medicine, and the social sciences. In these and other fields, κ can be used not only to estimate reliability but also to quantify the variance that can be attributed to the rating process. This entry provides a description of the development of κ and some extensions. It will also include a discussion of considerations that must be attended to when using this coefficient of agreement.

Development of κ

A simple, logical coefficient of agreement between two raters is the observed proportion of subjects the raters placed into the same category. This is called the observed proportion of agreement, denoted p_a. Table 1 is a contingency table containing hypothetical data that depicts the proportion of subjects and the categories in which each rater placed them. For example, 4% of subjects were placed in Category C by Rater 1 and in Category A by Rater 2. Based on Table 1, the proportion of agreement, which is the sum of the proportions along the main diagonal given by , is .70. Therefore, the raters agreed and placed 70% of subjects into the same category.



		Rater 2				
		Category A	*Category B*	*Category C*	*Category D*	*Total 1*
Rater 1	Category A	.21	.03	.02	.02	.28
	Category B	.01	.16	.03	.02	.22
	Category C	.04	.02	.15	.03	.24
	Category D	.00	.02	.06	.18	.26
	Total 2	.26	.23	.26	.25	1.00

Although the proportion of agreement is a simple and logical coefficient of agreement, it has been criticized by Cohen and others as being insufficient. The inadequacy stems from the idea that raters will have a certain level of agreement by chance alone, and p_a does not take that into consideration. Others before Cohen developed their own corrections for chance agreement; however, Cohen's correction is one of the few that involves the use of marginal proportions in its calculation with the assumption that each rater's marginal proportions are specific to that rater and not common across raters. To be exact, the expected proportion of chance agreement, denoted p_c, is given by , where κ is the number of categories, $p_{i.}$ is the proportion of subjects Rater 1 put into Category *i*, and $p_{.i}$ is the proportion of subjects Rater 2 put into Category *i*. Thus, Cohen's expected proportion of agreement by chance is the sum of the product of marginal proportions for each category. Using the data in Table 1, p_c = 0.2508. This is interpreted to mean that by chance alone, it is expected that the raters will agree on approximately 25% of the ratings. Cohen's κ uses both p_a and p_c. More specifically, the formula for κ is given by . Proper use of κ requires the following assumptions set forth by Cohen: (a) independence of subjects; (b) nominal, independent, mutually exclusive, and exhaustive categories; and (c) independence of raters. Assuming that these assumptions have been satisfied for the data in Table 1, κ = 0.60. This value represents the proportion of agreement between the raters after the removal of chance agreement.

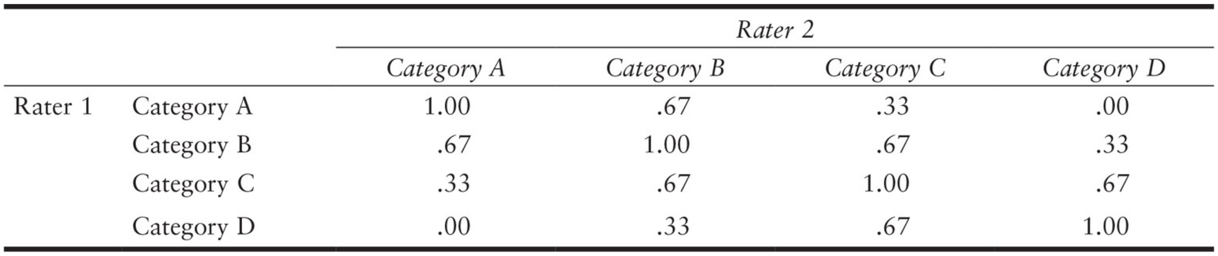
The value of κ ranges from −1 to 1. A κ value of zero indicates that the proportion of agreement observed is equal to the proportion of agreement expected by chance; a κ value that is greater than zero indicates that the raters agreed more than that which is expected by chance; a κ value that is less than zero indicates that the raters agreed less than that which is expected by chance; and a κ value equal to one indicates that the raters agreed on all ratings. Although the maximum possible value of κ is 1, it will only be obtained in rare cases. Cohen, in his original development of κ, introduced a formula for the maximum value of κ based on marginal proportions, which is given by , where and min ($p_{i.}$, $p_{.i}$) is the minimum value between $p_{i.}$ and $p_{.i}$. In the example depicted in Table 1, $\kappa_M = 0.96$. Therefore, in the current example, the maximum percentage of agreement that can be obtained after the removal of chance agreement permitted by the marginal proportions is 96%. Thus, the obtained κ is 36% points lower than the maximum possible value of κ, that is, 0.96–0.60.

Although Cohen did not set a standard for acceptable values of κ for reliability studies, others have. One of the most commonly used scales was developed in 1977 by J. Richard Landis and Gary Koch. Based on their cutoffs, any value of κ greater than 0.40 can be characterized as moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (0.81–1.00). All other values can be characterized as poor (<0.00), slight (0.0–0.20), or fair (0.21–0.40). Although this is one of the most commonly used scales, its cutoff values were subjectively selected. Other scales have been developed by A. G. Altman and Joseph Fleiss, Bruce Levin, and Myunghee Cho Paik.

Extensions

In 1968, Cohen developed the weighted κ coefficient, an extension of κ used for situations in which raters place subjects into ordinal or ordinal-like categories. In these cases, a disagreement between raters is no longer considered an absolute disagreement but rather partial agreement. Therefore, methods were developed to quantify the magnitude of disagreements by assigning weights to each cell in a cross tabulation of the rating data. The choice of weights can be arbitrary; however, for clear interpretation of the weighted κ coefficient, much consideration should be given in the selection of weights. This might require an expert panel with substantial knowledge on the differences in ratings to make those judgments. One method to assign weights for calculating weighted κ, denoted κ_w, is to weigh disagreements linearly by utilizing the ratio of the distance between two categories and the maximum distances between categories.

In this case, weights are given by , where *i*,*j* = 1, 2, …, *k* and *k* is the number of categories. Table 2 shows the linear weights for the data in Table 1. For instance, the weight for subjects assigned to Category A by Rater 1 and Category B by Rater 2 is calculated as .



		Rater 2			
		Category A	*Category B*	*Category C*	*Category D*
Rater 1	Category A	1.00	.67	.33	.00
	Category B	.67	1.00	.67	.33
	Category C	.33	.67	1.00	.67
	Category D	.00	.33	.67	1.00

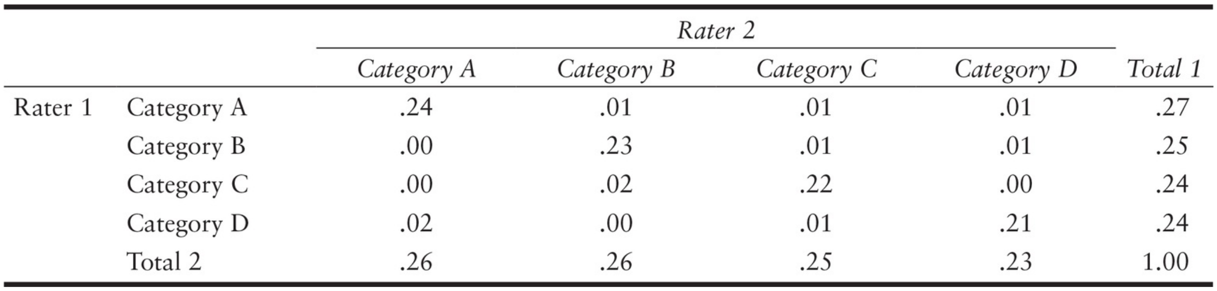
Note. Authors created specifically for this entry.

With the assignment of weights established, weighted κ is given by , where is the proportion of weighted agreement and is given by: , and , is the proportion of weighted chance agreement and is given by: . Using the proportions and weights listed in Tables 1 and 2, the proportion of weighted agreement using linear weights is , and the proportion of weighted chance agreement is . Therefore, the weighted κ coefficient is $\kappa_w = 0.66$. Note that if κ were used instead of κ_w for ordinal data, the result would have been a lower value. It is larger due to the partial agreement allowed with κ_w. Although this calculation involved linear weights, other weighting schemes include quadratic weights and weights defined by an expert. Regardless of the weights used, precautions should be taken when interpreting this statistic because its value is dependent on the magnitude of the weights.

There have been additional extensions of κ that include assigning subjects into ordinal categories, assigning interval values such as scores to subjects, and determining inter-rater reliability between more than two raters. More information on these topics can be found in the sources listed in further readings.

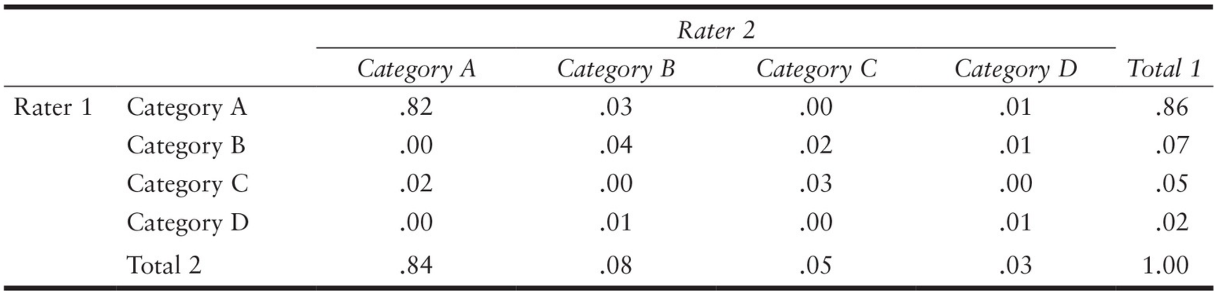
Considerations

Although κ is a widely used coefficient for inter-rater reliability, considerations should be made when deciding the appropriateness of its use. These considerations involve the structure of marginal proportions in the contingency table of the data. For instance, one should be mindful of the effect of the proportion of chance agreement when the marginal proportions are balanced versus when they are unbalanced. Table 3 depicts data in which the marginal proportions are balanced, that is, approximately equal proportions across all categories for both raters. Table 4 depicts data in which the marginal proportions are unbalanced. In both tables, the proportion of agreement is equal, $p_a = 0.90$. Because the proportion of agreement is equal for both sets of data, one would expect the value of κ to be the same or at least similar. For the data in Table 3, κ = 0.87, and in Table 4, κ = 0.63. The difference in these values is not due to the level of agreement between the raters but due to the calculation of expected agreement, which is dependent on the balance of the margins. Thus, interpreting this coefficient may be erroneous in the case of unbalanced margins.



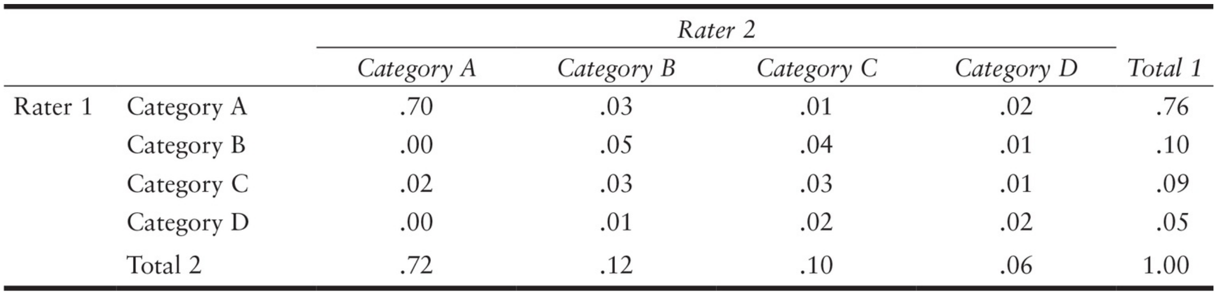
		Rater 2				
		Category A	*Category B*	*Category C*	*Category D*	*Total 1*
Rater 1	Category A	.24	.01	.01	.01	.27
	Category B	.00	.23	.01	.01	.25
	Category C	.00	.02	.22	.00	.24
	Category D	.02	.00	.01	.21	.24
	Total 2	.26	.26	.25	.23	1.00

Note. Authors created specifically for this entry.

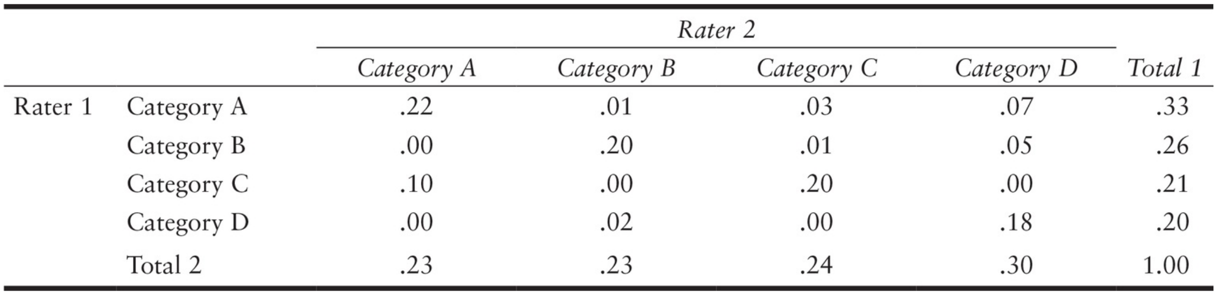


		Rater 2				
		Category A	*Category B*	*Category C*	*Category D*	*Total 1*
Rater 1	Category A	.82	.03	.00	.01	.86
	Category B	.00	.04	.02	.01	.07
	Category C	.02	.00	.03	.00	.05
	Category D	.00	.01	.00	.01	.02
	Total 2	.84	.08	.05	.03	1.00

Another consideration is related to whether the marginal proportions are symmetrical. Marginal proportions are symmetrical when the proportion that Rater 1 placed in each category is similar to the proportion that Rater 2 placed in each category. Table 5 contains data with symmetrical marginal proportions. The raters placed a similar proportion of subjects into each category. Table 6 contains data with asymmetrical marginal proportions. In both tables, the proportion of agreement is equal, $p_a = 0.80$. Because the proportion of agreement is equal for both sets of data, one would expect the value of κ to be the same or at least similar. However, for the data in Table 5, κ = 0.53, and in Table 6, κ = 0.73. This difference is partially due to the high marginal proportions for one category for both raters in the symmetrical margins' case and also due to the definition of chance agreement. Thus, the interpretation and use of κ may be erroneous for symmetrical data.



		Rater 2				
		Category A	*Category B*	*Category C*	*Category D*	*Total 1*
Rater 1	Category A	.70	.03	.01	.02	.76
	Category B	.00	.05	.04	.01	.10
	Category C	.02	.03	.03	.01	.09
	Category D	.00	.01	.02	.02	.05
	Total 2	.72	.12	.10	.06	1.00



		Rater 2				
		Category A	*Category B*	*Category C*	*Category D*	*Total 1*
Rater 1	Category A	.22	.01	.03	.07	.33
	Category B	.00	.20	.01	.05	.26
	Category C	.10	.00	.20	.00	.21
	Category D	.00	.02	.00	.18	.20
	Total 2	.23	.23	.24	.30	1.00

Although only two considerations were mentioned, others are noted in the literature and can be found in the further readings. In any research experiment or study, decisions have to be made regarding the appropriateness of the methods and tools used. There is no exception when it comes to using κ, weighted κ, or any of the extensions. Although κ is widely used, one may want to consider other indices of agreement if κ is inappropriate.

Robert L. Johnson and Kelvin Terrell Pompey

See also Inter-Rater Reliability; Reliability

Further Readings

Berry, K. J., & Mielke, P. W. Jr. (1988). A generalization of Cohen’s kappa agreement measure to interval measurement and multiple raters. Educational and Psychological Measurement, 48(4), 921–933. doi:10.1177/0013164488484007

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. doi:10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220. doi:10.1037/h0026256

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. Psychological Bulletin, 88(2), 322–328. doi:10.1037/0033-2909.88.2.322

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543–549. doi:10.1016/0895-4356(90)90158-L

Gwet, K. L. (2010). Handbook of inter-rater reliability (2nd ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Ryan J. McGill Ryan J. McGill McGill, Ryan J.

Kaufman-ABC Intelligence Test Kaufman-ABC intelligence test

919

922

Kaufman-ABC Intelligence Test

The Kaufman Assessment Battery for Children, Second Edition (KABC-II), is a measure of processing and cognitive ability for children and adolescents between the ages of 3 and 18 years. The KABC-II is a versatile instrument that can be used to assess for intellectual disability, learning disorders, developmental disabilities, more focal neurocognitive impairments, and intellectual giftedness, although it should be noted that the diagnosis of intellectual disability requires additional assessment of adaptive behavior. It can be administered as a complete measure of mental processing and general cognitive ability or more selectively to understand specific neuropsychological functioning. The following sections discuss the history, test structure and scoring, and validity of the KABC-II.

Historical Background and Development

The Kaufman Assessment Battery for Children (K-ABC)

The KABC-II is a conceptual and structural revision of the K-ABC that took 5 years to complete. It evolved out of the pioneering work of Alan Kaufman who served as the project manager for the revised version of the Wechsler Intelligence Scale for Children in 1974 by the Psychological Corporation, where he worked directly with David Wechsler. In 1979, Kaufman authored *Intelligent Testing With the WISC-R*, in which he introduced the concept of *intelligent testing* and suggested that examiners apply theoretical knowledge and clinical judgment flexibly in order to provide meaning to the scores obtained from intelligence tests. In 1983, he coauthored the K-ABC with his wife Nadine while working as a professor at the University of Georgia. Interestingly, the K-ABC development team also included several of his doctoral students, some of whom went on to publish psychoeducational tests on their own.

The K-ABC was a revolutionary instrument at the time of its publication and was praised for its rigorous standardization procedures and sophisticated validity studies. Psychometric researchers often credit its technical validation with setting the standard for future tests. The K-ABC was also heavily influenced by the neuropsychological theories of A. R. Luria and Roger Sperry during its conceptualization and was the first test to integrate cognitive psychology into intelligence testing. In contrast to the popular Wechsler Scales, which primarily emphasized the measurement of general intelligence (*g*), the K-ABC test structure emphasized multiple cognitive components, including sequential and simultaneous processing. Because of the de-emphasis of *g* and the exclusive focus on elements of cognitive processing, it was suggested that the K-ABC was a more useful instrument for appraising the cognitive abilities of culturally and linguistically diverse individuals who were thought to be the subject of bias in traditional IQ tests. To support this notion, validity studies in the K-ABC technical manual provided evidence that the difference in scores between Black and White examinees was less than that often reported in more conventional measures such as the Wechsler Scales. Additionally, the K-ABC provided users with a global achievement scale, making it the first cognitive test developed to be a comprehensive psychoeducational measure. Accordingly, it became a popular instrument in learning disability evaluations conducted by school and educational psychologists as well as for clinicians seeking an alternative to the Wechsler and Stanford-Binet Intelligence Scales.

From the outset, the K-ABC generated tremendous controversy and was subjected to numerous research investigations. As an example, in 1984, a special issue of the *Journal of Special Education* was devoted entirely to the K-ABC. The majority of those inquiries focused on the validity of the K-ABC theoretical framework and its structural model. As an example, an independent confirmatory factor analysis conducted by Tim Keith and Stephen Dunbar in 1984 suggested that K-ABC measures were consistent with contemporary models of intellectual ability, including a higher order general factor and that users should interpret the scores on the instrument with caution. Also in 1984, Arthur Jensen suggested that the diminished Black–White differences on the K-ABC were largely the result of statistical artifacts caused by the lower *g*-loadings on K-ABC measures. Nevertheless, the controversies engendered by the K-ABC remained largely unresolved when the process for its revision began in 1996.

The KABC-II

Published in 2004, the KABC-II was a major revision and restructuring of the K-ABC based on the hierarchical model of intelligence known as the Cattell–Horn–Carroll (CHC) model. Eight subtests were eliminated from the original K-ABC, and 10 measures were created and added to the current battery. Item discrimination and scale ranges were increased, and the theoretical foundation was updated from sequential-simultaneous processing theory. One of the unique features of the KABC-II is the flexibility that it affords the examiner in determining the interpretive model to administer to the examinee. Although examiners may select either the Luria or CHC models, the KABC-II manual advises users to interpret the KABC-II primarily from the CHC perspective.

Test Structure and Scoring

The KABC-II utilizes a dual theoretical foundation: (1) the CHC psychometric models of broad and narrow abilities and (2) elements of A. R. Luria’s neuropsychological theory of cognitive processing that were fixtures of the previous version of the instrument. The KABC-II core battery takes between 30 and 75 minutes to administer depending on the examinee’s age and the interpretive model that is selected.

CHC Model

The CHC model of intellectual abilities is hierarchical, with 50–60 narrow abilities at the bottom (Stratum I), 7–9 broad ability factors in the middle (Stratum II), and a general ability dimension (*g*) at the top (Stratum III). The model features 16 subtests (10 core and six supplemental), which combine to yield five first-order factor scale scores (short-term memory, long-term storage and retrieval, visual processing, fluid reasoning [Gf], and crystallized ability) as well as a second-order full-scale composite named the Fluid Crystallized Index (FCI) that is thought to represent psychometric *g*. Each CHC factor scale is composed of two subtest measures, and the FCI is derived from a linear combination of the 10 core subtests that compose the constituent factor scores.

Luria Model

The Luria interpretive model differs from the CHC model in terms of both factor structure (e.g., four vs. five factors) and content, specifically as it relates to the inclusion of measures of acquired knowledge. The Luria model emphasizes the role of cognitive processing while de-emphasizing acquired knowledge (i.e., it omits measures of crystallized ability from the CHC model). The factor-level scores and hierarchical structure mimic the CHC model; the only putative difference is how those variables are labeled. In keeping with the K-ABC lineage, the first-order factors are labeled as planning (Gf), learning (long-term storage and retrieval), simultaneous processing (visual processing), and sequential processing (short-term memory). The eight core subtests in the Luria model combine to form a second-order full-scale composite named the mental processing index (MPI).

According to the KABC-II manual, the Luria interpretive model is preferred in a variety of situations, including, but not limited to, examining individuals from culturally and linguistically diverse backgrounds, assessing individuals known or suspected of having autism spectrum disorder, and examining individuals with hearing or language deficits.

Nonverbal Index (NVI)

A separate Nonverbal Index, composed of subtests that do not require verbal output, is also provided as an alternative to the FCI/MPI. The NVI provides an estimate of overall cognitive ability for use with examinees who have severe hearing loss, limited English proficiency, or moderate to severe speech or language disorders.

Scoring

All composite and factor scores are based on a mean of 100 and an *SD* of 15. Full-scale composite scores range from 40 to 160, covering a wide range of intellectual abilities (±4 *SD*s). This allows for the assessment of intelligence from the lower levels of moderate intellectual disability to the higher levels of giftedness. Factor-level scores range from 48 to 160, providing a wide range of possible scores (sequential/short-term memory: 49–158, simultaneous/visual processing: 50–160, learning/long-term storage and retrieval: 48–160, planning/Gf: 51–160, and knowledge/crystallized ability: 48–160).

Recommendations for interpretation of KABC-II scores include the full-scale composite and comparisons of performance on the various combinations of first-order factor scores. In fact, the KABC-II manual suggests that users should focus most, if not all, of their interpretive weight on the factor-level scores.

Validity of the KABC-II

The total norming sample of the KABC-II (*N* = 3,025) was nationally stratified based on sex/gender, ethnicity, geographic region, and level of parent education (as a proxy for socioeconomic status) and was proportional to 2001 U.S. census estimates. Extensive normative and psychometric data can be found in the KABC-II manual. Mean internal consistency estimates were high for the factor scores (.88 to .93), the MPI/FCI (.95 to .97), and the NVI (.90 to .92). Validity evidence is provided in several forms in the KABC-II manual. It should be noted that not all of the CHC/Luria model factor scores could be replicated from ages 3 to 6; thus, the complete structural/theoretical models are only available from ages 7 to 18.

Independent Validity Studies

All independent reviews noted improvements over the K-ABC but also noted some problems; namely, no structural validity evidence to support the Luria interpretive model was provided in the KABC-II manual.

The authors of the KABC-II relied exclusively upon restricted confirmatory factor analyses to examine the structural validity of the instrument. For ages 7–18, a five-factor CHC measurement model was reported although standardized path coefficients between *g* and Gf were problematic (1.0 and 1.01) in the final models, suggesting that *g* and Gf were indistinguishable. Subsequent independent CFAs of the KABC-II have tended to support the structure described in the KABC-II manual. In 2007, Matthew Reynolds and colleagues found that the five-factor CHC measurement model was a better fit to the KABC-II data set than other rival measurement models and that the model was invariant across age-groups. Consistent with the results reported in the KABC-II manual, the path loading between *g* and the Gf factor in the final model approached unity. Additionally, Reynolds and colleagues utilized a latent variable approach to decomposition subtest variance and found that all of the measures contained nontrivial portions of *g* variance (16–53%). Not surprisingly, an incremental validity investigation conducted by Ryan McGill in 2015 found that the CHC factor scores consistently accounted for trivial proportions of criterion achievement score variance after controlling for the effects of the more parsimonious FCI score, challenging the interpretive recommendations in the KABC-II manual.

Given the limited evidence provided in the KABC-II manual to support use of the Luria interpretive model, McGill and Angelia Spurgin also conducted a series of psychometric investigations in 2015 to appraise the utility of Luria model scores. Use of the same exploratory factor analytic techniques employed by John Carroll failed to support the theoretical four-factor model posited by the test authors. In fact, forcing the theoretical model resulted in weak subtest loadings, impermissible factors, theoretically inconsistent subtest migration, and nontrivial cross-loading of measures. McGill was also able to replicate the CHC incremental validity results, finding that the Luria factor scores contributed weak increments of predictive achievement variance after controlling for the MPI.

In sum, the KABC-II has many strengths but also has some weaknesses. Although interpretation of the full-scale composites appears to have strong empirical support, as of 2016, more research is needed to support confident clinical interpretation of the lower order factor scores.

Ryan J. McGill

See also Buros *Mental Measurement Yearbook*; Cattell–Horn–Carroll Theory of Intelligence; Construct-Related Validity Evidence; Exploratory Factor Analysis; Intelligence Tests; School Psychology; Stanford–Binet Intelligence Scales; Wechsler Intelligence Scales; Woodcock-Johnson Tests of Cognitive Ability

Further Readings

Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Assessment Battery for Children–Second Edition. Circle Pines, MN: American Guidance Service.

McGill, R. J. (2015). Interpretation of KABC-II scores: An evaluation of the incremental validity of CHC factor scores in predicting achievement. Psychological Assessment, 27, 1417–1426. doi:10.1037/pas0000127

McGill, R. J., & Spurgin, A. R. (2015). Exploratory higher order analysis of the Luria interpretive model on the Kaufman Assessment Battery for Children–Second Edition (KABC-II) school-age battery. Assessment. doi:10.1177/1073191115614081

McGill, R. J., & Spurgin, A. R. (2016). Assessing the incremental value of KABC-II Luria model scores in predicting achievement: What do they tell us beyond the MPI? Psychology in the Schools, 53, 677–689. doi:10.1002/pits.2194

Reynolds, M. R., Keith, T. Z., Fine, J. G., Fisher, M. E., & Low, J. (2007). Confirmatory factor structure of the Kaufman Assessment Battery for Children–Second Edition: Consistency with Cattell–Horn–Carroll theory. School Psychology Quarterly, 22, 511–539. doi:10.1037/1045-3830.22.4.511

Mira B. Kaufman Mira B. Kaufman Kaufman, Mira B.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Kelly Grid

Kelly grid

922

923

Kelly Grid

The Kelly grid, also known as a repertory grid, is an interviewing method used to measure personality. The grid identifies the unique mental representations participants use to organize their life experiences, and it maps connections between them to determine how the participant interprets the world. This entry describes the process of using the Kelly grid and discusses the implications of this technique in personality assessment and within clinical, educational, and professional contexts.

The Kelly grid was developed in the 1950s by George Kelly and was adapted from his personal construct theory, which states that individuals organize their life experiences into a unique system of categorized mental representations or *constructs*. According to Kelly's theory, personality is composed of the specific constructs one uses to form a concept of reality. These constructs can then be applied to new situations to assess their validity in predicting or explaining certain experiences, thus helping the individual to make sense of the world. A Kelly grid identifies and connects these constructs to represent a unique and coherent system by which individuals interpret their life. In doing so, it might reveal a pattern of constructs that a participant considers to be related, thus demonstrating the participant's expectation to encounter these constructs in conjunction with the everyday world.

The Kelly Grid Process

During the Kelly grid process, the examiner first chooses one topic or domain relevant to the participant's life (e.g., family) and creates a set of elements representative of that domain (e.g., mother, father, sibling). The participants then compare groups of three elements, determining in which way two of them differ from the third; this difference represents a construct that they use to classify objects or experiences. After identifying as many constructs as possible, the participant rates every element on a 5-point scale as it pertains to each construct and its bipolar contrast (e.g., for the construct "good," the participant rates each element on a scale from "bad" to "good"). Finally, the examiner analyzes these ratings to determine how the different constructs are related and in what ways they represent the participant's attitudes toward the topic.

Implications of the Kelly Grid in Assessment

The Kelly grid is considered a relatively unbiased way to assess personality, as participants take an active part in their evaluation. Many standardized personality measures contain questions designed by examiners for a generalized population, which may therefore unintentionally reflect the constructs most important to examiners rather than to participants and disregard variation among individual participants. Kelly grids attempt to eliminate these biases by allowing participants to identify the constructs most relevant to themselves.

Since its establishment as a form of clinical assessment, the Kelly grid has been utilized in a variety of fields, including education (e.g., to assess teacher effectiveness) and the workforce (e.g., to assess attitudes about management styles). When applied in nonclinical settings, particularly to elicit responses from multiple participants about the same subject, Kelly grids have proven limited in their ability to generate cohesive patterns. Compared to standardized measures, however, the Kelly grid yields systematic feedback that is tailored to the perceptions and interpretations of the participant.

Mira B. Kaufman and Marc H. Bornstein

See also Cluster Analysis; Educational Psychology; Personality Assessment; Psychometrics; Tests

Further Readings

Faccio, E., Castiglioni, M., & Bell, R. C. (2012). Extracting information from repertory grid data: New Perspectives on clinical and assessment practice. New perspectives, 19, 177–196.

Fransella, F., Bell, R., & Bannister, D. (2004). A manual for repertory grid technique. Chichester, UK: Wiley.

Johnson, M., & Nádas, R. (2012). A review of the uses of the Kelly’s Repertory Grid method in educational assessment and comparability research studies. Educational Research and Evaluation, 18, 425–440. doi:10.1080/13803611.2012.689715

Kelly, G. A. (1963). A theory of personality: The psychology of personal constructs. New York, NY: W. W. Norton … Company.

Claudia A. Gentile Claudia A. Gentile Gentile, Claudia A.

Kindergarten

Kindergarten

923

925

Kindergarten

Kindergarten was developed in the late 18th century in what is now southern Germany as a preschool experience for children whose parents worked outside of the home. These original kindergarten programs consisted of playing, singing, drawing, and social activities intended to help young children transition from home to school. Today, in the United States, kindergarten is part of the K–12 education system. In some states, full-or half-day kindergarten is compulsory, whereas in others, it is offered and recommended but not required. Children are usually eligible for entry into kindergarten when they are 5 to 6 years old.

Kindergarten in the United States originally followed the European model and included programs that developed young children’s social and emotional skills as well as their readiness for academic learning (especially in reading and mathematics). More recently, with many children enrolled in preschools and an increased emphasis on academic achievement at all elementary grade levels, some kindergarten programs have become not a transition year but the first year of formal schooling, with the expectation that children will leave kindergarten having acquired some language skills (e.g., knowing the alphabet; identifying uppercase and lowercase letters; recognizing, blending, and segmenting letter sounds; and writing simple sentences) and numeracy skills (e.g., identifying and writing numbers 1 through 20 or higher, adding and subtracting numbers 1 through 10 or higher).

Research on Kindergarten

Much of the research on kindergarten has focused on the relationship between development in kindergarten and learning in later grades. Recent research has explored the relationship between students' socioeconomic status, their school readiness upon entering kindergarten, and their success trajectories through elementary school and has found that achievement and behavioral gaps between higher and lower income children are present at school entry and tend to increase over time.

Some longitudinal studies have found that children's social and emotional readiness for schooling contributes to their successful transition to and academic progress through school and have called for programs that support social and emotional development as a means of potentially reducing the achievement gap. Other studies have evaluated efforts designed to close the readiness gap through activities such as community outreach, family support services (including home literacy programs), and transitional kindergarten programs (in which an extra year of kindergarten study provides students time to develop fundamental skills needed for success in school). In addition, researchers have explored the relationship between children's home language and first language literacy with their English language skill development and academic development through kindergarten and into the first grade.

Other studies have focused on the structural aspects of kindergarten programs, such as what the optimal age is for beginning kindergarten and whether full-day or half-day programs provide a better foundation for students. Much of the early research on the full-versus half-day debate was inconclusive, but recent studies have found that students who attended full-day kindergarten had significantly greater achievement by the end of the kindergarten year and through the first grade. With the recent focus on academic achievement in kindergarten, researchers have studied specific approaches for teaching reading, writing, mathematics, and science, as well as how to best support English language learners.

One of the most comprehensive studies of kindergarten, the Institute of Education Sciences' Early Childhood Longitudinal Study, collected information about the early educational experiences of a nationally representative sample of children who were in kindergarten in the 2010–2011 school year; the children were observed through the spring of 2014, when most of them were in the third grade. The Institute of Education Sciences' Early Childhood Longitudinal Study explored differences in the performance of subgroups on reading, mathematics, and science assessments and found results that were similar to many other national studies of students at higher grade levels. Students from economically national studies of students at higher grade levels. Students from economically advantaged households outperformed those from economically disadvantaged households. Male students outperformed female students in mathematics, but not in reading and science. Also, students with a primary home language of English outperformed those with non-English home languages in all three subjects, whereas White and Asian students, and students of two or more races, outperformed Black and Hispanic students.

In addition to focusing on programs and student outcomes, researchers have studied the training, perspectives, and experiences of kindergarten teachers. A 2016 study by Aviva Sverdlov and Dorit Aram found that many kindergarten teachers believe that fostering children’s self-esteem is the most important goal of kindergarten, and promoting literacy and mathematics skills is the least important goal. Their views are consistent with recent calls from educators and parents to balance the recent academic focus in kindergarten with a more ecological approach, one that makes time for play, inquiry, and relationship building.

Challenges to Kindergarten Research

When exploring the relationship between different kindergarten programs and student development, one of the biggest challenges researchers face is that of how to validly and reliably assess young children. Researchers have found that some measures (such as the Teacher Rating Scale) appear to be sensitive to children’s social competence, whereas other measures (such as the Mathematics and Literacy Achievement Tests) are more sensitive to children’s anxiety. Researchers have also explored whether measures are reliable when used to rate the school readiness of children from different ethnic and language backgrounds, finding that the Kindergarten Student Entrance Profile exhibited measurement invariance across student ethnicities (Latino/White) and home languages (Spanish/English). Moreover, a logistical challenge to kindergarten research is that, while some assessments can be completed by teachers or observers using checklists or rubrics, most academic assessments require tests to be administered to students during one-on-one sessions.

Claudia A. Gentile

See also Childhood; Classroom Observations; Montessori Schools

Further Reading

Further Readings

Graue, E. (2011). Are we paving paradise? Educational Leadership, 68(7), 12–17.

Hammes, P. S., Bigras, M., & Crepaldi, M. A. (2016). Validity and bias of academic achievement measures in the first year of elementary school. International Journal of Research … Method in Education, 39(1), 3–18.

Mulligan, G. M., McCarroll, J. C., Flanagan, K. D., & Potter, D. (2016). Findings from the third-grade round of the early childhood longitudinal study, kindergarten class of 2010–11 (ECLS-K: 2011): First look. Washington DC: U.S. Department of Education. IES: NCES 2016-094.

Quirk, M., Mayworm, A., Edyburn, K., & Furlong, M. J. (2016). Dimensionality and measurement invariance of a school readiness screener by ethnicity and home language. Psychology in the Schools, 53(7), 772–782.

Sverdlov, A., & Aram, D. (2016). What are the goals of kindergarten? Teachers' beliefs and their perceptions of the beliefs of parents and of agents in the education system. Early Education and Development, 27(3), 352–371.

Ali S. Brian Ali S. Brian Brian, Ali S.

Jacqueline D. Goodway Jacqueline D. Goodway Goodway, Jacqueline D.

Kinesthetic Learning Kinesthetic learning

Kinesthetic Learning

Kinesthetic learning occurs as a result of exploring and discovering through movement. Also known as learning-by-doing, kinesthetic learning involves the whole body in gaining knowledge and skills both within and outside of the psychomotor domain. Kinesthetic learning is often misrepresented as tactile or hands-on learning; however, it requires a total body approach to be most effective. For example, someone learning to throw a ball should actually practice throwing for maximal force rather than walking through a slow-motion version of the skill. This entry discusses the two main ways of assessing kinesthetic learning: *process measures* and *product measures*. Examples of each assessment type are provided, supplemented by optimal strategies for teaching kinesthetic learners.

Assessment of Kinesthetic Learning

Process-Based Measures

Process-based measures look at *how* a performer engages in a movement. They typically involve a qualitative decision that is then quantified to express whether critical elements or components of a movement pattern are present when performed. For example, a process measure of the skill of throwing may include critical elements such as stepping with opposition, rotating the hips and shoulder with the forearm lagging, snapping the ball with the wrist, projecting the ball toward a target, and following through to the opposite hip. Using a process-based approach, an evaluator would score a 0 if a critical element was not present and a 1 if it was present. There are a number of valid and reliable process-oriented assessments in the kinesthetic domain that are often used in the research and diagnosis of kinesthetic/motor delays. These instruments include the Test of Gross Motor Development–3 and the Peabody Developmental Motor Scales–2.

Another type of process-based assessment incorporates a more developmental perspective. Developmental sequences of movement, which begin with inefficient movement patterns and progress to more proficient ones, have been identified for many fundamental motor skills. For example, there are five stages to learning how to throw a ball, each reflecting qualitatively different patterns of movement that progressively become more efficient in their performance. Throwing starts with a Stage 1 "dart throw" and ends with a Stage 5 throwing pattern similar to a baseball pitch. Overall, there are total body stages for 10 skills, including locomotor skills such as jumping, hopping, running, galloping, and skipping, as well as manipulative skills such as catching, kicking, punting, two-hand striking (e.g., hitting a ball with a bat or racquet), and throwing. For each skill, three to five developmental stages have been identified, and age-related norms from 0 to 120 months have been defined. Assessing the total body stage of a psychomotor skill and plotting it against an age-related norm can provide an instructor with a snapshot of a child's performance from a kinesthetic learning standpoint. Identification of a child's developmental stage for a skill also allows the teacher to determine whether an activity is appropriate, and it assists in aligning kinesthetic movement conditions to the child's developmental level. Finally, understanding total body stages can help an instructor modify task and environmental constraints, such as types of equipment, methods of feedback, and other visual and physical prompts, to a child's current level of performance to elicit a more proficient movement pattern.

Product-Based Measures

Product-based measures examine the *outcome* of a movement, not how a person performs it, to assess kinesthetic learning. Any defined result of a movement pattern can be included in a product measure. Examples include throwing a ball at a target, kicking a ball through a goal, and shooting a basketball into a hoop. Other types of product measures include distance (e.g., how far a child jumps or hops), speed (e.g., running, throwing, and kicking velocities), percentage (e.g., the number of times a ball thrown hits a target relative to the number of attempts), and duration (e.g., how long a child can hang from a bar in a pull-up position). Within the psychomotor domain, valid and reliable product measures typically used for diagnostics, research, and formal assessments include the FitnessGram, the Brockport Physical Fitness Test, the Movement Assessment Battery for Children–2, and the Bruininks-Oseretsky Test of Motor Proficiency.

Ensuring Success for Kinesthetic Learners

Should instructors focus on process measures or product measures of kinesthetic learning first? A general rule is that they should focus on process measures first, teaching an individual how to perform the skill with proficiency. As a person increases in proficiency, the person is more likely to attain product outcomes with consistency. Thus, improvements in process outcomes are associated with increases in product outcomes of kinesthetic learning. For example, as children develop a better throwing pattern (process), they are more likely to throw for longer distances and faster velocities (product outcomes).

How do children learn kinesthetically? One of the most important determinants of kinesthetic learning is frequent, high-quality movement experiences during early childhood. National guidelines suggest that children need to be given multiple, daily, structured and unstructured opportunities to move, with a specific focus on fundamental motor skills. These fundamental motor skills, such as throwing, running, and catching, are considered the ABCs of movement that make up sports, games, and lifetime activities. If children do not receive structured movement experiences facilitated by a knowledgeable adult, they often do not have the necessary kinesthetic learning to continue to be physically active across their lifespan. Thus, early opportunities for kinesthetic learning are critical for children.

Ali S. Brian and Jacqueline D. Goodway

See also Childhood; Puberty

Further Readings

Gagen, L. M., & Getchell, N. (2006). Using “constraints” to design developmentally appropriate movement activities for early childhood education. Early Childhood Education Journal, 34(3), 227–232. doi:10.1007/s10643-006-0135-6

Gallahue, D., Ozmun, J., & Goodway, J. D. (2012). Understanding motor development: Infants, children, adolescents and adults (7th ed.). New York, NY: McGraw-Hill.

Goodway, J. D., Brian, A., Chang, S., & Park, S. (2014). Methodological approaches to evaluate motor skill development, physical activity, health-related fitness, and perceived motor competence in young children. In O. N. Saracho, et al., Handbook of research methods in early childhood education. Charlotte, NC: Information Age Publishing.

Hildie Leung Hildie Leung Leung, Hildie

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Kohlberg’s Stages of Moral Development Kohlberg’s stages of moral development

927

930

Kohlberg’s Stages of Moral Development

Moral development is concerned with the emergence, change, and understanding of morality throughout the life span. In the context of moral development, morality is defined as the rightness or wrongness of action guided by ethical principles such as justice, equality, and dignity. This entry introduces Lawrence Kohlberg’s stages of moral development, describing how the theory grew out of early research on children’s development, then looking at each stage and how it is assessed. This entry then discusses empirical support for and critiques of the theory and looks at how the theory applies to education.

Kohlberg built on the ideas of Jean Piaget, who was the first to investigate moral reasoning in children by interviewing and observing children’s interactions during games with rules. He then proposed three stages of moral reasoning: the premoral stage, where infants and young children have no sense of obligation to rules; the heteronomous stage, where children aged 4–8 obey rules imposed by external parties to avoid punishment; and the autonomous stage, where children aged 8–12 begin to consider rules and motives behind actions critically based on principles of reciprocity, mutual respect, and cooperation.

Although Piaget’s attempts advanced the understanding of moral development at that time, his methodology lacked scientific rigor. As such, building on Piaget’s theorization, Kohlberg conducted systematic studies with children and adolescents in analyzing their responses to hypothetical moral dilemmas to provide evidence in a 1969 article that moral reasoning develops in a progressive fashion as one ages. Based on the findings, Kohlberg proposed six stages arranged in three levels to conceptualize moral development by outlining the arranged in three levels to conceptualize moral development by outlining the cognitive processes underlying the development of moral reasoning from childhood to adulthood.

Kohlberg stressed the cognitive basis of moral judgment in its relation to moral actions. Although his stages of moral development theory has been critiqued, it remains one of the most influential and cited theories in developmental psychology, as it provides an empirically tested structural framework to understanding moral development across the life span. The remainder of this section describes each of the six stages Kohlberg proposed.

Level I: Preconventional Morality

The first level in Kohlberg’s theory is the preconventional/premoral level, which is generally observed among children in early-to-middle childhood, beginning at around age 4. At this level, individuals are primarily concerned with themselves. Moral values are prescribed not by the individual but rather by authority figures. Actions are determined based on physical and hedonistic consequences, where right or good behaviors are associated with avoiding punishments, gaining rewards, or an exchange of favors.

The first stage in the preconventional level is known as punishment-obedience orientation. Individuals at this stage are compelled to behave in accordance with socially acceptable norms as externally prescribed by authority figures such as parents or teachers. This unquestioning obedience is bound merely by the threat or application of punishment, as one has yet to consider the underlying meaning or rationale (e.g., maintaining order of society) behind morality.

The second stage is the instrumental-relativist orientation. At this stage, individuals begin to become aware that others also have needs. They attempt to choose actions that satisfy others’ needs, given that their own needs are met. Human relations are perceived as an instrumental exchange in a pragmatic manner. Behaviors are guided by the “you scratch my back and I’ll scratch yours” principle, in an instrumental way, but not based on a sense of loyalty, gratitude, or justice. Children often consider whether something is “fair,” yet they are not actually concerned with upholding justice. They are primarily concerned with doing something right in order to obtain rewards.

Level II: Conventional Morality

The second level is the conventional/role conformity level, as seen among children in middle-to-late childhood. At this level, individuals are motivated to conform to conventions and rules of society, but no longer unquestioningly. One is able to comprehend and appreciate the value of morality aimed at maintaining order and expectations of one’s family, group, or nation, regardless of immediate consequences.

The third stage is the good boy–nice girl orientation. At this stage, one chooses to behave in ways that would please or impress others, especially authority figures or popular peers. Good behaviors are those that would result in praise and approval. One is concerned about maintaining positive relationships through trust, sharing, and loyalty and thus begins to engage in perspective taking, by considering the views of others and one’s intentions in the decision-making process.

The fourth stage is the law and order orientation. Individuals at this stage move beyond preoccupation with immediate groups (e.g., family, peers) to concern about the society. Right behaviors consist of performing one’s duties and obligations. One perceives society as a system of fixed rules and understands that any deviations from prescribed rules would result in chaos. However, at this stage, rules are perceived as inflexible, that is, no person or group is above the law.

Level III: Postconventional Morality

The third level is the postconventional/self-accepted moral principle level, which emerges in adolescence or adulthood. Kohlberg believed, however, that this final stage is rarely achieved, even by adults. Individuals at this level take an active role to define moral values. One conforms to standards that are internal where decisions are made based on thought and judgment of what constitutes right or wrong.

Stage 5 is known as the social-contract legalistic orientation, where morality is understood in terms of social mutuality. Norms of right actions are rationally analyzed and agreed upon by the whole society aimed at protecting individual rights and social order, as opposed to inflexible rules that must be strictly obeyed simply because they are the law. Rules are open to question and ones that are no longer able to serve the community’s best interest may be changed. Individuals at this stage adopt a utilitarian approach where the value of actions is guided by “the greatest good for the greatest number of people” principle. One holds respect for the law behind the law.

Stage 6 is the universal ethical principle orientation. Behaviors are guided by individual conscience consisting of abstract, universal principles such as justice, equality, and dignity, which transcend social norms and rules. Individuals answer to their inner conscience and are willing to disobey the law if it happens to violate their personal ethical principles.

Kohlberg asserted that individuals go through the stages sequentially, progressing from Stage 1 to Stage 2, then Stage 3 and so forth. One does not skip any stages or move in a mixed-up sequence. Also, Kohlberg argued that the stages are hierarchically integrated. Individuals who are at a higher stage of moral reasoning are still able to comprehend the insights and motives from earlier stages. As one progresses from a lower to a higher stage of moral reasoning, one’s framework is broadened to consider more factors and abstract values in one’s decision-making and judgment process.

Assessing Moral Reasoning

The most widely used measurement of moral reasoning is the Moral Judgement Interview, where hypothetical moral dilemmas are presented to participants who are then asked to make a decision and provide justification and reasoning for the decision. In the dilemmas, competing values are pitted against each other (e.g., obeying the law vs. upholding life). The Heinz dilemma is one of Kohlberg’s most famous moral dilemmas:

> Heinz’s wife was dying from a unique type of cancer. A local druggist discovered a new drug that might save her but was charging for it at an extremely high price. Heinz could not afford the drug. He was only able to raise half of the money after seeking help from his family and friends. He explained the situation to the druggist and asked if he could purchase the drug at a lower cost or pay the remaining amount to him subsequently. However, the druggist refused.

Respondents are asked to indicate whether the husband should violate the law and break into the druggist’s laboratory to steal the otherwise unobtainable drug that could save his dying wife. The emphasis is not placed on whether a participant's response is to "steal" or "not to steal" the drug but rather the reasoning behind the decision. The following are examples of responses reflective of each stage.

Stage 1 (punishment-obedience orientation)

"Heinz should not steal the drug because he will be sent to prison if caught."

Stage 2 (instrumental-relativist orientation)

"Heinz should steal the drug because he could save his wife, then she can take care of him."

Stage 3 (good boy–nice girl orientation)

"Heinz should steal the drug as he will be praised as a good husband."

Stage 4 (law and order orientation)

"Heinz should not steal the drug as it is illegal to steal."

Stage 5 (social-contract legalistic orientation)

"Heinz should steal the drug as individuals have the right to life regardless of what the law stipulates. If Heinz is caught for stealing, then the law should be reconsidered given that one's life is at stake."

Stage 6 (universal ethical principle orientation)

"Heinz should steal the drug, as preserving human life is a more fundamental value than property rights."

Empirical Support for and Critiques of Kohlberg's Stages of Moral Development

Scholars have empirically tested the theory proposed by Kohlberg. For instance, studies have been conducted on and found support for Kohlberg's claim that children’s moral reasoning followed an invariant sequence, and there was no evidence of regression or skipping of stages. A review of cross-cultural studies has also supported universality.

Generally speaking, the content of the moral dilemmas in Kohlberg’s model and interview were adapted to fit local context, and participants were interviewed in their native language. However, an interesting finding was that participants from traditional tribal or folk societies rarely achieved postconventional level of moral reasoning. Kohlberg reasoned that this may be due to a lack of exposure to the cognitive and social experiences needed for mature reasoning.

While Kohlberg’s theory provides a clear structure of moral development and a framework to understand how individuals develop a sense of morality, the theory has also been criticized. In terms of methodology, Carol Gilligan argued that Kohlberg’s all-male sample was biased and problematic. She argued that males often conceptualize morality in terms of abstract principles of law and justice, whereas females may take into consideration the notions of care and compassion, which are neglected in Kohlberg’s theory.

Kohlberg’s use of moral dilemmas to assess moral reasoning has also been challenged. Particularly, scholars argued that the dilemmas are hypothetical, meaning that one’s decision will result in no real consequences. Yet in real life, when one is confronted with moral dilemmas, the decision one makes often results in actual consequences that may yield fundamental impact on one’s interest or well-being. It is unclear whether participants would make the same decisions in reality. In addition, the dilemmas are artificial. Kohlberg’s sample consisted of children and adolescents who have never been married or experienced any situations that may resemble the presented moral dilemma (e.g., the Heinz dilemma), so it may be difficult for them to identify with the protagonist, which may influence their moral reasoning.

In terms of theorization, studies have found a lack of consistency in one’s stages of moral development across contexts. For instance, a participant who reasons based on postconventional moral principles at Stage 5 to one particular moral dilemma may regress to reasoning using conventional principles in another dilemma. Therefore, scholars have questioned whether distinct stages of moral development exist and if all individuals progress as proposed by Kohlberg. Finally, one of Kohlberg’s major assumptions is that moral reasoning is based fundamentally on the principle of justice. However, Gilligan asserted that principles of care and compassion are important, especially for females principles of care and compassion are important, especially for females.

Moral Development and Education

A traditional approach to moral development is character education, where teachers teach universal moral values to students based on a list of virtues predetermined by the school. However, this indoctrinating method has been critiqued for the fact that the preached values are defined and determined by teachers and reflective of the curriculum. Thus, students adopt a rather passive role.

Studies have shown that when children merely listened to adults' moral judgments, their resultant change in moral reasoning was quite small. Therefore, in order to reorganize children's thinking, they must take on a more active role. The values clarification approach has been proposed as a rational approach to moral education, where students are encouraged to form personal opinions and make judgments surrounding issues where values are in conflict. This pedagogy stresses open discussion of moral dilemmas in the classroom. It is believed that this approach enables students to become more aware of their own and others' moral values and stimulates progression from lower to higher stages of moral reasoning.

Hildie Leung and Daniel Tan-lei Shek

See also Bayley Scales of Infant and Toddler Development; Behaviorism; Cognitive Development, Theory of; Erikson's Stages of Psychosocial Development

Further Readings

Gibbs, J. C. (2013). Moral development and reality: Beyond the theories of Kohlberg, Hoffman, and Haidt. New York, NY: Oxford University Press.

Kohlberg, L. (1975). The cognitive-developmental approach to moral education. The Phi Delta Kappan, A Special Issue on Moral Education, 56(10), 670–677.

Kohlberg, L. (1981). Essays on moral development, Vol. I: The philosophy of moral development. San Francisco, CA: Harper … Row.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. Theory Into Practice, 16(2), 53–59.

Lapsley, D., & Carlo, G. (2014). Moral development at the crossroads. Developmental Psychology, 50(1), 1–7.

Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. Psychological Bulletin, 97(2), 202–232.

Walker, L. J. (1982). The sequentiality of Kohlberg’s stages of moral development. Child Development, 53(5), 1330–1336.

Lori A. Thombs Lori A. Thombs Thombs, Lori A.

Kolmogorov–Smirnov Test Kolmogorov–smirnov test

930

932

Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov (KS) test is a statistical procedure for comparing the distribution of random samples. The one-sample KS test can be used to determine whether a data set follows any hypothesized (but fully specified) continuous density. Perhaps its most common use is to verify whether a data sample follows the normal (or Gaussian) density, such as checking the assertion that residuals from a fitted regression model follow the normal density. In the two-sample case, the KS procedure tests whether two data samples have equal underlying distributions. An example is comparing assessment scores for two different schools, when it is known that the scores do not follow the normal distribution.

This entry describes the basic principles of the KS test, including the cumulative distribution function (CDF) and its role in both the one-and two-sample settings. Details about the null hypothesis (H_0), alternative hypothesis (H_a), test statistic, and decision rule for the test are presented. Illustrative examples are also provided in each section.

CDF

The KS test is based on the idea of the CDF (denoted by F_z). For a continuous random variable X, the CDF is the probability that X is less than or equal to x:

$$F_x = P\left(X \leq x\right).$$

A parametric density f_x, such as the normal mean (μ) and variance (σ^2), or the exponential θ, has a known CDF. Note that a sample is not needed to specify F_x

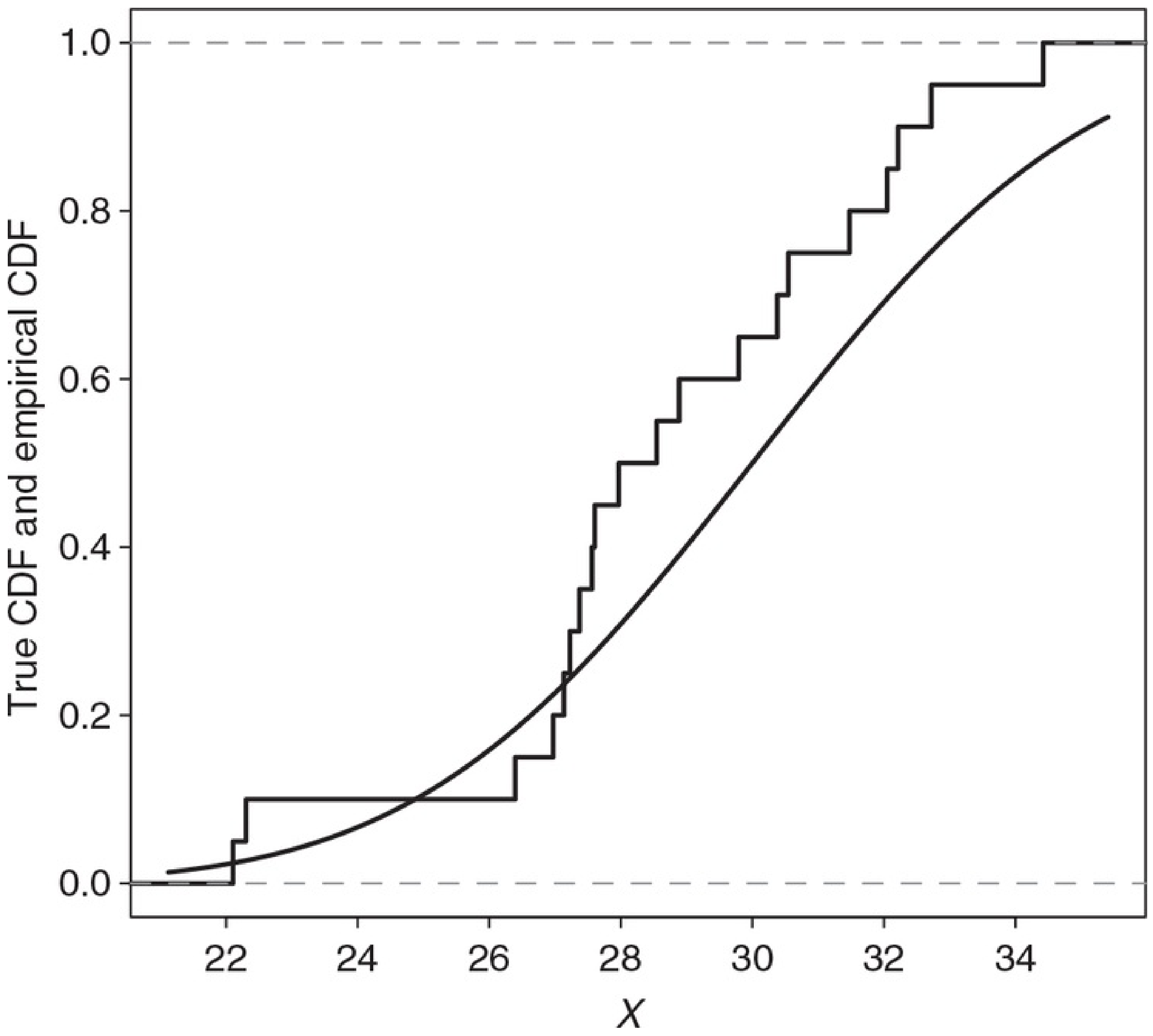
in this case because it uses only the functional form of the density f_x. The function $F(x)$ is sometimes referred to as the *true CDF* or the *theoretical CDF* for the random variable X.

Now suppose a random sample of size n, denoted by $(x_1, x_2, \ldots, x_n)$ is available, and an estimate of the true CDF is desired. The sample CDF $F_n(x)$ is defined as follows:

$$F_n(X) = \text{number of observations} \leq x_n = x_i \leq x_n.$$

If the sample values x_i are unique, the sample CDF is a step function, with jumps of height $1/n$ at each observation. Another term for the sample CDF is the *empirical distribution function*. To illustrate these concepts, a plot of a normal CDF and a sample CDF computed from a sample of size n = 20 are presented in Figure 1.

Figure 1 Plot of the empirical and theoretical CDF for one sample



One-Sample Case

The idea behind the one-sample KS test is to compare the sample CDF to the theoretical CDF and determine whether the largest difference between these two functions is statistically significant. The null and alternative hypotheses are H_0: $X \sim F$ (i.e., the sample X has the same underlying density as the theoretical CDF) versus the general “not equal to” alternative, H_a, where sample X does not follow the distribution defined by CDF F. The test statistic for the KS test (D) is

$$D = \frac{\max}{\text{all } x_i} \mid F_n(x_i) - F(x_i) \mid .$$

Graphically, this statistic is the largest vertical distance observed between F and F_n. It is sufficient to compute the distance between F_n and F at each x_i because the maximum always occurs at an observed data value.

Deciding if the observed value D of the test statistic is large enough to reject H_0 involves computing its p value. Statistical computing packages such as SPSS, SAS, and R are used here because the null distribution of the KS statistic is complicated. If the p value is small, say less than a significance level of α (e.g., an error rate of 5% or α = .05), this implies that the observed difference is too extreme, and the null hypothesis is rejected and H_a is accepted.

Revisiting the data from Figure 1, suppose we wish to test whether this sample comes from a normal density with mean of 30 and an *SD* of 4, which were values used to plot the true CDF. The value of the KS test statistic is $D = .2096$, and the p value associated with this computed test statistic is $p = .2999$. Because this p value is not small, we would conclude that there is not sufficient statistical evidence to refute the assertion that these data are from a normal(30,4) density.

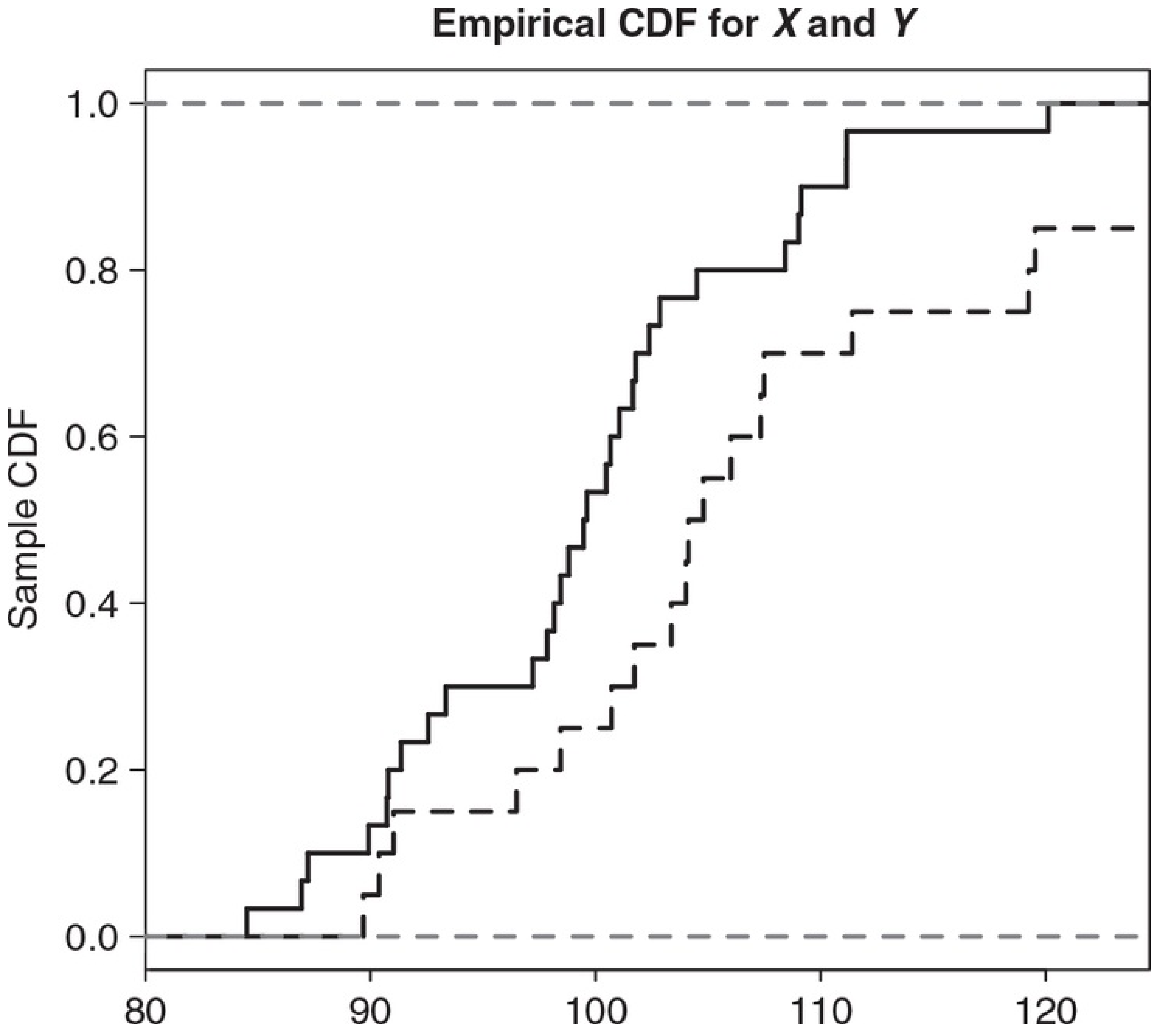
Often, one wishes to test an underlying distribution but does not know the values of the parameters of this density. In this case, the KS test is technically called the Lilliefors test, which tests the closeness of the data to the hypothesized distribution when the parameters have been estimated. Returning again to the example, if the mean and standard deviation are not known, they must be estimated to compute the assumed CDF F, and this slightly changes the theory of the null distribution of D. If we carry out the KS test on the data from the previous example, this time not specifying the values of the parameters, the test statistic and p value change to $D = .1439$, $p = .3412$. Although these results are slightly different from the first example that assumed known mean and variance, the conclusion is the same—insufficient statistical evidence to refute the assertion that these data are from a normal density. This method is often still called the KS test but with Lilliefors correction.

Two-Sample Case

For a two-sample case, the null hypothesis asserts that the two underlying densities (or CDFs) for samples are the same, or H_0: $F_X = F_Y$, and the alternative hypothesis is H_a: $F_X \neq F_Y$. The test statistic is the maximum difference that is observed between the two sample CDFs. Denote these two estimates, computing from samples (x_1, x_2, …, x_n) and (y_1, y_2, …, y_m) by F_n, X and F_m, Y, respectively. The test statistic is again the largest observed difference between the two estimated functions, D = max(F_n, X – F_m, Y), which is usually computed using a statistical computing package. The decision rule is to reject H_0 if the p value associated with D is less than α.

Suppose two data sets X (sample size n = 28) and Y (sample size m = 16) are available, and we wish to test the null hypothesis that these two samples have the same underlying density. Figure 2 depicts the sample CDFs for these data, where the largest observed difference between X (solid line) and Y (dotted line) is D = .4167, with a computed p value of .0244. Using a significance level of 5%, we would reject H_0, because the p value is less than α = .05, and conclude that the distributions for X and Y are not the same.

Figure 2 Plot of the sample CDF for two data sets



Concluding Remarks

The KS procedure is not the only statistical method that can be used for testing normality of a single sample. Other well-known methods are the Shapiro–Wilk test, the Anderson–Darling test, and the Cramér–von Mises test. The statistical computing packages SPSS, SAS, and R can be used to carry out all of these tests. Finally, it is important to remember that the KS test is appropriate for continuous distributions only. For discrete data and testing for a specific discrete probability mass function, the chi-square goodness-of-fit test should be used.

Lori A. Thombs

See also Chi-Square Test; Goodness-of-Fit Tests; Normal Distribution

Further Readings

Birnbaum, Z. W., (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. Journal of the American Statistical Association, 47, 425–441.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribution. Giornale dell' Istituto Italiano degli Attuari, 4, 83–91.

Lilliefors, H. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62, 399–402.

Massey, F. J. (1951). The Kolmogorov–Smirnov test for goodness of fit. Journal of the American Statistical Association, 46, 68–78.

Jie Chen Jie Chen Chen, Jie

KR-20

KR-20

932

936

KR-20

The Kuder and Richardson Formula 20 (KR-20), published in 1937, checks the internal consistency of items on a test. The internal consistency, or *reliability*, refers to the degree to which all of the test items measure a common characteristic of the examinees and are free from measurement error, which can be random (e.g., an examinee’s mood or health condition) or systematic (e.g., the sound of traffic outside). As an indicator of stability of performance, reliability is desired in any test. This entry defines and explains how to interpret the KR-20, describes its applications and limitations, and provides an example of calculating the KR-20 statistic using the software program SPSS.

Definition and Interpretation

The KR-20 is a special case of Cronbach’s α in which the items are binary variables (i.e., test answers that are either right or wrong, as opposed to answers graded on a scale). Correct answers are scored as 1 and incorrect as 0. The formula for KR-20 for a test with *k* items is:

$$\rho_{KR-20} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i q_i}{\sigma_x^2}\right),$$

where *k* = number of questions, p_i = proportion of correct responses to test item *i*, q_i = proportion of incorrect responses to test item *i*, σ^2 = variance of the total scores of all the examinees.

The variance is expressed as:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n},$$

where *n* is the total sample size.

Values of ρ KR-20 generally range from 0 to 1, but they can be below 0 if the sample size is small. The closer the value is to 1, the more reliable the test. A value of .90 or more indicates a homogeneous test where every item measures the same general trait of ability or personality as every other item. However, homogeneity (or unidimensionality) is an assumption, not a conclusion, of reliability coefficients. It is possible to have a high KR-20 with a multidimensional scale, especially with a large number of items.

The interpretation of the KR-20 depends on the purpose of the test. Most high-stakes examinations are intended to distinguish students who have mastered the material from those who have not. For these, a KR-20 of .50 or higher is desired. A KR-20 of less than .30 is considered poor no matter the sample size. If the purpose of the test is to ensure that all students have mastered essential skills or concepts, a KR-20 close to 0 is desired.

Applications and Limitations

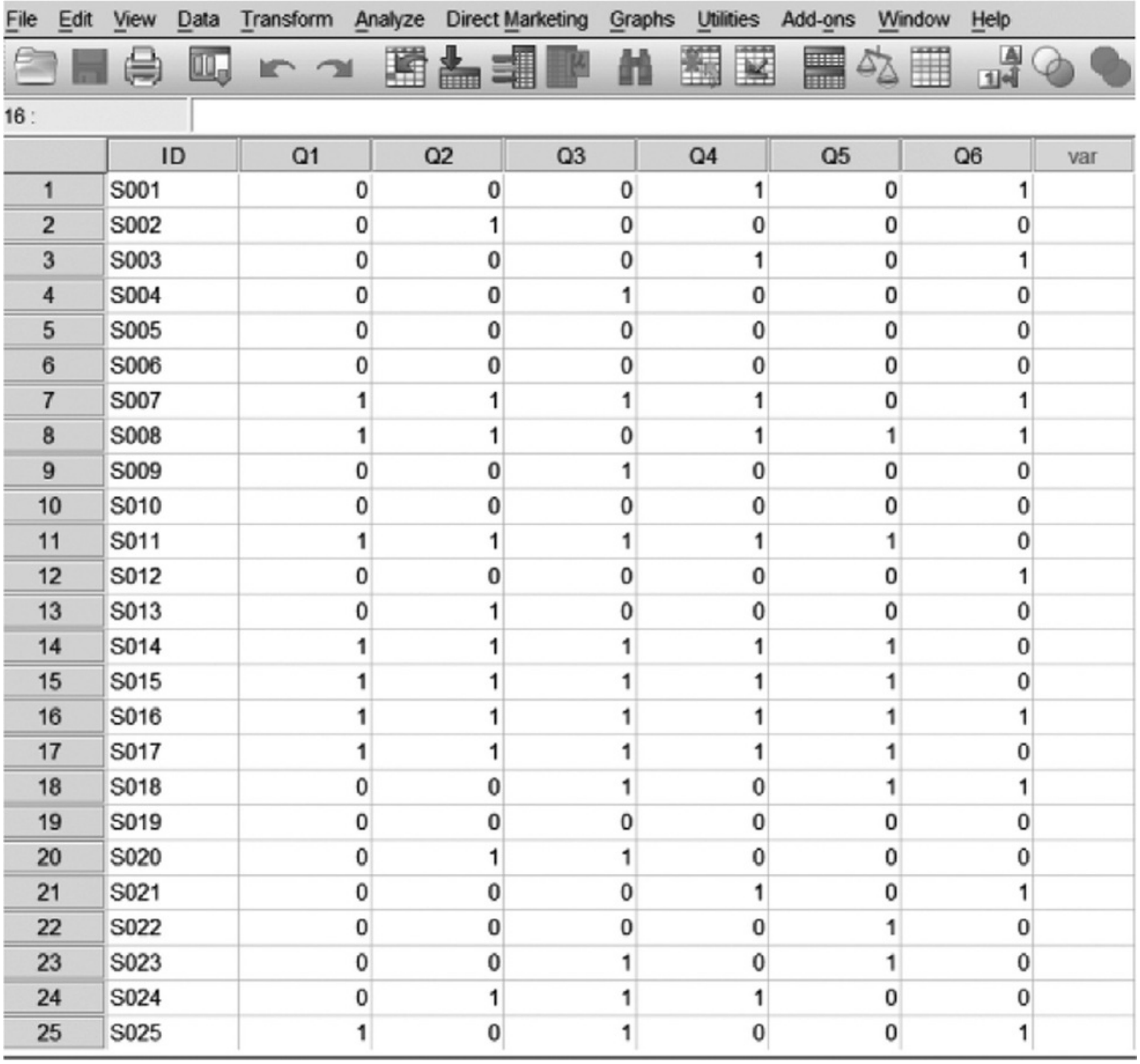
The KR-20 is used for items that have varying difficulty (i.e., some items are easy and some are challenging). If a test has questions with more than two answer possibilities, Cronbach's α should be used to measure reliability. If all questions in a binary test are equally challenging, a simplified version of KR-20, called the KR-21, should be used. The KR-20 may be affected by difficulty of the test, the spread in scores, and the length of the test.

According to Robert Thorndike, the KR-20 has several limitations. First, it only provides evidence on the precision with which we can assess an examinee at a specific moment. Variation from day to day cannot be reflected. Second, a set of items based on common reference material (e.g., reading items based on a single passage) are more alike than truly independent items. Thus, an examinee who succeeds on 1 item of the set is more likely to succeed on the other items of the set than items that are not in the set, resulting in an artificially high reliability coefficient. Third, a single-administration reliability coefficient becomes meaningless in a timed test because a low reliability would primarily be due to the difference in test-taking speed instead of a difference in performance.

Using SPSS to Calculate the KR-20

Figure 1 shows a small data set with 25 participants and their responses to six test questions in SPSS. All questions are scored as 1 if the answer is *correct* and 0 if *incorrect*.

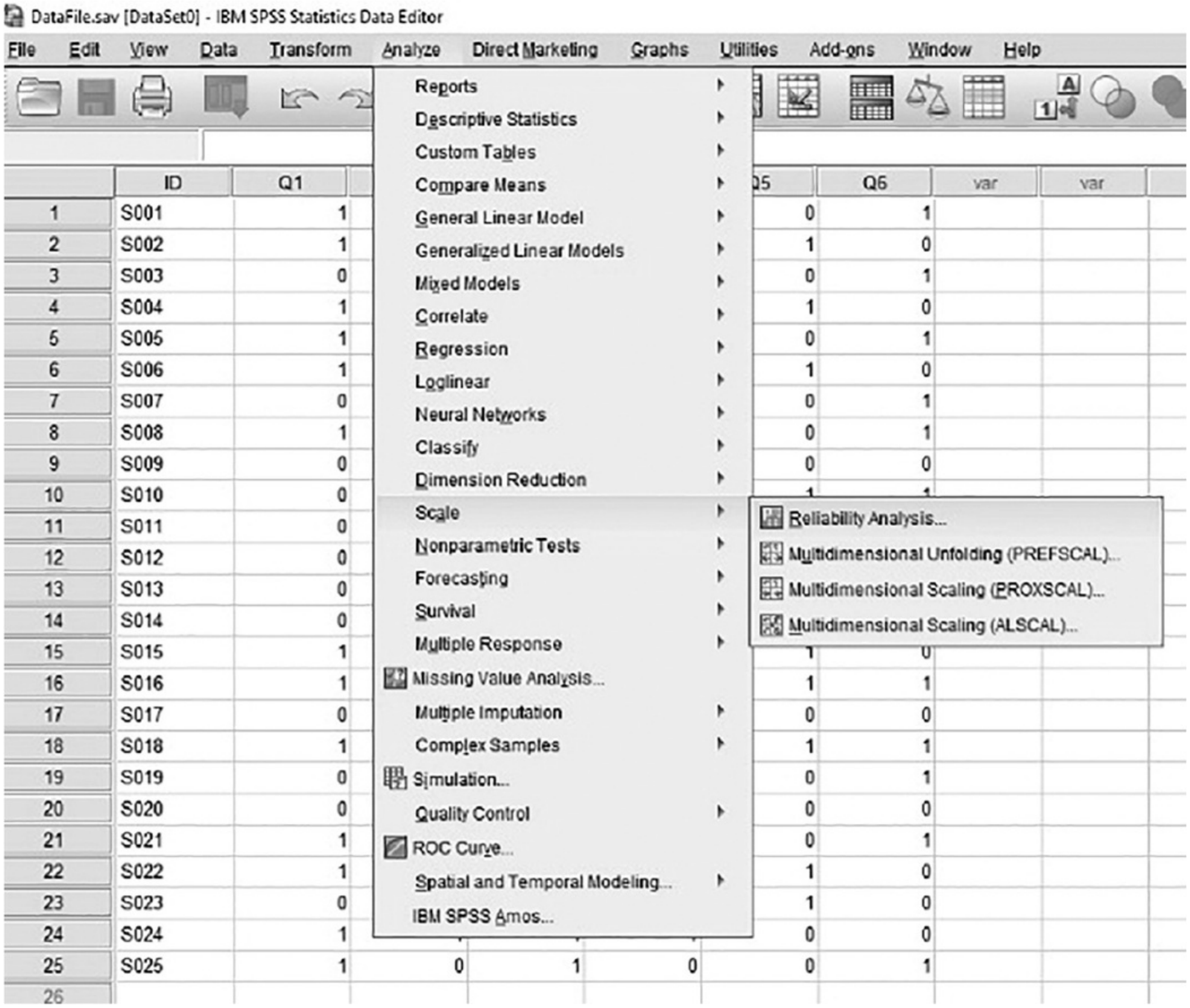
Figure 1 An example of a data set in SPSS



	ID	Q1	Q2	Q3	Q4	Q5	Q6	var
1	S001	0	0	0	1	0	1	
2	S002	0	1	0	0	0	0	
3	S003	0	0	0	1	0	1	
4	S004	0	0	1	0	0	0	
5	S005	0	0	0	0	0	0	
6	S006	0	0	0	0	0	0	
7	S007	1	1	1	1	0	1	
8	S008	1	1	0	1	1	1	
9	S009	0	0	1	0	0	0	
10	S010	0	0	0	0	0	0	
11	S011	1	1	1	1	1	0	
12	S012	0	0	0	0	0	1	
13	S013	0	1	0	0	0	0	
14	S014	1	1	1	1	1	0	
15	S015	1	1	1	1	1	0	
16	S016	1	1	1	1	1	1	
17	S017	1	1	1	1	1	0	
18	S018	0	0	1	0	1	1	
19	S019	0	0	0	0	0	0	
20	S020	0	1	1	0	0	0	
21	S021	0	0	0	1	0	1	
22	S022	0	0	0	0	1	0	
23	S023	0	0	1	0	1	0	
24	S024	0	1	1	1	0	0	
25	S025	1	0	1	0	0	1	

To conduct the reliability analysis, select “reliability analysis” from the pop-up menu under Analyze/Scale (Figure 2) and follow the steps in the subsequent windows to generate the output.

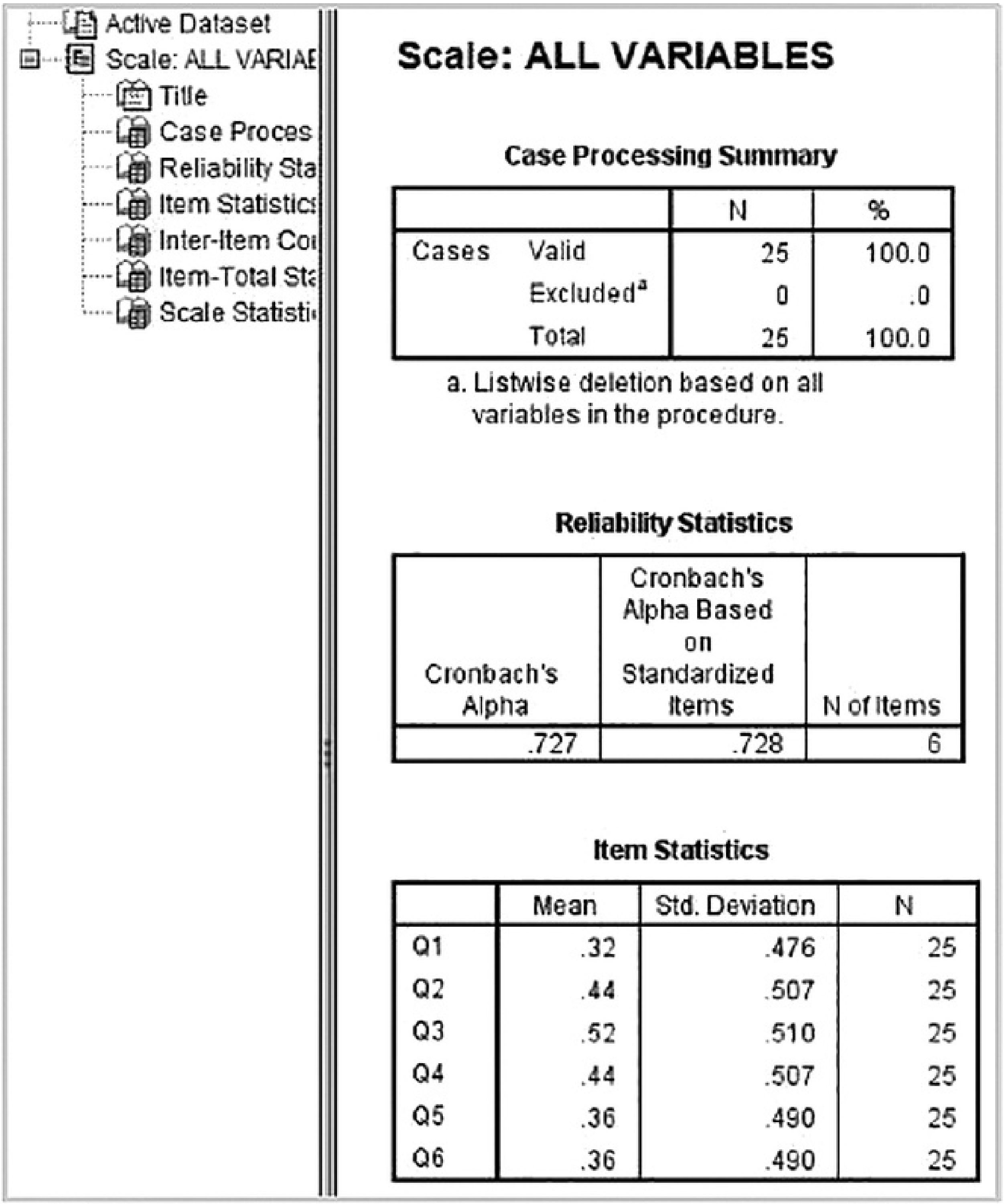
Figure 2 Steps to conduct a reliability analysis



	ID	Q1				Q5	Q6	var	var
1	S001	1				0	1		
2	S002	1				1	0		
3	S003	0				0	1		
4	S004	1				1	0		
5	S005	1				0	1		
6	S006	1				1	0		
7	S007	0				0	1		
8	S008	1				0	1		
9	S009	0				0	0		
10	S010	0				1	1		
11	S011	0							
12	S012	0							
13	S013	0							
14	S014	0							
15	S015	1				1	0		
16	S016	1				1	1		
17	S017	0				0	0		
18	S018	1				1	1		
19	S019	0				0	1		
20	S020	0				0	0		
21	S021	1				0	1		
22	S022	1				1	0		
23	S023	0				1	0		
24	S024	1				0	0		
25	S025	1	0	1	0	0	1		
26									

Figure 3 shows the first part of the analysis output. The top table confirms there are 25 observations in the data set. The middle table shows that the Cronbach’s α for the binary variables (i.e., the KR-20 statistic) is .727, indicating that all items are measuring the same construct. A value above .70 is considered adequate, a value above .80 is considered optimal, and anything closer to 1 is even better. The bottom table shows the descriptive statistics of each item.

Figure 3 The first part of the reliability analysis output



Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	25	100.0
	Excluded[a]	0	.0
	Total	25	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.727	.728	6

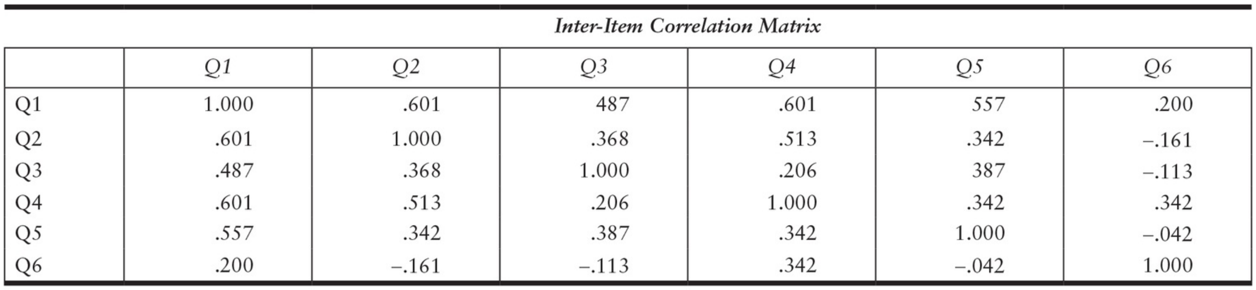
Item Statistics

	Mean	Std. Deviation	N
Q1	.32	.476	25
Q2	.44	.507	25
Q3	.52	.510	25
Q4	.44	.507	25
Q5	.36	.490	25
Q6	.36	.490	25

In Figure 4, the matrix table at the top shows interitem correlations. A correlation coefficient smaller than .30 indicates a low correlation and is not desired. A negative value indicates that 2 items are measuring different constructs. The table shows that Item 6 has low correlations with the other items.

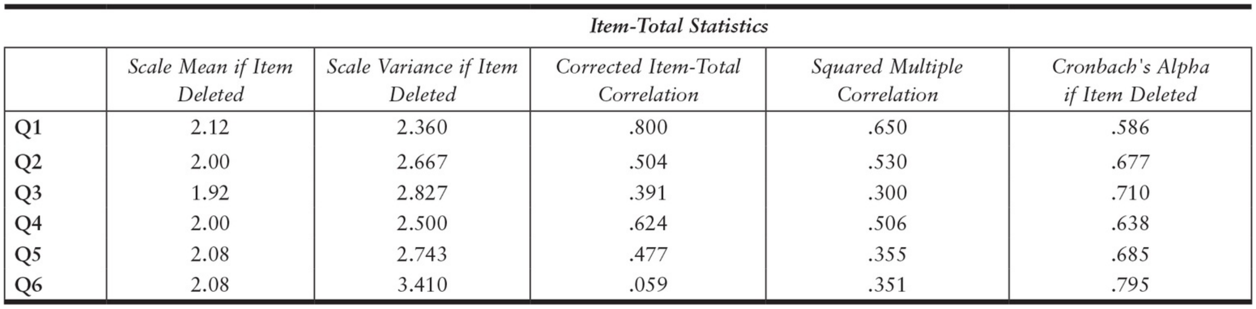
The bottom table presents the overall internal consistency after deleting each of the items; the column on the far right tells us which variable is lowering the Cronbach’s α.

Figure 4 The second part of the reliability analysis output



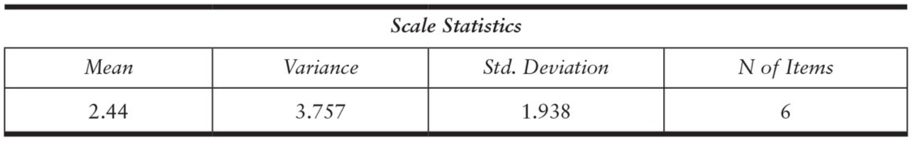
Inter-Item Correlation Matrix

	Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*
Q1	1.000	.601	487	.601	557	.200
Q2	.601	1.000	.368	.513	.342	–.161
Q3	.487	.368	1.000	.206	387	–.113
Q4	.601	.513	.206	1.000	.342	.342
Q5	.557	.342	.387	.342	1.000	–.042
Q6	.200	–.161	–.113	.342	–.042	1.000



Item-Total Statistics

	Scale Mean if Item Deleted	*Scale Variance if Item Deleted*	*Corrected Item-Total Correlation*	*Squared Multiple Correlation*	*Cronbach's Alpha if Item Deleted*
Q1	2.12	2.360	.800	.650	.586
Q2	2.00	2.667	.504	.530	.677
Q3	1.92	2.827	.391	.300	.710
Q4	2.00	2.500	.624	.506	.638
Q5	2.08	2.743	.477	.355	.685
Q6	2.08	3.410	.059	.351	.795



Scale Statistics

Mean	*Variance*	*Std. Deviation*	*N of Items*
2.44	3.757	1.938	6

The reliability analysis of this test shows that Question 6 might be measuring a different construct from the other test questions. The overall test reliability will improve from .727 (adequate) to .795 (nearly optimal) if we remove this question.

Jie Chen

See also Reliability; SPSS

Further Readings

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160.

Thorndike, R. M. (2005). Measurement and evaluation in psychology and education (7th ed.). Upper Saddle River, NJ: Pearson Education.

Wright, B., & Stone, M. (1999). Measurement essentials. Wilmington, DE: Wide

Range.

John T. E. Richardson John T. E. Richardson Richardson, John T. E.

Kruskal–Wallis Test

Kruskal–wallis test

936

939

Kruskal–Wallis Test

One of the simplest designs in quantitative research involves the random allocation of a sample of *N* individuals to *k* different groups. The groups are exposed to different treatments, and the research question is whether there is any variation among the groups on some criterion variable. Classically, this question is addressed using the one-way between-subjects analysis of variance. However, this procedure assumes that the criterion variable in question (a) is measured on an interval or ratio scale, (b) is normally distributed, and (c) has the same variance in all of the groups. The Kruskal–Wallis one-way analysis of variance by ranks (to give the test its full name) was developed for use in situations in which one or more of these assumptions is not met. This entry describes the original derivation of the Kruskal–Wallis test, provides a simple worked example, discusses exactly what the test is measuring, assesses the test's power and power efficiency, and considers alternative tests that might be used as well as or instead of the test.

Analysis of Variance by Ranks

In the 1940s, a number of statisticians had suggested the use of ranks to compare two different groups (as in the Mann-Whitney test). In 1952, two American statisticians, William H. Kruskal and W. Allen Wallis, built upon this previous work to propose that the same approach could be used to compare three or more groups. In their procedure, the original observations are replaced by the numbers from 1 to *N*, where 1 refers to the smallest observation and *N* refers to the largest observation across all of the *k* groups.

Suppose that n_i is the number of cases in the ith group, that R_i is the sum of the ranks in the ith group, and that R is the sum of the ranks across all k groups. The deviation of the mean of the ranks in the ith group from the overall mean is $[(R_i/n_i) - (R/N)]$. Kruskal and Wallis defined a statistic which they denoted as H as the sum of the squared standardized deviations across the k groups. However, this can be simplified computationally because the ranked data are simply the integers from 1 to N. As a consequence, the sum of all the ranks, R, is $N(N + 1)/2$, the mean of all the ranks is $(N + 1)/2$, and the variance of all the ranks is $(N^2 - 1)/12$. This enabled Kruskal and Wallis to express their test statistic in terms of the following formula:

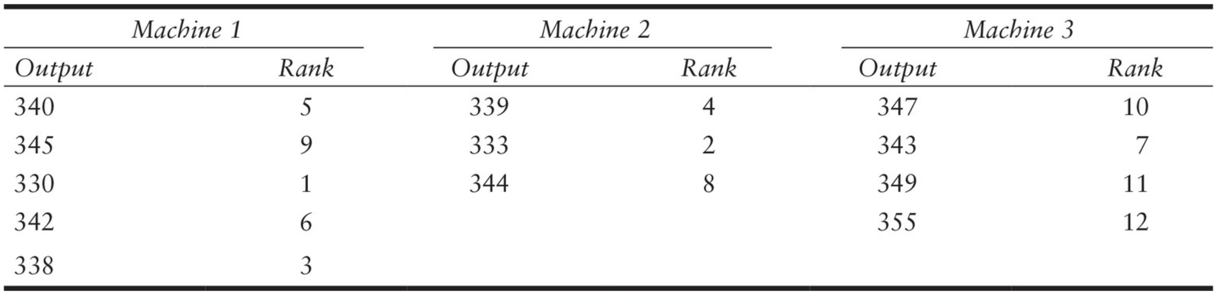
$$H = \{12/[N(N+1)]\}\left[S\left(R_i^2/n_i\right)\right] - 3(N+1)$$

where summation is carried out across the k groups.

For values of n_i up to 5, Kruskal and Wallis presented tables showing the true probability of obtaining a value of H or a higher value under the null hypothesis that the observations in the different groups were drawn from identical populations. For larger values, they proposed that the mean ranks (i.e., R_i/n_i) would be approximately normally distributed under the null hypothesis; consequently, H itself would be distributed as chi-squared (X^2) with $(k - 1)$ degrees of freedom, and it could be evaluated using readily available tables of X^2. They themselves called this “the H test.” It assumes that the original observations are measured on at least an ordinal scale, that they are independent of one another, and that those within each group are drawn from the same population. However, it does not make any assumptions about the parameters of the populations from which the data are drawn; hence, it is an example of a nonparametric statistical test.

A Worked Example

Kruskal and Wallis gave the hypothetical example of three machines that had been designed to produce large numbers of bottle caps. Machine 1 was a standard machine, but Machines 2 and 3 had been modified in different ways. Table 1 shows the total number of bottle caps produced by each machine on a number of different days, together with the ranks of the observations from 330 (the lowest, ranked 1) to 355 (the highest, ranked 12). For these data, the sums of the ranks for Machines 1, 2, and 3 are 24, 14, and 40, the value of is 115.2 + 65.333 + 400 = 580.533, and the value of *H* is 5.656. Kruskal and Wallis noted that the exact probability of obtaining a value of 5.656 or greater under the null hypothesis that the output of the three machines was identical was .049, whereas the approximation to that probability that was yielded by the X^2 distribution with 2 degrees of freedom was .059.



Machine 1		*Machine 2*		*Machine 3*	
Output	*Rank*	*Output*	*Rank*	*Output*	*Rank*
340	5	339	4	347	10
345	9	333	2	343	7
330	1	344	8	349	11
342	6			355	12
338	3				

Source: Adapted with permission from Kruskal and Wallis (1952). Copyright by the American Statistical Association.

If *k* = 2, then the Kruskal–Wallis test is formally equivalent to the Mann-Whitney test. At the same time, it can be generalized from *k* = 3 to any number of groups. The example in Table 1 was deliberately chosen to avoid tied observations. Kruskal and Wallis proposed that tied values should be assigned the mean of the ranks in question. For instance, the two lowest observations in Table 1 are 330 and 333, which are assigned the ranks of 1 and 2. If they had both been 330, they would both have been assigned the rank of 1.5. Kruskal and Wallis described a procedure that could then be used to adjust the value of *H* for tied values. However, they noted that in many situations, the difference between the original value of *H* and the adjusted value was negligible, and so the adjustment made very little practical difference to the outcome. Even so, it is a simple matter to apply the adjustment using modern statistical packages, which typically report values of *H* and their associated significance levels, both uncorrected and corrected for ties.

What Is the Kruskal–Wallis Test Measuring?

Kruskal and Wallis noted that, strictly speaking, the statistic *H* was measuring a tendency for observations in at least one of the populations to be larger (or smaller) when paired randomly with observations in all of the other populations. András Vargha and Harold D. Delaney called this situation one of *stochastic heterogeneity*. In principle, this might arise if the *k* populations differed in their variability rather than in their means. Nevertheless, Kruskal and Wallis argued that their procedure was relatively insensitive to any differences in variability across the groups. In most practical situations, therefore, the statistic *H* was measuring whether the mean of at least one population differed from those of the others. However, this has been called into question.

The situation of *stochastic homogeneity* is one where the probability of an observation in each of the populations being larger or smaller when paired randomly with the observations in all of the other populations is exactly .5. In this case, Vargha and Delaney showed that the expected values of the mean ranks R_i / n_i were the same. They went on to show that the Kruskal–Wallis test was a valid test of the hypothesis of stochastic homogeneity only if the variance of the ranks was the same across the *k* groups. As they commented, this is not as restrictive as the assumption of homogeneity of variance in the observations themselves. Even so, if this assumption is violated, Vargha and Delaney recommended the use of a robust parametric test on the ranks instead. They also showed that, if the relevant distributions were symmetrical, then the stochastic homogeneity would imply equality of the group medians. However, it would only imply equality of the group *means* if the samples were drawn from populations with the same shape and variance.

Power and Power Efficiency

The *power* of a statistical test is the probability of rejecting the null hypothesis when it is false. (Its complement is the probability of *not* rejecting the null hypothesis when it is false, in other words the probability of making a Type II error.) In general, nonparametric tests tend to be less powerful than the corresponding parametric test because they use less of the information that is contained in the data. (For instance, the Kruskal–Wallis test employs only the ranks of the observations, whereas the parametric analysis of variance employs the actual values of the observations.) In their original account, Kruskal and Wallis admitted that they knew very little about the power of their *H* test. However, for the bottle-cap production data shown in Table 1, they noted that a parametric analysis of variance yielded the value of *F* with 2 and 9 degrees of freedom of 4.2282, and that the probability of obtaining this or a higher value under the null hypothesis was .051. This figure was close to the probability of .049 that had been yielded by their own test, suggesting that its power was similar to that of the parametric analysis of variance.

The power of two different statistical tests in the same research design can be compared using the notion of *power efficiency*. This notion relies upon the fact that the power of a test in a particular situation depends (other things being equal) on the sample size. Suppose that Test 1 is the most powerful statistical test when used in a particular research design with data that meet its underlying assumptions. Test 2 is a less powerful test in the same design, in that it would need to be used with a sample of N_2 cases to match the power that is achieved by Test 1 with N_1 cases (where $N_2 \geq N_1$). The power efficiency of Test 2 is N_1/N_2, often expressed as a percentage. In 1954, Fred C. Andrews demonstrated that the power efficiency of the Kruskal–Wallis test in comparison with the parametric analysis of variance approached a value of $3/\pi$ or 95.5% as the overall sample size increased. Accordingly, the Kruskal–Wallis test can be recommended as a powerful distribution-free test.

Further Tests

If the Kruskal–Wallis test yields a statistically significant result, this implies that at least one of the *k* samples is different from the other samples. However, in itself, it does not indicate where such differences may have arisen. (There are, of course, a number of procedures for carrying out post hoc tests in the context of a parametric analysis of variance.) Olive J. Dunn described a procedure for carrying out $k(k-1)/2$ pairwise comparisons among the *k* groups that incorporated a Bonferroni adjustment to maintain the overall Type I error rate. This in turn can be adapted for situations in which one of the samples is regarded as a control group with which the other $(k-1)$ samples are to be compared.

As just mentioned, the alternative hypothesis in the Kruskal–Wallis test is simply that at least one of the *k* samples is different from the others. However, in some contexts, a researcher may have more specific expectations regarding the pattern of differences that might arise among the *k* groups. The Terpstra–Jonckheere test constitutes a procedure that was originally based upon the Mann-Whitney test, for examining a specific trend across a series of ordered alternatives.

John T. E. Richardson

See also Analysis of Variance; Bonferroni Procedure; Mann-Whitney Test; Power; Rankings; Type I Error; Type II Error

Further Readings

Andrews, F. C. (1954). Asymptotic behavior of some rank tests for analysis of variance. Annals of Mathematical Statistics, 25, 724–736. doi:10.1214/aoms/1177728658

Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6, 241–252. doi:10.2307/1266041

Hettmansperger, T. P. (1984). Statistical inference based on ranks. New York, NY: Wiley.

Jonckheere, A. R. (1954). A distribution-free *k*-sample test against ordered alternatives. Biometrika, 41, 133–145. doi:10.1093/biomet/41.1-2.133

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583–621. doi:10.2307/2280779

Ruxton, G. D., & Beauchamp, G. (2008). Some suggestions about appropriate use of the Kruskal-Wallis test. Animal Behaviour, 76, 1083–1087. doi:10.1016/j.anbehav.2008.04.011

Siegel, S., & Castellan Jr., N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York, NY: McGraw-Hill.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall’s test against trend, when ties are present in one ranking. Indagationes Mathematicae, 14, 327–333. doi:10.1016/S1385-7258(52)50043-X

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics, 23, 170–192. doi:10.3102/10769986023002170

John T. E. Richardson John T. E. Richardson Richardson, John T. E.

Kurtosis

Kurtosis

939

946

Kurtosis

Kurtosis is a Greek word (κυ´ ρτωσις) denoting curvature, from *kurtos* (κυρτο´ ς) meaning convex or curved. (It is used in geometry to refer to the appearance of convex figures and in medicine to refer to the curvature of the spine seen in kyphosis. Greek writers prefer to render the word in English as *kyrtosis*.) Kurtosis was adopted by the British mathematician and statistician Karl Pearson in 1905 to describe the shape of frequency distributions in comparison with that of a normal curve: “If more flat-topped I term them *platykurtic* [i.e., of broad curvature], if less flat-topped *leptokurtic* [of thin curvature], and if equally flat-topped *mesokurtic* [of intermediate curvature]” (p. 173). Pearson went on to provide a quantitative definition of kurtosis that has since been widely used by statisticians. This entry discusses the statistical definition of kurtosis, the relation between kurtosis and “peakedness,” and the practical usefulness of kurtosis.

Statistical Definition of Kurtosis

The shape of a distribution can be characterized by its *moments*. If a variable is denoted by *X* and its distribution has the mean μ, its *r*th moment about the mean is in the case of a frequency distribution or for a probability distribution. Either way, the first moment about the mean is zero by definition, and the second moment about the mean is the variance, σ^2. Subsequent moments about the mean can be standardized by dividing them by . Pearson labeled the standardized third moment about the mean, , as β_1, and this is often used as a measure of the skewness of the distribution. Similarly, Pearson labeled the standardized fourth moment about the mean, , as β_2.

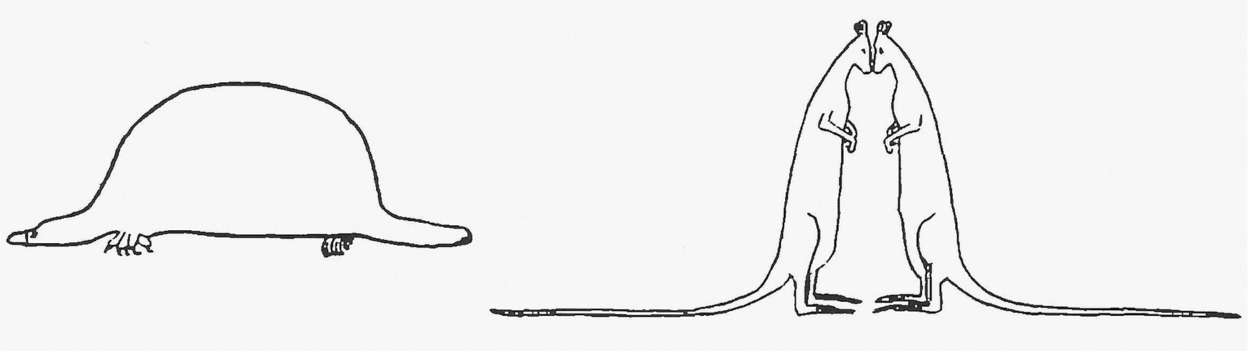
Pearson noted that $\beta_2 = 3$ for a normal distribution, and he defined the degree of kurtosis of the distribution as $\eta = \beta_2 - 3$ (other writers have used the symbol γ_2). This is sometimes known as “excess kurtosis” (i.e., beyond that of the normal distribution). Different writers have adopted either β_2 or γ_2 to refer to kurtosis itself, and for clarity, these symbols will be used for the rest of this entry. Pearson identified platykurtic distributions as those for which β_2 was less than 3 and leptokurtic distributions as those for which β_2 was greater than 3. It can be shown that , and thus $\beta_2 \geq 1$ and $\gamma_2 \geq -2$, with equality when a random variable takes only one of the two different values with equal probabilities. It can also be shown that $\beta_2 > 1 + \beta_1$ and, in the case of unimodal symmetric distributions, that $\beta_2 \geq 1.8$. However, there is no upper limit to β_2, which may be infinite for certain distributions.

Student’s *Memoria Tecknica*

In 1927, another British statistician, William S. Gosset (who published under the pseudonym of “Student”), suggested a *memoria tecknica* or mnemonic for remembering the difference between platykurtic and leptokurtic distributions. He presented the illustration shown in Figure 1 along with the following explanation:

> Platykurtic curves have shorter “tails” than the normal curve of error and leptokurtic longer “tails.” I myself bear in mind the meaning of the words by the above *memoria tecknica*, where the first figure represents platypus, and the second kangaroos, noted for “lepping,” though, perhaps, with equal reason they should be hares! (p. 160)

Figure 1 Student’s Memoria Tecknica



Source: Reprinted with permission from Student (1927), Copyright 2017 by Oxford University Press.

Kurtosis Versus Peakedness

During the rest of the 20th century, Student's mnemonic was reproduced by other writers when endeavoring to explain the notion of kurtosis. However, the illustration may itself have encouraged a fundamental misunderstanding about the nature of kurtosis by distracting the reader's attention from the size of the tails of distributions (as in Student's original explanation) to the size of their peaks. Certainly, over the next 30 years, the idea grew that the kurtosis of a distribution was essentially a matter of its "peakedness." This notion persisted in textbooks and online resources well into the 21st century.

Nevertheless, it has been known since the 1940s that is not a reliable indicator of the peakedness of a distribution. For instance, it was shown that distributions for which $\beta_2 > 3$ and distributions for which $\beta_2 < 3$ could both have peaks that were higher or lower than that of a normal curve. It was also shown that distributions with a mean of 0, a variance of 1, a skewness of 0, and $\beta_2 = 3$ may exhibit widely varying degrees of peakedness. Some of these distributions do approximate a normal distribution, but many do not. The main problem is that the values of β_2 and γ_2 are determined mainly by the shape of the tails and hardly at all by the shape of the peak. This is unsurprising because the formulas for kurtosis raise the deviations $(X - \mu)$ to the fourth power, thus accentuating the contribution of the larger deviations in the tails of the distribution at the expense of the smaller ones closer to the mean of the distribution.

How Useful Is Kurtosis?

How Useful Is Kurtosis?

The normal distribution has an important role in statistics and the applied sciences, and it is clearly important to be able to determine whether a distribution is normal in shape. $\beta_2 = 3$ or $\gamma_2 = 0$ is a necessary condition for a distribution to be normal in shape, but it is not a sufficient one. Consequently, the evaluation of kurtosis is of only limited practical value in determining whether a distribution is normal or not. Nevertheless, precisely because it is mainly influenced by the tails of a distribution, β_2 proves to be a useful statistic for the detection of outliers in small samples. It is probably for this reason that kurtosis should continue to be used as a property of distributions.

John T. E. Richardson

See also Moments of a Distribution; Normal Distribution; Skewness; Variance

Further Readings

Livesey, J. H. (2007). Kurtosis provides a good omnibus test for outliers in small samples. Clinical Biochemistry, 40, 1032–1036. doi:10.1016/j.clinbiochem.2007.04.003

Pearson, K. (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson [The law of error and its generalizations by Fechner and Pearson]: A rejoinder. Biometrika, 4, 169–212. doi:10.1093/biomet/4.1-2.169

Student. (1927). Errors of routine analysis. Biometrika, 19, 151–164. doi:10.1093/biomet/19.1-2.151

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. The American Statistician, 68, 191–195. doi:10.1080/00031305.2014.917055

L

Uta Jüttner Uta Jüttner Jüttner, Uta

Dorothea Schaffner Dorothea Schaffner Schaffner, Dorothea

Laddering

Laddering

943

944

Laddering

Laddering is a qualitative research technique in which the interviewer asks a series of questions with the aim of identifying higher order drivers of consumer decision making. This entry describes the theoretical foundations of laddering, elaborates its implementation, and discusses different applications.

The laddering method is derived from means-end theory. Central to means-end theory is the notion that individual behavior is driven by personal values or end states that individuals strive for in their lives. It draws on insights from motivation and cognitive psychology and is premised on the belief that product or service attributes serve as a means to satisfy individuals’ higher level consequences (functional and psychosocial) that derive their importance from satisfying personal values or goals. The following means-end chain from a car purchasing context shows a ladder as a sequence of elements at all four levels of abstraction:

- powerful engine (product attribute)
- ride comfort and sporty handling (functional consequences)
- confidence (emotional consequence)
- a comfortable life/hedonism (personal values).

The goal of the laddering interviewing procedure is to uncover how respondents translate attributes of products or services into the motivating elements that define why they are personally relevant. Moving respondents up the “ladder of abstraction” from attributes to functional and psychosocial consequences to personal values requires experienced interviewers to follow a general questioning protocol.

In a first step, personally meaningful service attributes are identified using elicitation-questioning techniques such as distinguishing between and among brands of products or services. In a second step, the facilitator uses a series of probes for each of the attributes elicited, typified by asking different versions of the question “Why is this important to you?” This continues until complete ladders with verbatim responses at all levels of abstraction are obtained.

Analyzing the laddering data across a sample of customers starts with a coding process in which each response is classified and assigned to the four levels. Subsequently, a summary table is constructed that indicates the number of linkages between the elements. Finally, the dominant associations within the summary table can be graphically represented in decision maps or hierarchical value maps. Overall, the qualitative data collection and analysis of the laddering technique enables the researcher to reveal the customers’ way of thinking and represent a discovery process. Therefore, hard-laddering advances that aim to standardize the process with self-administered questionnaires have generally been viewed critically. In contrast to the more widely used qualitative approach, hard laddering makes it impossible to detect the perceptual processing by the respondent.

Originally, laddering was developed within marketing to uncover drivers of consumer decision making. Within-marketing laddering has been most commonly used to design and develop new products, to improve marketing and advertising strategies, and to understand consumer satisfaction and choice. Additionally, laddering has been applied widely across a substantive number of research domains, ranging from American presidential politics to environmental behavior to health and safety issues.

The laddering technique has been criticized as a time-consuming and expensive technique as well as for its lack of standard statistical measures to assess data and solution quality. Online approaches based on new technologies might help to overcome some of these obstacles.

Uta Jüttner and Dorothea Schaffner

See also Interviews; Motivation; Qualitative Research Methods

Further Readings

Gutman, J. (1991). Exploring the nature of linkages between consequences and values. Journal of Business Research, 22(2), 143–148.

Olson, J. C., & Thomas, J. R. (2001). The means-end approach to understanding consumer decision making. In J. R. Thomas & J. C. Olson (Eds.), Understanding consumer decision making: The means-end approach to marketing and advertising strategy (pp. 3–20).

Reynolds, T., & Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. Journal of Advertising Research, 28(1), 1–31.

Reynolds, T. J., & Phillips, J. M. (2008). A review and comparative analysis of laddering research methods: Recommendations for quality metrics. Review of Marketing Research, 5, 130–74.

Bo Hu Bo Hu Hu, Bo

Latent Class Analysis Latent class analysis

944

948

Latent Class Analysis

The term *latent class* (LC) analysis refers to a class of statistical analyses that use the LC model to explain the associations among a set of observed variables. The LC models are advantageous generally because they bring in unobserved (latent) categorical variables, each category of which is defined as a subgroup. Thus, in LC models, the associations among observed variables are explained by the relationships between the latent categorical variables and each observed variable. LC models have been used in many applications in statistical analysis, such as clustering, diagnostic classification, density estimation, and dealing with unobserved heterogeneity. Further, LC models assume that observations within each subgroup are generated from an independent random process, and therefore, the distribution of overall observations can be seen as a mixture of the distribution of observations from each subgroup. In this sense, LC models are more generally referred to as finite mixture models.

This entry begins with a brief review of the history of LC analysis. Then, it introduces two approaches (i.e., probabilistic and log-linear) in parameterizing LC models. Further, the entry focuses on the basic methods for model estimation and model evaluation.

History of LC Analysis

The interest in LC models can be traced back to the late 19th century, when American philosopher Charles Sanders Peirce discussed the use of a latent structure model in measuring the success of prediction. However, the formal use of LC models in statistical analysis began in the 1950s, as Paul Lazarsfeld first applied LC models to clustering analysis. In 1974, Leo Goodman extended LC analysis to dealing with observed polytomous variables and multiple latent variables. Moreover, Goodman’s work on model estimation and identification greatly boosted the application of LC analysis in a number of different areas. In the 1990s, a general framework for categorical data analysis with discrete latent variables was proposed by Jacques Hagenaars. Ever since, LC models were extended to the settings that involve continuous covariates, ordinal variables, and longitudinal data.

LC Model

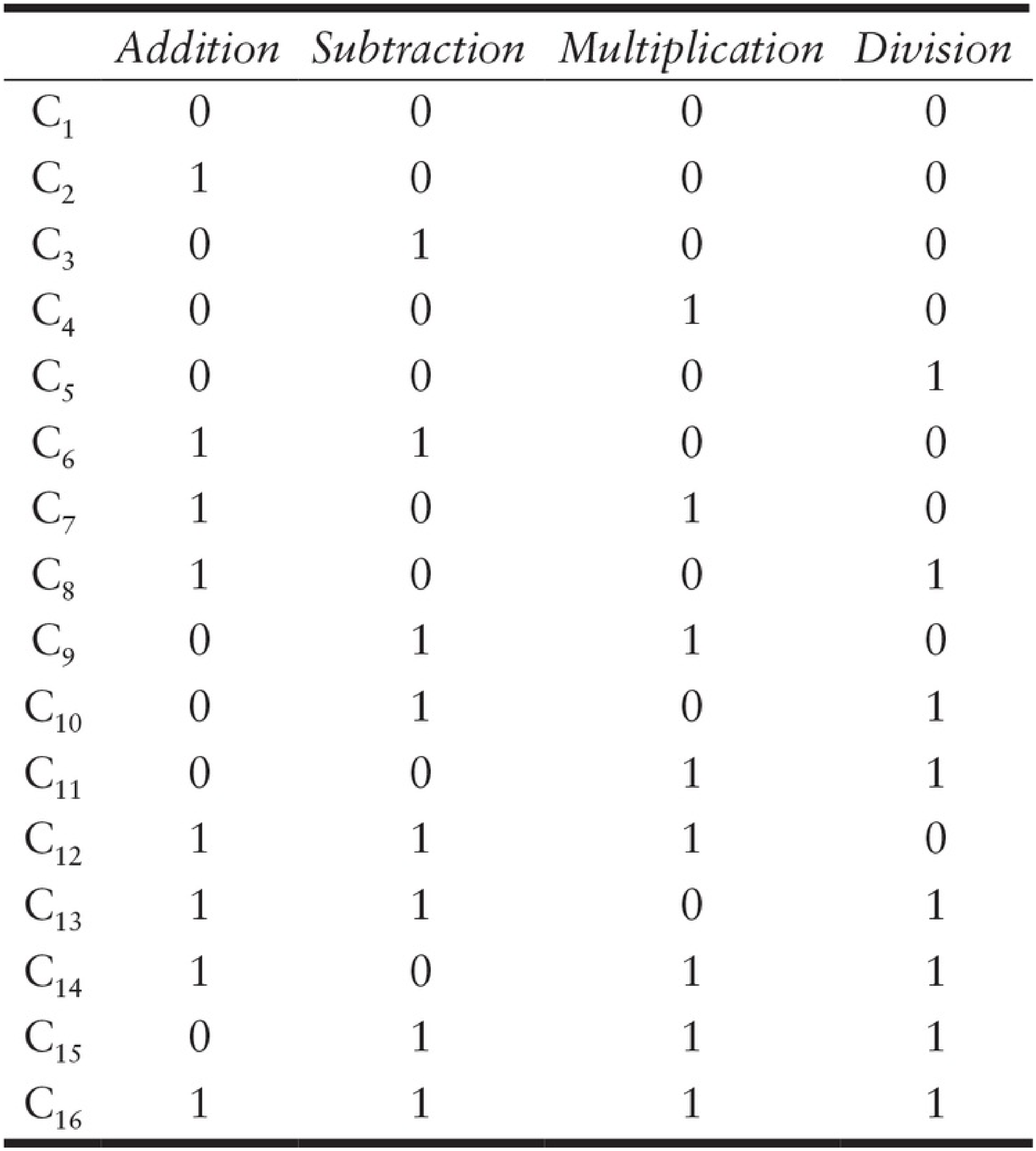
A basic LC model includes two types of categorical variables: observed categorical variables and latent categorical variables. This section introduces two approaches in parameterizing the LC model: probabilistic and log-linear. As an illustrative example, consider a hypothetical test designed to measure students’ mathematical ability in four areas (attributes): addition, subtraction, multiplication, and division. Students’ performance in each area is categorized as either “mastery” or “nonmastery.” The test consists of 20 multiple-choice items with each attribute being measured by 5 items. As such, in this case, there are 4 binary latent variables and 20 binary observed variables. This hypothetical test represents the application of LC models in diagnostic classification. The primary goal of diagnostic classification models is to classify respondents according to multiple latent characteristics representing the knowledge state of a respondent.

Probabilistic Parameterization

Let X_i represent the ith element of attribute (latent variable) vector X, and Y_l the lth elements of item (observed variable) vector Y, where $1 \le i \le N$, and $1 \le l \le L$. In this case, $N = 4$, $L = 20$. Further, let C_j be the jth class of the LC vector C (as shown in Table 1), where $1 \le j \le 2^N$, and y a complete response pattern. Then, the probability of obtaining the response pattern y, $p\ (Y = y)$, can be expressed as the sum of the weighted class-specific probabilities, $p\ (Y = y \mid X = C_j)$. Specifically,

$$p\left(Y = y\right) = \sum_{j=1}^{2^N} p(X = C_j)p\left(Y = y \mid X = C_j\right),$$

where $p\ (X = C_j)$ is the probability of the membership in LC C_j.



	Addition	*Subtraction*	*Multiplication*	*Division*
C_1	0	0	0	0
C_2	1	0	0	0
C_3	0	1	0	0
C_4	0	0	1	0
C_5	0	0	0	1
C_6	1	1	0	0
C_7	1	0	1	0
C_8	1	0	0	1
C_9	0	1	1	0
C_{10}	0	1	0	1
C_{11}	0	0	1	1
C_{12}	1	1	1	0
C_{13}	1	1	0	1
C_{14}	1	0	1	1
C_{15}	0	1	1	1
C_{16}	1	1	1	1

The LC models require conditional independency among observed variables (i.e., the assumption of local independence). That is, the *L*-observed variables are assumed to be mutually independent given the latent variables, which can be formulated as follows:

$$p\left(Y = y \mid X = C_j\right) = \prod_{l=1}^{L} p(Y_l = y_l \mid X = C_j),$$

where $p\ (Y_l = y_l \mid X = C_j)$ is the probability of a response on item l by a respondent from LC C_j. In clustering analysis, these probabilities can be used to name the classes. Combining Equations 1 and 2 gives the following model for $p\ (Y = y)$:

$$p\left(Y = y\right) = \sum_{j=1}^{2^N} p(X = C_j) \prod_{l=1}^{L} p(Y_l = y_l \mid X = C_j).$$

To classify respondents (i.e., assign individuals to LCs), one needs to know the probability that a respondent belongs to LC C_j (also referred to as posterior membership probability), $p\ (X = C_j | Y = y)$, which can be obtained by the Bayes's rule:

$$p(X = C_j \mid Y = y) = \frac{p\left(X = C_j\right) p\left(Y = y \mid X = C_j\right)}{p\left(Y = y\right)}.$$

According to the classification rule of modal assignment, respondents are assigned to the LC with the highest $p\ (X = C_j | Y = y)$.

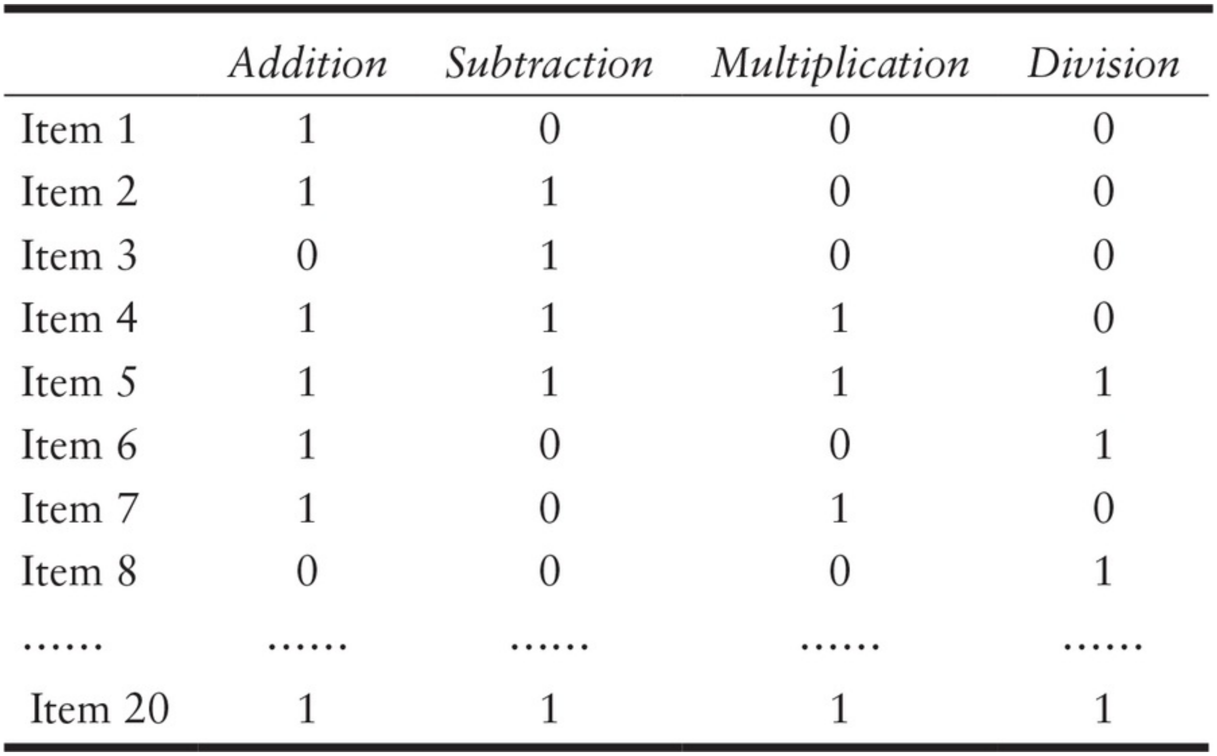
Log-Linear Parameterization

Loglinear model is a general framework to model the relationship among a set of discrete variables. The model focuses on the predicted frequency of respondent in each category. In a general form, the logarithmic transformation of the predicted frequencies, *ln* (*F*), is modeled as the linear combination of a set of discrete variables with the weights of the variables that capture how *ln* (*F*) changes across categories. This model can be expanded to the situations involving latent discrete variables by simply treating them as additional dimensions with missing frequencies.

In the log-linear parameterization, the probability of a response on item *l* by a respondent from LC C_j, $p\ (Y_l = y_l\ |X = C_j)$, can be expressed as follows:

$$p\left(Y_l = y_l \mid X = C_j\right) = \frac{\exp\left(\lambda_{l,0} + \lambda_l^T \boldsymbol{h}\left(C_j, \boldsymbol{q}_l\right)\right)}{1 + \exp\left(\lambda_{l,0} + \lambda_l^T \boldsymbol{h}\left(C_j, \boldsymbol{q}_l\right)\right)},$$

where $\boldsymbol{q_l}$ is the set of Q-matrix entries for item *l*. As shown in Table 2, Q-matrix is item-by-attribute matrix, in which the relationships between items and attributes are indexed with 1 (present) and 0 (absent). Also, λ_l,0 represents the logit of response pattern y_l when all Q-matrix indicated attributes equal zero. Moreover, the λ_l represents a vector of size $(2^{N-1}) \times 1$ with main effect and interaction effect parameters for item *l*, and $\boldsymbol{h}\ (\boldsymbol{C_j}, \boldsymbol{q_l})$ a vector of size $(2^{N-1}) \times 1$ with linear combinations of the $\boldsymbol{C_j}$ and $\boldsymbol{q_l}$. Note that, in Equation 5, the log-linear model takes a particular form with logit (log odds) as the link function.



	Addition	*Subtraction*	*Multiplication*	*Division*
Item 1	1	0	0	0
Item 2	1	1	0	0
Item 3	0	1	0	0
Item 4	1	1	1	0
Item 5	1	1	1	1
Item 6	1	0	0	1
Item 7	1	0	1	0
Item 8	0	0	0	1
……	……	……	……	……
Item 20	1	1	1	1

An effective way to parameterize the probability of the membership in LC C_j, $p\ (X = C_j)$, is to bring in the kernel expression μ_j, which can be transformed to $p\ (X = C_j)$ through the following formula:

$$p\left(X = C_j\right) = \frac{\exp\left(\mu_j\right)}{\sum_{j=1}^{2^N}(\mu_j)}.$$

Specifically, the kernel expression μ_j is the linear combination of a set of attributes with the parameters representing the main effects associated with each attribute as well as all possible interaction between attributes. In general, the kernel expression can be written as

$$\mu_j = \sum_{i=1}^{N} \gamma_{1,(a)} C_{ja} + \sum_{a'=a+1}^{N} \gamma_{2,(a,a')} C_{ja} C_{ja'}$$

$$+ \ldots + \gamma_{N,(a,a',\ldots)} \prod_{a=1}^{N} C_{ja},$$

where μ_j is the kernel expression used to determine the membership probability for LC C_j; $\gamma_{1,}(_a)$ is the main effect parameter associated with attribute a; $\gamma_{2,(a,a')}$ is the two-way interaction effect parameter associated with attributes a and a'; $\gamma_{N,}(_{a,a',\ldots})$ is the N-way interaction effect associated with all attributes; and C_{ja} is the nonzero attribute a in LC C_j. Combining Equations 5 and 6, the probability of obtaining the response pattern y, p $(Y = y)$, in the log-linear form, can be written as

$$p\left(Y = y\right) = \sum_{j=1}^{2^N}\left[\frac{\exp\left(\mu_j\right)}{\sum_{j=1}^{2^N}\left(\mu_j\right)}\right]$$

$$\prod_{l=1}^{L}\left[\frac{\exp\left(\lambda_{l,0} + \lambda_l^T h\left(C_j, q_l\right)\right)}{1 + \exp\left(\lambda_{l,0} + \lambda_l^T h\left(C_j, q_l\right)\right)}\right].$$

Model Estimation

The parameters of LC models are typically estimated using the maximum likelihood estimation (MLE) method. The goal of MLE is to find the parameter values that maximize the likelihood of the observations given the parameters. In estimation, the logarithm of likelihood function is maximized, which, for the illustrative example, can be written as

$$ln\left[p\left(Y = y\right)\right] = ln\left[\sum_{j=1}^{2^N} p(X = C_j)\right]\prod_{l=1}^{L} p(Y_l = y_l \mid X = C_j).$$

There are two main MLE-based iterative approaches to find the ML estimates: the expectation–maximization (EM) and the Newton–Raphson (NR) algorithms. Both algorithms begin with a set of “start values” and go through a series of estimation–reestimation iterations until a desired criterion (for convergence) is reached.

The EM algorithm consists of two basic steps: the expectation (E) and the maximization (M) steps. The desired estimates are found by iteratively processing these two steps. Specifically, in estimating LC models, the goal of the E step is to calculate the expected value of the log-likelihood function with respect to the latent variables, conditional on the observed data and the current estimates. Then, in the M step, the (expected value calculation) function is maximized to find the updated estimates for the parameters.

The NR method also takes an iterative approach to find the MLEs of the LC model parameters. The NR method begins with a set of initial parameter values (θ) and then modifies them by the product of the inverse of the Hessian matrix (*H*) and the gradient vector (*g*) of the log-likelihood function, . That is,

$$\theta_{i+1} = \theta_i - H_i^{-1} g_i.$$

Note that, when the initial parameter estimates (θ) are close to the ML estimates, the convergence of the NR method is quite fast.

There are several computer programs that can be used to estimate LC models. The first LC program is MLLSA, which was developed by Clifford Clogg in 1977. LEM by Jeroen K. Vermunt is a general command-based program for the analysis of categorical data. LEM can estimate a variety of MC models. Mplus is a command-based interface developed by Linda Muthén and Bengt Muthén for modeling and analyzing data with latent variables at any levels of measurement. Mplus can solve most types of LC models, expect for those with nominal indicators. Unlike command-based programs, Latent GOLD provides a graphical user interface and can deal with LC models with any type of indicators and covariate.

Model Evaluation

The goodness of fit of the LC models can be evaluated by four commonly used criteria: Pearson chi-square, likelihood-ratio chi-square (G^2), Akaike information criteria, and Bayesian information criteria (BIC). Chi-square and G^2 are regarded as *absolute* model fit statistics, which rely on the comparison between the model-implied cell frequency count (f_m) based on the (observed variable) contingency table and the observed cell frequency count (f_o). In contrast, AIC and BIC are regarded as *relative* model fit indices, indicating either of the statistics itself does not serve as a criterion for model fit but can suggest the desirable model in model comparison.

Specifically, Pearson chi-square is calculated by the formula as follows:

$$\chi^2 = \sum_{r=1}^{n} \frac{\left(f_{o.r} - f_{m.r}\right)}{f_{m.r}},$$

where n is the total number of cells, $f_{o.r}$ the observed frequency in the rth cell, $f_{m.r}$ the model-implied frequency in the rth cell. The degree of freedom (df) for Pearson chi-square statistic can be calculated as

$$df = (n-1) - \left[n_c \left(\sum_{l=1}^{L} n_l + L - 1 \right) - 1 \right],$$

where n_c is the total number of LC, n_l the number of category for the lth observed variable. In the illustrative example, $n = 2^{20}$, $n_c = 2^4$, and $n_l = 2$.

The likelihood ratio chi-square statistic is expressed as a function of the ratio of the observed to model-implied cell counts and can be written as

$$G^2 = 2\sum_{r=1}^{n} f_{o.r} ln(f_{o.r} / f_{m.r}).$$

Similar to chi-square, G^2 has an asymptotic chi-square distribution. Since both chi-square and G^2 are based on the entire response pattern, they are also called full-information goodness-of-fit statistics. Note that, as the chi-square–based statistics, reliable tests for chi-square and G^2 require sufficient sample size for each cell of the contingency table. In practice, when the number of observed variables is relatively large (e.g., a 100-item test), having a sufficient sample size for each cell becomes unrealistic. Therefore, the chi-square and G^2 statistics reported under such situation are problematic.

In model evaluation, AIC and BIC are also known as information criteria. They penalize the log-likelihood for the increase of parameters due to model complexity. To do so, a particular term associated with the number of parameters is reduced from the log-likelihood value. Specifically,

$$\text{AIC} = -2\ ln(L) - 2df$$

and

$$\text{BIC} = -2\ ln(L) - df \times [ln(N)],$$

where ln (L) is the log-likelihood value obtained from each analysis, N the sample size. Both AIC and BIC represent the trade-off between model fit and model parsimony. For both statistics, lower values are desirable.

Bo Hu

See also Classification; Exploratory Factor Analysis

Further Readings

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. American Journal of Sociology, 79, 1179–1259.

Hagenaars, J. A. (1990). Categorical longitudinal data—Loglinear analysis of panel, trend and cohort data. Newbury Park, CA: Sage.

Hagenaars, J. A., & McCutcheon, A. L. (2002). Applied latent class analysis. Cambridge, UK: Cambridge University Press.

Heinen, T. (1996). Latent class and discrete latent trait models: Similarities and differences. Thousand Oaks, CA: Sage.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer *et al.* (Eds.), Measurement and prediction (pp. 362–472). Princeton, NJ: Princeton University Press.

Rupp, A., Templin, J., & Henson, R. (2010). Diagnostic measurement: Theory, methods, and applications. New York, NY: Guilford Press.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. Educational Measurement: Issues and Practice, 32(2), 37–50.

John T. E. Richardson John T. E. Richardson Richardson, John T. E.

Latin Square Design

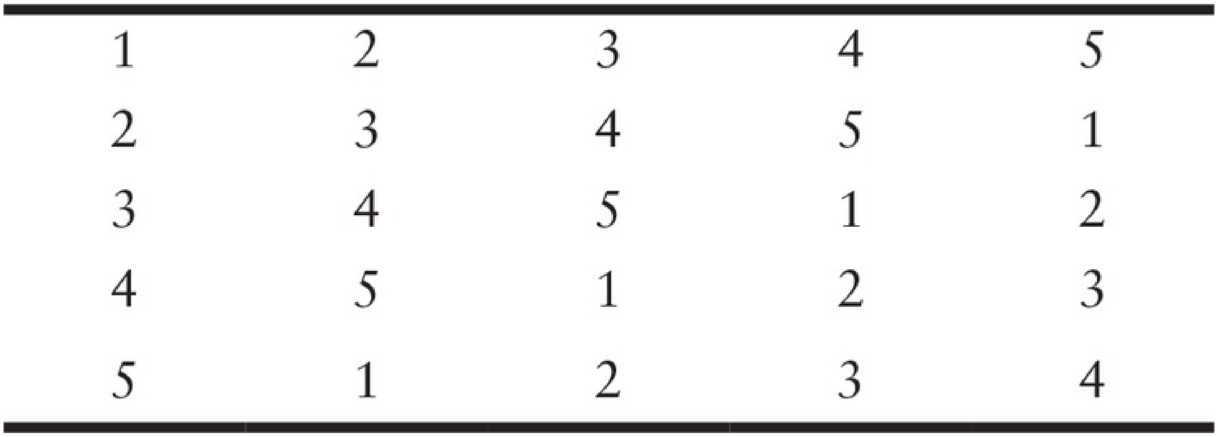
Latin square design

948

949

Latin Square Design

A Latin square is a grid or matrix containing the same number of rows and columns (*k*, say). The cell entries consist of a sequence of *k* symbols (for instance, the integers from 1 to *k*) inserted in such a way that each symbol occurs only once in each row and once in each column of the grid. By way of an example, Table 1 shows a Latin square that contains the numbers from 1 to 5. This entry describes the classification of Latin squares, their origins in agricultural experiments, and their applications in the social and behavioral sciences.



1	2	3	4	5
2	3	4	5	1
3	4	5	1	2
4	5	1	2	3
5	1	2	3	4

Table 1 is an example of a *standard form*, in that the numbers in the first row and the numbers in the first column are in their natural order. Other, nonstandard Latin squares can be constructed by interchanging different rows in the table, by interchanging different columns in the table, or both. In research practice, it is typically recommended that a standard form of the relevant size should be drawn at random from published tables of Latin squares and that its rows and its columns should be interchanged at random using published tables of random sequences.

In 1925, Ronald A. Fisher, a British statistician, proposed that Latin squares could be used to arrange plots in agricultural experiments so as to control for differences in soil fertility. For instance, Table 1 might represent the arrangement of 25 plots available for an experiment, and five different treatments would be applied corresponding to the numbers in the cells in the table. An analysis of variance carried out on some criterion variable would then identify the variation among the rows and the variation among the columns, leaving an unbiased estimate of the effect of the treatments, controlling for any differences among the rows and columns.

Fisher's writings proved very influential in the social and behavioral sciences, especially in the years after World War II. Benjamin J. Winer, Donald R. Brown, and Kenneth M. Michels identified four main applications of Latin squares in such research: to control nuisance variables, to counterbalance order effects in repeated measures designs, to confound treatment conditions with group main effects, and as balanced fractional replications from a complete factorial design. The second of these applications is probably the most common, although researchers may well not take this feature into account in their analysis of their results. In some fields, the use of Latin square designs has declined, but they remain a potentially important experimental technique.

John T. E. Richardson

See also Analysis of Variance; Repeated Measures Analysis of Variance; Repeated Measures Designs

Further Readings

Grant, D. A. (1948). The Latin square principle in the design and analysis of psychological experiments. Psychological Bulletin, 45, 427–442. doi:10.1037/h0053912

Hamlin, R. P. (2005). The rise and fall of the Latin Square in marketing: A cautionary tale. European Journal of Marketing, 39, 328–350. doi:10.1108/03090560510581809

Reese, H. W. (1997). Counterbalancing and other uses of repeated-measures Latin-square designs: Analyses and interpretations. Journal of Experimental Child Psychology, 64, 137–158. doi:10.1006/jecp.1996.2333

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical principles in experimental design (3rd ed.). New York, NY: McGraw-Hill.

Socorro Herrera Socorro Herrera Herrera, Socorro

Latinos and Testing

Latinos and testing

949

952

Latinos and Testing

Assessment refers to a broad range of procedures used by educators to collect information, from which inferences are drawn about a student's knowledge, understanding, skills, or abilities. Assessment may be undertaken to improve student learning, to judge student performance, or to evaluate the effectiveness of an educational system. Ideally, the information or data derived from assessment is used to promote equity in learning, assessment, and educational opportunity.

This entry focuses on the impact of assessment, particularly high-stakes testing, on Latino students. It starts with an overview of the demographic characteristics of Latinos, followed by an explanation of assessment and testing. Next, the implications of the No Child Left Behind Act (NCLB) of 2001 are discussed, highlighting how its high-stakes testing legacy has impacted Latino youth.

Demographic Profile of Latinos

Latinos form the largest ethnic/racial minority group in the United States with a recorded 54 million residents, accounting for 17% of the nation's population. States that each have over 1 million Latino residents are Arizona, Illinois, New Jersey, and New York, with particularly large Latino populations in Florida, California, and Texas.

Spanish is the most common first language or heritage language among English language learners (ELLs); as of 2013, Spanish was spoken at home among 71% of ELLs. In early education programs, Latino children lag behind in preschool enrollment, compared to their African American and White peers. This places Latino children in a precarious situation for school readiness and ultimately for academic success and employment opportunity later in life. Latinos constitute a significant population in elementary and high schools. In 2014, they made up 25% of students enrolled in public elementary and secondary schools. Increase in the Latino student population is anticipated; by 2036, they are projected to constitute one-third of the nation’s children aged 3 through 17 years.

Poverty is unequivocally a detriment to the educational advancement of Latino youth. Almost one-third of Latino children live in poverty. Furthermore, in school year 2014–2015, nearly half of Latinos attended high-poverty schools. High-poverty schools are typically underfunded and overcrowded and have fewer trained teachers and services for ELLs. Researchers have found high levels of segregation by race, poverty, and ELL status in the schools that many Latinos attend, referring to this as “triple segregation.”

High dropout rates and low college attendance are issues that have plagued Latinos, although in recent years they have made gains in both areas. The Latino high school dropout rate decreased from 32% in 2000 to 10% in 2015. Although this is marked progress, the rate is still much higher than the overall dropout rate of 6%. Latinos have also made notable advancements in two-and four-year college enrollment. Between 2000 and 2015, Latino enrollment in degree-granting postsecondary institutions more than doubled, increasing from 1.4 million to 3 million students, while enrollment for other racial/ethnic groups fluctuated during this time. Latinos made up 18% of students in postsecondary institutions by fall 2015. In spite of increased enrollment, Latinos still trail behind their White, Asian, and African American peers in graduating with a four-year degree.

Despite the recent educational gains Latinos have made in postsecondary education, there are still significant systemic inequities affecting them in PreK–12 settings. Most notable are the tests that act as a gatekeeper to academic progression, graduation, admission, and ultimately employment among Latino youth.

Assessment Versus Testing

Assessments are conducted for formative or summative purposes. The purpose of formative assessment is to observe, evaluate, and provide ongoing feedback on student learning, so that instruction can be modified accordingly to enhance learning. Formative assessment is gradual, occurs over time to extend student learning, and is generally low-stakes because emphasis is placed on mastery of content rather than on earning a grade. Examples of informal formative assessments are observations, daily work, conversations with students, and graphic organizers, whereas more formal formative assessments include quizzes, portfolios, and performance assessments.

In contrast to formative assessment, the purpose of summative assessment is to evaluate and measure student learning at a given point of time, usually after a period of instruction, such as a lesson, theme, unit, semester, or school year. The outcome of summative assessment is a measurement such as a score or grade, which is typically referenced against specific learning outcomes, standards, or benchmarks. Because summative assessments are often used as indicators of school, teacher, or student accountability, they carry significant value. Examples of postinstructional classroom summative assessments are midterm and final exams, as well as papers. Standardized tests are a type of summative assessment, often associated with high-stakes testing. Simply put, formative assessment is assessment *for* learning, whereas summative assessment is assessment *of* learning.

Although assessment is frequently associated with testing, they are not synonymous. Testing is a specific, evaluative procedure used to sample, measure, and judge a student’s performance. The outcome of a test is a score or grade, which is intended to reflect the student’s level of competency or achievement. Test scores or grades are typically used to make important decisions regarding a student’s educational path, such as retention, grade promotion, entry into programs (e.g., honors, Advanced Placement [AP], or gifted programs), access to special services (e.g., special education or remedial programs), and graduation.

Testing Implications of No Child Left Behind

The U.S. elementary and secondary school system is characterized by extensive high-stakes testing, much of which is federally mandated. NCLB had a defining impact on U.S. education. Although the law was replaced in 2015 by the Every Student Succeeds Act (ESSA), ESSA still mandates annual testing as an accountability measure. There is an expectation, however, that the law will lead to a reduction in other standardized testing. Although NCLB is no longer in place, it wrote the parameters and set the agenda for how assessment was defined.

NCLB emerged out of growing economic concerns that the United States was not producing an internationally competitive workforce. This was partly attributed to deficits in the U.S. educational system. Through federal measures and a top-down approach, NCLB sought to reform K–12 education by holding schools accountable for the academic progress and performance of their students. Furthermore, NCLB endeavored to improve the educational standing of disadvantaged students or *subgroups*—such as minorities, ELLs, those in special education programs, and children from low-income homes—by addressing the achievement gap.

Although NCLB neither mandated a national set of standards nor a testing scheme, it did mandate states to develop and adopt standards and to administer standardized tests in math and reading from Grades 3 through 8 and in high school. These test scores had to be reported for the entire student population and by subgroups. Standardized test scores in conjunction with other measures—such as attendance rate, graduation rate, and participation rate on state assessment—constituted the basis for adequate yearly progress, a metric used to assess the performance of schools and school districts. In sum, NCLB was a test-driven accountability measure, characterized by high-stakes testing based on standards, which was used at state, district, and school levels with the intention of monitoring and improving student performance.

The requirement that states and school districts report on the test performance of subgroups gave visibility to disadvantaged populations. However, a significant negative consequence of NCLB was that the pursuit of accountability led to an overemphasis on testing and to schools orienting their instruction to the test. Teaching to the test was fueled in large part by pressures to meet adequate yearly progress. Furthermore, the emphasis on high-stakes testing inadvertently eroded the emotional well-being of many students, especially because deficit discourse—that is, hyperfocus on what the student couldn’t do—was normative.

The Effect of Testing on Latinos

High-stakes testing disadvantages Latino students in the following ways. First, because most standardized tests are normed for native English populations, when these tests are administered to Latino ELLs, they inadvertently test English proficiency as opposed to strictly content. Furthermore, ELLs and, by extension, many Latinos are typically subjected to more testing than native English speakers. For low-performing students, this increases the likelihood of retention or not being promoted; tracking or placement into less competitive classes; (mis)- placement into special education or remedial programs; denial of access to extracurricular activities and programs; and dropping out or not graduating from high school.

Whether by default or by design, the data procured through NCLB high-stakes testing were used punitively to make crucial decisions about students' educational path, which would later impact their chances at educational, occupational, and socioeconomic advancement. Furthermore, although NCLB did not require students to pass an exit exam to graduate from high school, the number of states with this requirement increased after NCLB was adopted, adding to the hurdles in the way of Latinos' opportunities for postsecondary education. Exit exams have come under increasing criticism in recent years and many states have dropped them, but as of early 2017, at least 12 states still required students to pass a test to graduate from high school.

The passage of ESSA has begun an era of redefining assessment. Although high-stakes standardized testing remains, there is room for reform because ESSA endorses flexibility in how and when annual assessments are administered. Furthermore, ESSA promotes assessments that test what students are actually learning. This change in legislation can be leveraged to the advantage of all students but particularly for those in subgroups. The flexibility envisaged in ESSA has the potential to lead to improvements in the operationalization of formative and summative assessment.

Socorro Herrera

See also African Americans and Testing; Asian Americans and Testing; Minority Issues in Testing

Further Readings

Bartlett, L., & García, O. (2011). Additive schooling in subtractive times Dominican immigrant youth in the heights. Nashville, TN: Vanderbilt University Press.

Contreras, F. (2011). Achieving equity for Latino students: Expanding the pathway to higher education through public policy. New York, NY: Teachers College Press.

Gathercole, V. C. M. (Ed.). (2013). Issues in the assessment of bilinguals. Buffalo, NY: Multilingual Matters.

Leal, D. L., & Meier, K. J. (Eds.). (2011). The politics of Latino education. New York, NY: Teachers College Press.

Menken, K. (2008). English learners left behind: Standardized testing as language policy. Buffalo, NY: Multilingual Matters.

Noguera, P. A. (2008). The trouble with black boys: And other reflections on race, equity, and the future of public education. San Francisco, CA: Jossey-Bass.

Pandya, J. Z. (2011). Overtested: How high-stakes accountability fails English language learners. New York, NY: Teachers College Press.

Valencia, R. R. (Ed.). (2004). Chicano school failure and success: Past, present, and future. New York, NY: Routledge.

Valenzuela, A. (1999). Subtractive schooling: US-Mexican youth and the politics of caring. Albany State University of New York Press.

Ser Hong Tan Ser Hong Tan Tan, Ser Hong

Gregory Arief D. Liem Gregory Arief D. Liem Liem, Gregory Arief D.

Learned Helplessness Learned helplessness

952

953

Learned Helplessness

Individuals with learned helplessness are characterized by their learned inclinations to see that their responses to escape unpleasant situations have no bearings on the outcome. This leads the individuals to hold the expectation that they have no control over the occurrence of the negative stimuli or outcome. Consequently, these individuals adopt a passive, pervasive self-defeating attitude as they fail to unlearn their preconceived mind-sets and relearn new ways to overcome other aversive situations.

The reformulated learned helplessness theory distinguishes between universal versus personal helplessness, global versus specific helplessness, and chronic versus transient helplessness. External attributions are made in universal helplessness, and individuals believe that no one has control over the outcome. In contrast, internal attributions are made in personal helplessness, and individuals believe that other people have control over the outcome even though they themselves do not. Global helplessness happens when individuals extrapolate their passive, helplessness attitude to divergent situations, whereas specific helplessness limits the passive attitude to situations that resemble the original situation. On the other hand, chronic helplessness is continuous, while transient helplessness is temporal in time.

Learned helplessness is applied to explain continued poor performance in students who had experienced failure in school achievement tasks. Accordingly, learned helplessness is manifested in a dysfunctional cognition–behavior–affect system where the components feed into each other.

Fundamentally, students lack motivation as they perceive that they have no control and their actions do not translate into results. Failures may or may not stem from students' actions or the lack thereof; students do not see themselves as being responsible for their failures. The perceived dissociation between students' actions and outcomes leads students to believe that their failures cannot be overcome. This, however, may not reflect students' objective ability. Students attribute their failures to fixed or uncontrollable reasons such as poor ability, rather than malleable and controllable reasons such as inadequate effort. These attributions constitute universal and chronic helplessness. Furthermore, students ruminate about the reasons behind their failures, leading to strong negative and depressive affect.

As further attempts are deemed to be futile, students who display learned helplessness give up easily and do not persist to find solutions to their problems. Such behavioral withdrawal leads to repeated failure outcomes that result in poor psychological adjustment such as shame, resignation, and despair as students condemn their own competencies.

Interventions such as attribution retraining and positive self-regulation skills can potentially help students who suffer from learned helplessness. As it is inevitable for students to experience failures in their learning journey, efforts should be invested to refrain students from slipping into learned helplessness or to provide opportunities for students to experience mastery and success on achievement tasks. Additional support is also needed to help students who display learned helplessness break the cycle of the dysfunctional cognition–behavior–affect system and pull them out of their rut.

Ser Hong Tan and Gregory Arief D. Liem

See also Attributional Theory; Locus of Control; Motivation; Resilience; Self-Regulation

Further Readings

Abramson, L. Y., Seligman, M. E. P., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. Journal of Abnormal Psychology, 87, 49–74. doi:10.1037/0021-843X.87.1.49

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. Journal of Personality and Social Psychology, 31, 674–685. doi:10.1037/h0077149

Peterson, C., Maier, S. F., & Seligman, M. E. P. (1993). Learned helplessness: A theory for the age of personal control. New York, NY: Oxford University Press.

Jessica Namkung Jessica Namkung Namkung, Jessica

Peng Peng Peng Peng Peng, Peng

Learning Disabilities

Learning disabilities

953

956

Learning Disabilities

Under the Individuals with Disabilities Education Act, (IDEA), the term *specific learning disability* is defined as a specific disorder in one or more areas of psychological processes involved in understanding and using spoken or written language, which results in deficits in the ability to listen, think, speak, read, write, spell, or do mathematics. Specific learning disability, also referred to as *specific learning disorder*, does not include learning problems that are attributable to sensory disorders, emotional disturbance, intellectual disabilities, or cultural or economic disadvantages. Today, learning disabilities account for as much as 50% of all students receiving special education. This entry first discusses the causes of and identification of learning disabilities. It then describes the characteristics of students with learning disabilities and types of learning disabilities. Finally, it looks at instructional strategies that are thought to be effective with students with learning disabilities.

Although a single cause of a learning disability is not known, the possible causes include physiological factors (e.g., heredity, brain injury, and biochemical imbalance) and environmental factors (e.g., poor nutrition and exposure to environmental toxins, such as lead). Students with learning disabilities are a heterogeneous group, meaning that they may have problems in reading, mathematics, written language, or oral language.

Identification of Learning Disabilities

IQ-Achievement Discrepancy

Traditionally, learning disabilities are identified based on the discrepancies between a composite measure of IQ and academic achievement, such as mathematics achievement. That is, if students show at least a two standard deviation difference between their intelligence, indexed by an IQ test, and academic ability, indexed by an academic achievement test, the students are identified as having a learning disability. Thus, a student who exhibits unexpected learning difficulties as indicated by academic achievement far below what would be expected by the IQ score (e.g., average IQ score of 100 and below-average reading achievement score of 70) would be identified as having a learning disability. On the other hand, a student with a below-average IQ score of 85 and a below-average mathematics achievement score of 80 would not be identified as having a learning disability because there is not a large enough discrepancy.

The IQ-achievement discrepancy model has been often criticized for being a wait-to-fail model. That is, students are not identified as having learning disabilities until there are substantial differences between their IQ and achievement scores, thereby delaying early intervention opportunities. This model also does not assess or inform the quality of instruction received by students, thereby not allowing discrimination between those who are low achievers as a result of poor instruction and those with true learning disabilities.

Response to Intervention

As an alternative to the IQ-achievement discrepancy model, response to intervention is a three-tier approach to providing high-quality, research validation instruction with ongoing progress monitoring and data-based decision making. In general, all students are given a universal screening measure at the beginning of the year to identify those who are at risk. In Tier 1, teachers provide evidence-based instruction in the general education, whole-class setting and monitor student progress on a weekly basis. Students who do not make adequate progress (generally around 20–30% of students) move to the more intensive level. In Tier 2, students receive additional support, such as targeted interventions, in a small-group setting (three to five students) from either the general classroom teacher or other educational personnel, such as a reading specialist, special education teacher, or tutor. Student progress is monitored continuously throughout. Those who make adequate progress may return to Tier 1 or continue to receive Tier 2 instruction, and those who still do not make sufficient progress in Tier 2 move to the most intensive level (5–10% of students).

In the most intensive level, Tier 3, students receive even more differentiated and individualized support, preferably in a one-on-one setting. At this stage, students may be identified with a learning disability and qualify for special education services. In this way, response to intervention is based on a preventive framework by providing increasingly more intensive instruction to help struggling learners early on before they fall too far behind their peers and before they are identified as having learning disabilities.

Characteristics of Students With Learning Disabilities

Although students with learning disabilities have average to above-average intelligence, they show deficits in cognitive abilities, such as attention, working memory, long-term memory, and processing speed, compared to their typically developing peers. For example, students with learning disabilities have significant difficulty attending to classroom instruction and ignoring other stimuli (e.g., auditory and visual) in the classroom. In fact, a significant number of students with learning disabilities have co-occurring attention-deficit/hyperactivity disorder.

Students with learning disabilities show weaknesses in memory. In particular, students with learning disabilities have poor working memory, which refers to the ability to simultaneously process and store information to support ongoing cognitive tasks (e.g., keeping track of the contents of the text just read while processing new text when reading a story), and long-term memory, which refers to the permanent storage of information in the brain (e.g., automatically recalling 5 + 2 = 7 from memory without having to count up).

Students with learning disabilities also show weakness in information processing. They tend to be slower at processing words or numbers and executing multiple steps compared to typically developing peers. They also have difficulty with metacognition, lacking the ability to relate what they have learned to new information they are learning. Besides the cognitive difficulties, students with learning disabilities tend to have lower self-esteem, likely to be due to repeated academic failure they experience, poor social skills, and lower motivation compared to their typical peers.

Types of Learning Disabilities

Reading Disabilities

It is estimated that reading disabilities is the most prevalent type of learning disabilities. As many as 90% of school-aged children with learning disabilities have reading disabilities, and even the low estimates are approximately 60%. The most common subtypes are dyslexia, specific comprehension problems, or a combination of dyslexia and specific comprehension deficits.

Students with dyslexia are characterized by having difficulties in recognizing words accurately and fluently but having language comprehension abilities appropriate for their age level. Those with dyslexia have deficits in phonological awareness, the ability to understand that speech flow can be broken into smaller sound units, such as words, syllables, and phones. These deficits, in turn, lead to word recognition problems. On the other hand, students with specific comprehension deficits have no problems in recognizing words accurately and fluently but have difficulties in language comprehension skills, such as vocabulary and listening comprehension. These difficulties may stem from broad language difficulties that are present before developing reading skills, such as weak vocabulary knowledge, difficulty processing grammar, and poor oral language comprehension. The combination of dyslexia and specific comprehension deficits is characterized by difficulty in word recognition skills as well as language comprehension skills. It is estimated that 5–17% of the population has dyslexia, and 10–15% of primary school-aged students have specific reading comprehension deficits.

Mathematics Disabilities

Students with mathematics learning disabilities constitute approximately 5–8% of the school-aged population. It is estimated that approximately 40% of students with learning disabilities have mathematics disabilities. The two most common deficit students with mathematics learning disabilities experience are computations and problem solving.

Computations include number combinations (e.g., 3 + 4 = 7) and procedural

Computations include number combinations (e.g., 3 + 4 = 7) and procedural computations (e.g., 16 + 28 = 44), in which students find the answer to already setup problems. Students with mathematics learning disabilities are slower at counting to figure out the answers, rely on inefficient counting strategies, such as finger counting and counting all (counting 1, 2, 3, 4, 5, 6, 7 for 3 + 4 = 7), and have difficulty automatically recalling number combinations from long-term memory. By contrast, problem solving refers to the thinking required to solve word problems, which require students to understand the problem narrative, identify missing information, construct a number sentence, and solve for the missing information to find the answer. Because of the combination of language requirements and complex processes involved in problem solving, problem solving is the most difficult area of mathematics for many with learning disabilities.

Written Expression Disabilities

Students with written expression disabilities have a particular difficulty communicating through writing. Although the prevalence of written expression disabilities has not been well studied and varies greatly across studies (1–20%), it is estimated that at least 10% of school-aged children may be affected given the high rates of reading disabilities (10–15%) and language disorders (8–15%). Those with written expression disabilities have deficits in many related abilities required in writing, such as grammar and punctuation. Common deficits are in handwriting, spelling, and composition. Students with written expression disabilities often have poor handwriting; handwriting requires the knowledge of orthography and planning ability as well as motor skills. Students with written expression disabilities also experience difficulty with spelling, which requires integration of visual processing, phonological or orthographic representations, and motor skills. Difficulties with composition involve problems with organization, coherence, clarity, and revision processes.

Instructional Strategies

Explicit Instruction

Explicit instruction is a systematic instructional approach, in which the teacher directly shares the information students need to learn using concise and specific language, but is also didactic in that there is a high level of teacher and student interaction. Explicit instruction involves a focused lesson, in which complex skills are broken down into smaller, targeted parts, a clear explanation of the targeted skill, teacher-led step-by-step demonstrations (I do), guided practice (we do) with corrective feedback, and independent practice (you do) with immediate feedback.

Strategy Instruction

Strategy instruction focuses on teaching rules and techniques that guide students to learn new skills, complete tasks independently, recall the skills later, and apply/generalize the skills in new settings and situations. This is particularly effective for students with learning disabilities because they often do not develop efficient learning strategies on their own. Strategy instruction includes instruction for more broad domains, such as teaching study skills (e.g., note taking, summarizing, and self-questioning) and mnemonics to assist with remembering and recalling a specific strategy, and academic content-specific strategies, such as using self-regulated strategy development to improve writing.

Scaffolding Instruction

Scaffolding refers to the process through which a teacher adds support for students in mastering tasks. The teacher systematically builds on students' experience and prior knowledge when teaching a new skill, then removes the support gradually as students become more proficient. Content scaffolding refers to choosing content that is not too difficult or unfamiliar for students learning a new skill, whereas material scaffolding refers to the use of written prompts or cues to guide students to perform a task independently. In task scaffolding, a teacher specifies steps in a task and models the steps by verbalizing thought processes, followed by student practice. During task scaffolding, the teacher gradually releases the responsibility of completing the task to the students.

Jessica Namkung and Peng Peng

See also Ability-Achievement Discrepancy; Response to Intervention; Special Education Identification; Special Education Law

Further Readings

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2006). Learning disabilities: From identification to intervention. New York, NY: Guilford Press.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? Reading Research Quarterly, 41, 93–99.

Geary, D. C. (2004). Mathematics and learning disabilities. Journal of Learning Disabilities, 37, 4–15.

Graham, S., Harris, K. R., MacArthur, C. A., & Schwartz, S. (1991). Writing and writing instruction for students with learning disabilities: Review of a research program. Learning Disability Quarterly, 14, 89–114.

National Reading Panel. (2000). Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implication for reading instruction. National Institute of Child Health and Human Development, National Institutes of Health.

Swanson, H. L. (1999). Instructional components that predict treatment outcomes for students with learning disabilities: Support for a combined strategy and direct instruction model. Learning Disabilities Research & Practice, 14, 129–140.

Michael Tang Michael Tang Tang, Michael

Arunprakash T. Karunanithi Arunprakash T. Karunanithi Karunanithi, Arunprakash T.

Learning Maps

Learning maps

956

957

Learning Maps

Learning maps, a subset of concept maps, are organizational graphs to represent and present information in terms of relationships between the different parts of a map constructed as a gestalt system where the whole is greater than the sum of its parts. Concept maps and learning maps are visual abstractions of mental “places” and their connections to each other to give users a better grasp of knowledge landscapes. Such cognitive maps can be divided into two intersection types: maps to organize and analyze information and maps as education tools, with the latter further subdivided into maps for learning and maps for assessment. Learning maps therefore are spatial instructions to facilitate more self-modifying learning through dynamic adaptive processes with multiple alternate nodes and pathways to realize specific goals. This entry further discusses how learning maps are used for learning and assessment and the relationship of learning maps to systems thinking.

In education, learning maps as a learning aid can be used to lay out a study or lesson plans, to help with problem solving, and to master concepts and content. For example, these maps can be used for constructing main elements of a problem or problems with the steps needed for arriving at a solution and for making summaries of content and of key concepts contained in textbooks. Unlike traditional outlines that order content, concepts, and procedures as linear lists with little association with the listed items other than perhaps temporal relations, learning maps organize information as dynamic structures in which different content, concepts, and procedural nodes are interconnected by logical, different content, concepts, and procedural nodes are interconnected by logical, heuristic, associative, hierarchical, and other relationships.

It has been argued that learning maps teach a different way of acquiring knowledge that can be correlated with systems thinking. This open-ended and holistic cognitive mode is important because students and teachers cannot rely exclusively on thinking processes that are linear, fragmented, and reductionist to deal with today’s complex problems. Learning maps can be a tool to promote learning as a process that allows for taking different pathways toward different ends.

By regarding the overall process of learning as a systems map, learning maps have the ability to organize information according to structures as changing holistic parts that serve as interacting nodes as parts of a dynamic system that can contract or expand depending on learning objectives. Learning maps complement and reinforce systems thinking in that these maps are graphic devices that represent system characteristics such as open and infinite expandability, interconnected dynamic structures, and multiple parallel flows to lay out different paths leading the student to higher orders of thinking such as evaluating, hypothesis making, and thinking in highly abstract patterns.

Although there have been numerous studies on concept maps as assessment instruments, few studies have been specifically dedicated to learning maps as such. Nevertheless, what research exists on learning maps as a tool to measure academic achievement is not without significance. For example, a group of researchers at the University of Kansas has been researching the use of Dynamic Learning Maps (DLM) to assess the cognitive ability of students with significant cognitive challenges, with results that may be extended to measure cognitive ability, in general. The DLM alternate assessment researchers constructed learning maps as assessment maps because traditional multiple-choice testing does not allow students to fully demonstrate their academic knowledge or knowledge potential.

DLM’s use of learning maps has parallels to Joseph Novak’s development of concept mapping, an analytical and learning technique. Novak and his team at Cornell University originated the idea of concept maps as learning tools from their need to record the academic progress of science, technology, engineering, and mathematics students in the 1970s. To measure their students’ academic progress, Novak and his research team constructed visual maps, based on the constructivist theory of active learning, which became concept maps used primarily for teaching and analysis. Since then, concept maps have been used as primarily for teaching and analysis. Since then, concept maps have been used as a means to increase meaningful learning in education, government, and business settings.

Although Novak's concept maps grew out of assessment needs and are essentially analytical and learning tools, DLM's learning maps system is an assessment tool that can be used as a learning guide with concept maps embedded in the system's structure.

DLM maps are representations of how academic skills are acquired as reflected in education research literature. Nodes in the maps represent specific knowledge areas, skills, and understanding in specific disciplines along with the basic knowledges needed as prerequisites or understanding in the specific academic areas. The system is unique in that students and teachers can choose different pathways through different nodes to define how they, the students, think they can be best assessed for optimal results. Moreover, because the multiple and interconnected assessment rubrics represent achievement stages, the same rubrics can be used concurrently as hierarchical objectives and goals, a learning guide, and for formative evaluation of student progress. Learning maps therefore can be used both as assessment and teaching tools that organize the process of knowledge acquisition into changing and alternate pathways.

Michael Tang and Arunprakash T. Karunanithi

See also Alternate Assessments; Cognitive Neuroscience; Concept Mapping; Constructivist Approach; Curriculum Mapping; Information Processing Theory; Learning Theories

Further Readings

Luhmann, N. (2012). Introduction to systems theory (P. Gilgen, Trans.). Cambridge, UK: Polity Press.

Novak, J. D. (2010). Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. Oxford, UK: Routledge.

Okada, A., Shum, S. B., & Sherborne, T. (Eds.). (2014). Knowledge cartography: Software tools and mapping techniques. London, UK: Springer-

Verlag.

Websites

Dynamic Learning Maps: http://dynamiclearningmaps.org/

Nancy Butler Songer Nancy Butler Songer Songer, Nancy Butler

Learning Progressions

Learning progressions

958

959

Learning Progressions

A learning progression is a clearly articulated sequence of knowledge in a particular domain that communicates the vertical development of target knowledge over an extended period of time. Some aspects of learning progressions can be traced back to work in the 1960s by Jerome Bruner, who recognized that strong domain knowledge builds on previous strong domain knowledge and that deep conceptual knowledge development requires time, guidance, and multiple exposures. A well-defined learning progression can be useful in several ways. These include (a) providing a big picture view of intended knowledge, (b) serving as a resource for organized curriculum planning, and (c) serving as a reference point for gathering and using evidence for formative or summative assessment purposes. This entry provides basic information on the importance, characteristics, and purposes for the use of learning progressions, with emphasis on assessment-related uses.

Why Learning Progressions Are Important

Research informs us that the development of new knowledge builds on and from prior knowledge. However, many teaching and learning resources do not reflect this understanding. Many textbooks present content in an illogical sequence that is neither coherent nor organized within a grade or across multiple grades. Most content standards emphasize lists of learning targets with little attention to prioritization or an organized sequence. Most state and national assessments do not provide clear evidence of how learning is expected to develop over time or topic. As a result, most educators do not have resources that allow for systematic and sequential planning or assessment over time and topic.

Learning progressions are important as a template for the development and use of curriculum materials, instructional practices, and assessment. Using learning progressions in a coordinated manner shifts thinking about the nature of learning. With a progression approach, learning shifts from the process of mastering a body of disconnected facts and vocabulary to the organized building and revisiting of concepts over time. The role of assessment also shifts to a process of gathering evidence of progress along a continuum of increasing sophistication.

Characteristics of Learning Progressions

All learning progressions share a few common features. These include a series of levels from less sophisticated knowledge or skills (called the lower anchor) to more sophisticated knowledge or skills (called the upper anchor). Learning progressions vary in these dimensions: (a) the amount of time or material in the span, (b) the level of detail or granularity of knowledge at each level, (c) the nature of the knowledge itself (e.g., declarative facts, skills or ways of knowing, or a hybrid that combines both), and (d) whether the knowledge is representative of actual knowledge or aspirational knowledge of the target audience.

Most researchers describe the development of strong learning progressions as a difficult process that requires prioritization. In other words, learning progressions that contain fewer big ideas will support increased coherence in curriculum and instruction, opportunities for deeper learning and revisiting of concepts by students, and more valid assessment instruments. Learning progressions are used in many domains including mathematics, history, language arts, science, and communication.

Purposes of Learning Progressions

Learning progressions are valuable for curriculum development, instruction, and assessment. Using a learning progression as a template for textbook or curriculum development can support a systematic presentation of concepts that emphasizes explicit connections between concepts and ideas. Classroom instruction benefits when teachers are able to use evidence about students’ current knowledge for tailored feedback or to make sound decisions about the next instructional step.

Amelia Wenk Gotwals and Nancy Butler Songer have described both the challenges and benefits of designing a suite of assessment tasks that provide evidence of student knowledge at multiple points along a learning progression. Challenges include the difficulty of designing and providing validity evidence of a suite of tasks that can be used to pinpoint where students' learning lies on the progression and to identify common errors of misconceptions. Benefits include the ability of teachers to obtain first-hand evidence of common misconceptions or pivotal points in an instructional sequence that might require greater instructional emphasis. Called gatekeeper concepts, these points along a progression signify high-priority knowledge that students should provide evidence of mastery prior to moving on to the next area of knowledge development. A strong learning progression will identify a small number of gatekeeper concepts that can be used as points for evidence gathering, calibration of instruction, and student feedback or tutoring.

Another important tip in learning progression use is to make sure that the learning progression includes both (a) articulate descriptions of accomplishment and (b) representative student answers at each level. Including student exemplars in the progression levels can help teachers to more accurately identify incomplete or partial answers and to guide particular instructional interventions to challenge and clarify areas of misconceptions.

Assessment tasks coordinated with a learning progression can be empirically evaluated to provide information on both students' proficiency levels and assessment task difficulty. A Wright map presents assessment tasks' data on a continuum with most difficult tasks on top and easier tasks toward the bottom. Wright maps also map student proficiency on the same scale allowing teachers and others to identify a great deal of information on student progress along a learning progression. In these ways, new developments in educational measurement can combine with new developments in learning progressions to reveal more information about student progress, challenges, and knowledge development.

Nancy Butler Songer

See also Learning Maps; Learning Theories

Further Readings

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. The Journal of Research in Science Teaching, 50(5), 597–626.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Washington, DC: The Council of Chief State School Officers (CCSSO). Retrieved May 3, 2016, from http://www.ccsso.org/Resources/Publications/Learning_Progressions_Supporti

National Research Council. (2006). Systems for state science assessment. Committee on Test Design for K-12 Science Achievement. In M. Wilson & M. Bertenthal (Eds.), Board on testing and assessment, center for education. Division of behavioral and social sciences and education. Washington, DC: National Academies Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. Journal of Research in Science Teaching, 46(6), 716–730.

Kelley Buchheister Kelley Buchheister Buchheister, Kelley

Learning Styles

Learning styles

959

961

Learning Styles

There are different theories on learning styles, but they generally posit that each person has a dominant learning style and can learn best by using that style (e.g., visual learners can learn best through a visual presentation of information). Related to these theories is the idea that teachers can be most successful when they understand their individual students' learning styles and effectively match each student's unique, individual style to the strategies and methods they employ in the classroom. The idea that instruction is best provided in a way that matches the individual's learning style is known as the meshing hypothesis.

Research evidence for these ideas is lacking. Although there is empirical evidence that children and adults express preferences about how they want to present information or have information presented to them, these preferred styles or techniques may exhibit themselves in an eclectic mix of styles depending on the content area, environment, or experiences of the individual. This entry further defines learning styles and discusses research into efforts to identify students' learning styles and match instructional approaches to each student's style. It then looks at the idea that it is optimal for all learners that information be presented in multiple ways and describes an instructional approach corresponding to this idea.

Learning styles identify common ways that individuals gravitate toward acquiring knowledge and skills. Lynne Baldwin and Khaled Sabry define learning styles as the way in which "learners approach the task of learning differently, and use a pattern of behaviour that they have developed over time" (p. 325). An individual's preferred style is said to guide the way in which the individual learns by directing (a) how the individual internally or externally represents experiences or knowledge, (b) the manner through which the individual recalls or applies information, and/or (c) the communication style—including word choice and mannerisms—of the individual. Learning style models and methods of measuring individual learning styles include David Kolb’s model and the Index of Learning Styles developed by Richard Felder and Barbara Soloman.

Although ideas about learning styles are often at the forefront of educational discussions and emphasized in the context of teaching and learning, few studies have used rigorous methods to test the validity of the claims surrounding learning styles theories when applied to educational settings. Studies have explored the connection between students’ motivation and teaching practices that are matched to their specific, preferred learning styles. Although research shows that individuals express personal preferences for learning in specific ways, there is a lack of research-based evidence that indicates identifying students’ particular learning styles produces better educational outcomes.

Furthermore, results from studies using rigorous research methods and statistical analyses contradict the popular claims directly connecting an individual’s academic success to the individual’s identified learning style (e.g., visual, kinesthetic, auditory) and teachers’ instructional methods. There is a dearth of existing evidence regarding students’ reliable exhibition of particular learning styles over an extended period, or that learning outcomes or progress are significantly increased when teachers put forth the effort to match their instruction to students’ learning preferences.

It may be necessary for learners to experience modes of learning they do not prefer in order for them to fully develop certain skills or cognitive qualities. In fact, Felder has argued that education and learning experiences should incorporate various learning styles in order to achieve balance between accommodating and enriching an individual’s unique learning experience. Catering exclusively to a single learning preference can put the learner at a disadvantage because the learners are not provided with opportunities that encourage their full development.

How individuals learn is affected by factors such as engagement in educational or developmental activities, differences in prior knowledge or experiences, and what they attend to within each presentation of information. It is important to consider how to successfully engage all students in experiences that attend to consider how to successfully engage all students in experiences that attend to students’ strengths, interests, and needs and consider how to structure the classroom environment to provide students with opportunities to achieve success.

The philosophy of universal design for learning corresponds to this idea, as through this design, teachers consider how flexible methods of *presentation* (e.g., multiple modes of representation, varied contexts or situations), *expression* (e.g., sharing mathematical thinking through various modalities and mediums) and *engagement* (e.g., being aware of learners’ interests and strengths) can be incorporated into instructional practice. When using a universal design for a learning approach, teachers do not specifically attempt to mesh a unique, preferred learning style to particular instructional methods. Instead, universal design for learning emphasizes the need for teachers to present information in multiple ways and for students to have multiple ways of asking questions and showing what they learned.

Kelley Buchheister

See also Cattell–Horn–Carroll Theory of Intelligence; Emotional Intelligence; *g* Theory of Intelligence; Intelligence Quotient; Learning Theories; Multiple Intelligences, Theory of; Universal Design in Education

Further Readings

Baldwin, L., & Sabry, K. (2003). Learning styles for interactive learning systems. Innovations in Education and Teaching International, 40(4), 325–340. doi:10.1080/1470329032000128369

Claxton, C., & Murrell, P. (1987). Learning styles: Implications for improving educational practice (ASHE-ERIC Higher Education Rep. No. 4). College Station, TX: ASHE.

Felder, R. (1993). Reaching the second tier: Learning and teaching styles in college science education. Journal of College Science Teaching, 23(5), 286–290.

Felder, R. (1996). Matters of style. ASEE Prism, 6(4), 18–23.

Gilakjani, A. P. (2012). A match or mismatch between learning styles of the learners and teaching styles of the teachers. International Journal of Modern Education and Computer Science, 11, 51–60. doi:10.5815/ijmecs.2012.11.05.

Honey, P., & Mumford, A. (1992). The manual of learning styles (2nd ed.). Maidenhead, UK: Author.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. Psychological Science in the public interest, 9(3), 105–119.

Zhang, L. (2006). Does student-teacher thinking style match/mismatch matter in students' achievement? Educational Psychology, 26, 395–409.

Kelley Buchheister Kelley Buchheister Buchheister, Kelley

Learning Theories

Learning theories

961

965

Learning Theories

Learning theories describe views regarding how one acquires knowledge and creates connections among the items of information encountered in the world. In the field of education and child rearing, learning theories do not prescribe exact pedagogical strategies or instructional methods. However, the types of experiences a teacher, caregiver, or mentor provides the learner are influenced by his or her understanding of how people learn. This entry provides an overview of the learning process, describes three major learning theories and their instructional implications, and discusses the key principles in these theories that can be used to provide opportunities for learning.

The Learning Process

Advances in science over the last 50 years have allowed researchers to make considerable strides in documenting the thinking and learning process. Cognitive scientists have begun collaborating with educators, child development experts, and researchers to bring together ideas within a formal educational context.

Although children’s minds were once believed to be essentially blank slates that must be filled with information, observations of young children have indicated that human learning begins at early stages of infancy through an active, rather than a passive, learning process. In fact, infants attend to particular information such as language, number, physical attributes or properties, and spatial sense prior to formal instruction of these concepts. These findings provide the foundation of the idea that learning is based on experience and that knowledge develops as individuals build on the ideas generated from these experiences, then subsequently reflecting on their conclusions. Therefore, the three principles that define the process of thinking and learning are as follows: (1) humans glean isolated bits of information from their experiences—both natural encounters and formal instruction; (2) the individual generates networks of related knowledge to prioritize, connect, and organize the collected information; and (3) the learners take control of learning by identifying a purpose, defining a plan to achieve the purpose, and monitoring their process toward achieving the desired goals.

Consider a 4-year-old child playing with wooden blocks. Her exploration leads her to arrange the blocks in various ways, from a horizontal line of blocks on the floor, to a more complex cognitive feat of balancing blocks vertically to construct a castle tower. She begins by attending to bits of information she collects during the experience as she observes the attributes and properties of the wooden blocks. Then, through trial and error, she begins to notice features of the blocks and her construction methods that either inhibit her building process or contribute to her architectural success. She continually reflects on her experiences, talking to herself during the play scene, as she organizes and connects information she gleaned through her observations to become a more productive builder. The young engineer reminds herself of these conclusions as she finalizes her construction after earlier attempts to build a tower ended in the blocks collapsing. She begins to place a long, wide rectangular block atop a smaller rectangular prism near the apex of the tower, then seeing it wobble, removes it and puts it at the base.

To understand this as a learning experience, one must examine how the individual acquired knowledge. In this case, the young child generated information regarding the attributes of the blocks that she later could relay to her teacher: “I need a flat one like this, but this [pointing to the long, wide rectangular prism] is too big and too heavy for this skinny one on top [points to the top block in the tower]. It keeps falling.” Through her construction attempts, she was able to prioritize and relate the properties of the blocks to her experiences with trial and error as she built the tower. Finally, by reflecting on her experience with her teacher, she was able to apply the newly acquired knowledge in a subsequent building experience as she and a friend used blocks to build a castle later in the week.

This scenario provides a glimpse into the three general principles of learning (i.e., generating experiences, developing relations among gathered elements of information, reflecting on the connections and events). However, in 1996, as Ernst von Glasersfeld argued, a single theory “cannot claim to be anything but

Ernst von Glasersfeld argued, a single theory "cannot claim to be anything but one approach to the age-old problem of knowing. Only its application in contexts where a theory of knowing makes a difference can show whether or not it can be considered a viable approach" (p. 309). Thus, it is imperative that stakeholders in cognitive science, human development, and education navigate divergent perspectives and frameworks on learning set forth by various theorists. The next three sections define three major learning theories: behaviorism, constructivism, and sociocultural theory. These theories are later described in the context of children using blocks to further accentuate the characteristics of each theory.

Behaviorism

Behaviorism, in which learning is achieved through trial and error and measured by the observable relationship between a stimuli and a response, was the predominant learning theory in the early 20th century. Behaviorists of the era emphasized that psychological studies must solely attend to observable behaviors and the contributing stimuli. Specifically, John B. Watson maintained that human psychology through the behaviorist perspective is governed by the actions or behaviors of the human being and that consciousness or introspectiveness are neither definitive nor usable concepts. Motivating factors in behaviorism are comprised of external forces such as reward and punishment. As a result, the identification and definition of learning could be explained without focusing on internal, unobservable mental actions or images.

Constructivism

The ideas behind constructivism began to emerge in the early 20th century. One of the key ideas of constructivism is Jean Piaget's theory that the development of thinking and learning occurs in stages. Based on extensive observations of his own children, Piaget developed four major stages of cognitive development that humans move through from more concrete thinking to abstract or symbolic interpretations and logical reasoning.

Constructivist theorists perceive learning as an active, individual process that occurs as a result of resolving problematic situations. In this perspective, a young child learns by setting a goal and planning to accomplish this goal. Learning occurs as the child experiences change that follows disequilibrium or an internal imbalance that requires the construction of a new structure to regain balance. This process of cognitive self-regulation equilibration corresponds to the ability of the child to either assimilate new knowledge into his existing mental structure or accommodate the mental structure to incorporate new knowledge. In this way, as the child encounters novel information, he is able to assimilate the experience or idea within his existing knowledge base. However, if the existing schema, or internal, mental image does not coincide with the novel information, the child must alter his mental structures to accommodate the new experience. Thus, a child facing disequilibrium spurs this process of assimilation and accommodation and revises his actions to regain equilibration.

Sociocultural Theory

Similar to the constructivist theory of learning, sociocultural theory, sometimes referred to as social constructivism, emphasizes active, rather than passive, thought and views learning in the context of activity. Yet while Piaget attended to the construction of knowledge, his stage theory implies that cognitive development promotes language development because language reflects the individual’s existing knowledge with little contribution to new understanding. Sociocultural theory, attributed to several theorists and often associated with Lev Vygotsky, is rooted in socially negotiated knowledge where language facilitates cognitive development. In fact, Vygotsky emphasized the role of language within the social environment as integral to cognitive development.

Martin J. Packer and Jessie Goicoechea further highlight a key premise of sociocultural theory indicating that cognitive development is shaped by the cultural environment and founded in purposive activity. In 2000, the researchers argue that when one assumes the individualistic nature of learning, one

> fails to grasp the affective, relational, and cultural dimensions of activity, … [for example] constructivism also can take for granted the objective appearance of the world and fail to recognize its cultural and historical basis; the objects we know are also products of human activity. (p. 235)

Thus, sociocultural theory separates itself from constructivism and further distances itself from behaviorism, as the sociocultural perspective situates the act of learning within a position of cooperative participation in cultural practices.

What is learned is rooted in the context in which the experience takes place—including the tools, people, discursive practices, or cultural practices with whom the individual interacts. Thus, learning—according to sociocultural theorists—is a social activity that exists only through linguistic interactions with other humans or cultural artifacts. In particular, Vygotsky identified the role of (a) interactions with a more knowledgeable other and (b) culturally developed sign systems within the sociocultural perspective.

Instructional Implications

While each of these theories constitutes a theory of learning, rather than an epistemological perspective, one's theoretical view has powerful implications for the types of experiences and support provided to the learner. In this section, the discussion returns to the preschool classroom block center and provides an account of how an educator, mentor, or caregiver might organize and support the learning environment on the basis of each learning theory.

Behaviorism

As a child builds a tower, the teacher supports the experience by reinforcing desired behaviors and providing opportunities for the child to copy the teacher's modeled actions. For instance, when a child attempts to build his tower and successfully balances and stacks the blocks, the teacher acknowledges his accomplishment with positive affirmations or with nonverbal gestures such as a smile or a thumbs-up. Moreover, when the child experiences difficulties stacking the blocks with different shapes, the classroom teacher models the desired action and encourages the child to copy the steps until he is successful. In this case, there is little interaction between the teacher and the learner.

Constructivism

Constructivist theory is grounded in the fact that knowledge is not passively received but actively constructed. In the case of building a block tower, from this perspective, the teacher attends to how the child assimilates and accommodates information as she experiences disequilibrium. For instance, as the child builds the tower, her actions seem to indicate that she recognizes that the attributes of the rectangular prism deem that shape as a prime structure for building. However, as she attempts to put a large, wide rectangular prism on top of a narrow tower of blocks, she is faced with a conundrum, as the prism is causing narrow tower of blocks, she is faced with a conundrum, as the prism is causing the tower to fall. Here, the child must coordinate this novel information with prior knowledge and experience. Her learning, through the constructivist perspective, is seen as the process of assimilating and organizing information about the shape of the block into her existing schemas.

Sociocultural theory

From the sociocultural perspective, influential instruction includes interactive talk and scaffolding where a more knowledgeable other assists and guides the novice learner toward more efficient strategies. As in the constructivist perspective, a child might exhibit learning by testing out hypotheses, analyzing data, and experimenting; however, learning through the sociocultural lens is explicitly encouraged by meaningful exchanges through questioning and reflection that allow the learner to negotiate meaning and develop relational understanding.

In the context of the block play, the play experience itself provides an opportunity to cognitively benefit the child; however, more powerful possibilities occur when teachers engage the learner in reflecting on the ideas that emerged through the experience. Thus, while it may seem that through the sociocultural perspective a child is constructing knowledge as he assimilates information into his existing schemas during his building experience, the substantial difference in sociocultural theory is the role of the more knowledgeable other and the cultural artifacts in contributing to the learner’s negotiation of meaning. In this case, a teacher who ascribes to this theory of learning aids the child in making connections to lived experiences, assists the child in negotiating meaning surrounding the attributes of the shape, and encourages reflection by providing questions for the child to consider about his observations and actions.

Key Principles for Encouraging Learning

While variation among learning theories exists, there are several considerations that remain constant with regard to cognitive development. For instance, all three of the learning theories described earlier start with the premise that all learning is based on experience and that knowledge develops as individuals build on the ideas generated from these experiences while reflecting on their conclusions. In behaviorism, behavior changes based upon the reinforcement of desired behaviors previously exhibited. Constructivist theory emphasizes the active construction of knowledge through changes that occur following a period of disequilibrium where a novel experience does not coincide with existing knowledge. Therefore, the learner must assimilate or accommodate the newly gained information into his or her existing schemas. Finally, sociocultural theory builds knowledge through the negotiation of the meaning of new experiences among more knowledgeable others, peers, and individuals within communities of practice. Furthermore, in both constructivism and sociocultural theory, the individual generates networks of related knowledge to prioritize, connect, and organize the collected information, whether it be through assimilation and accommodation or through negotiating what is taken to be a shared understanding.

All of these theories would indicate that experiences that involve authentic, engaging interactions provide the greatest opportunities for learning. Moreover, by creating representations that allow them to express, clarify, and reflect upon their ideas and by working to understand others' representations, learners can develop integrated networks of knowledge that promote knowledge transfer and deep conceptual understanding.

Kelley Buchheister

See also Behaviorism; Cognitive Development, Theory of; Constructivist Approach; Educational Research, History of; Learning Styles

Further Readings

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). Learning: From speculation to science. In National Research Council, How people learn: Brain, mind, experience, and school (expanded ed., pp. 3–28). Washington, DC: National Academies Press. doi:10.17226/9853

Glasersfeld, E. von. (1996). Aspects of radical constructivism and its educational recommendations. In L. P. Steffe, P. Nesher, P. Cobb, G. A. Goldin, & B. Greer (Eds.), Theories of mathematical learning (pp. 307–314). Mahwah, NJ: Erlbaum.

Lerman, S. (1989). Constructivism, mathematics and mathematics education.

Educational Studies in Mathematics, 20, 211–223.

Packer, M. J., & Goicoechea, J. (2000). Sociocultural and constructivist theories of learning: Ontology, not just epistemology. Educational Psychologist, 35(4), 227–241.

Piaget, J. (1977). The development of thought: Equilibration of cognitive structures. New York, NY: Viking.

Skinner, B. F. (1938). The behavior of organisms. New York, NY: Appleton-Century-Crofts.

Thomas, N. J. (2014). Founders of experimental psychology: William Wundt and William James. Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/mental-imagery/founders-experimental-psychology.html

Thorndike, E. L. (1913). Educational psychology: Vol. 1. The original nature of man. New York, NY: Teacher’s College Press.

Van Scoy, I. J. (1995). Trading the 3 r’s for the 4 e’s: Transforming curriculum. Childhood Education, 72(1), 19–23.

Vygotsky, L. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman., Eds.). Cambridge, MA: Harvard University Press.

Pamela Williamson Pamela Williamson Williamson, Pamela

James McLeskey James McLeskey McLeskey, James

Least Restrictive Environment Least restrictive environment

965

968

Least Restrictive Environment

The passage of the Education for All Handicapped Children Act of 1975 (EAHCA), or Public Law 94-142, entitled students with disabilities, aged 3–21 years, to a free appropriate public education in the least restrictive environment (LRE). The reauthorization of Public Law 94-142, later renamed the Individuals with Disabilities Education Act (IDEA), conceptualized LRE as follows:

> [t}o the maximum extent appropriate, students with disabilities, including students in public or private institutions or other care facilities, are educated with students who are not disabled, and that special classes, separate schooling, or other removal of students with disabilities from the regular educational environment occurs only when the nature or severity of the disability is such that education in regular classes with the use of supplementary aids and services cannot be achieved satisfactorily (IDEA, 20 U.S.C. § 1412 (612)(a)(5)(A)).

This definition presumes that students with disabilities will be served in general education settings with appropriate support and services and moved to more restrictive settings without general education students only if the student’s disability prevents this. The LRE mandate leaves room for interpretation; thus, the definition continues to evolve through legislation and case law. This entry describes the legislative origins of LRE, the legal influences on this concept, and national trends in student placements.

Legislative Origins and Evolution of LRE

Legislative Origins and Evolution of LRE

Following the *Brown v. Board of Education* landmark civil rights case that determined a separate school system violated the rights of African American students, two right-to-education cases for students with disabilities, the *Pennsylvania Association for Retarded Children (PARC) v. Commonwealth of Pennsylvania* (1971) and *Mills v. District of Columbia* (1972), successfully argued that the exclusion of students with disabilities also violated the Equal Protection Clause of the 14th Amendment of the U.S. Constitution. In 1972, Congress launched an investigation into how students with disabilities were being educated. This investigation revealed that fewer than half of all students with disabilities were receiving an appropriate education, whereas 1.75 million of these students were receiving no education at all. This prompted Congress to pass the Education for All Handicapped Children Act in 1975, which mandated a free appropriate public education and LRE for all students with disabilities.

When Education for All Handicapped Children Act was reauthorized in 1990, the act was renamed the IDEA. Regulations of the 1990 reauthorization specified that local education agencies needed to provide a continuum of placements outside the general education setting, should individuals’ needs dictate more restrictive settings. IDEA was reauthorized in 1997 and in 2004. In 2004, Congress clarified that state funding should not include mechanisms that influenced placements, such as providing more state dollars for educating students in more restrictive placements.

Although IDEA has been reauthorized multiple times, the definition of LRE continues to lack clarity. This has resulted in disagreements between individual families and local education agencies regarding the implementation of LRE. Through due process hearings, disagreements might be mediated locally, and if disagreements persist, they may be challenged in district courts and circuit courts. Mediations and district court rulings impact individual cases, whereas circuit court rulings become case law in that circuit. Although circuit court rulings only have direct influence in the circuit where they were settled, circuit courts tend to notice the findings in other circuits. The Supreme Court, which would influence every circuit, has not heard a case involving the implementation of LRE. The remainder of this section provides summaries of circuit court cases that settled disputes in the areas of (a) placement, (b) proximity of placement to home-zoned school, (c) educational methodologies, and (d) educational benefit.

Placement Cases

Placement Cases

In the case of *Roncker v. Walter* in 1983, the evaluation results for a student with an intellectual disability suggested the student would benefit from being educated with general education students. The school district instead placed the student in a setting with other students with intellectual disabilities. The court ruled that the school district abused its discretion by placing this student in a more segregated placement.

In the case of *Kerkam v. McKenzie* in 1988, the court sided with a school district that placed a student with an intellectual disability in a day program within the district, rather than the private residential setting preferred by the family. The court noted that the district was responsible for providing an appropriate education for the student and not necessarily a placement that maximized the student's potential, as was argued by the family. In the case of *Oberti v. Board of Education of the Clementon School District* in 1993, the court ruled in favor of the family who wanted their child with Down syndrome to be placed more than half of the time in the general education setting. The court noted that the school district failed to provide appropriate support services to the student in the general education setting to help the student be successful.

Proximity to Home-Zoned School

In the case of *Schuldt v. Mankato Independent School District* in 1990, the family of a student with spina bifida who was paralyzed from the waist down challenged the school district's decision to bus their child to a school other than the home-zoned school to accommodate the child's physical needs. The court upheld the school district's decision, noting that the law did not require the district to modify the student's home school, as long as the district provided the student placement in a suitable setting where the student's needs could be met.

In the *Barnett v. Fairfax County School Board* 1991 case, the circuit court decided that placing students where there was a special speech program instead of offering the speech program of the home school was not in violation with the law. In the case of *Poolaw v. Bishop* in 1995, the courts found in favor of a school district that placed a student who was deaf in a school 280 miles from his home to learn American Sign Language. This was the closest school that could provide these services, and it was determined that in order for instruction to be appropriate, the student needed to learn American Sign Language. The courts also ruled that a day school, some distance from the student’s home, was an appropriate placement for a student who was deaf in the *Flour Bluff Independent School District v. Katherine M* in 1996. This ruling affirmed that appropriateness of a placement was more important than distance from home.

Educational Methodologies

In the case of *Lachman v. Illinois State Board of Education* in 1988, the court ruled in favor of a school district that placed a student who was deaf in a placement for at least half of the day with other students with hearing impairments instead of what the parents preferred—full-time placement in a general education setting. The ruling noted that it was up to state boards of education, not courts, to determine the use of educational methodologies.

Educational Benefit

Daniel R. R. v. State Board of Education in 1989 challenged a district’s placement of a student with an intellectual disability in a special education placement after determining that the student was not making progress in a combined general education and special education setting. The court ruled in favor of the district. This case is the origin of the so-called two-prong test for determining a student’s placement. First, can the student be successful in a general education setting with the use of appropriate supports? Second, is the student educated with general education peers appropriate to the student’s needs?

The two-prong test was subsequently applied in the court ruling in *Greer v. Rome City School District* in 1991. Records from the individualized education program meeting demonstrated that only two placement options were considered—the district wanted to place the student in a self-contained classroom, whereas the family wanted their child placed in a general education classroom. The court ruled against the district on the basis of the first test—the district did not share appropriate support options to help the student remain in the general education setting.

In 1994, the court developed a four-factor test to determine a student’s placement in the case of the *Sacramento City Unified School District v. Rachel H*. In this case, the parents of a child with a moderate disability disagreed with the district’s decision to place their child in a self-contained classroom for academic instruction. The court considered (a) comparisons of the educational benefits between general and special education classrooms, (b) comparisons of the social benefits between general and special education classrooms, (c) effects of the student with disabilities on the teacher and other students in the classroom, and (d) the cost of educating the student with nondisabled peers. After consideration of the answers to these questions, the court ruled in favor of the family on the basis that the school district failed to provide evidence that the student’s needs would be better met in the self-contained classroom.

The case of *Clyde v. Puyallup School District* in 1994 asked the court to settle a temporary placement dispute. The district temporarily moved a student with Tourette syndrome from a general education setting to a resource room in response to the student’s increased disruptive behaviors. The court ruled in favor of the district, noting that the student’s behavior prohibited his academic learning and social learning in the classroom. The court also noted that the behaviors were disrupting peers. Thus, the student’s lack of progress was used to support the change in placement.

Hartmann v. Loudoun County Board of Education in 1997 settled a case that involved a student with autism spectrum disorder whose district placed him in a self-contained classroom against the wishes of his family. The circuit court ultimately ruled that the preference for educating students with disabilities alongside students without disabilities should be weighed against the progress of both the student with disabilities and general education peers.

National Placement Trends

Since the implementation of Public Law 94-142, later named IDEA, states have collected student placement data in compliance with the law. In 2009, the U.S. Department of Education defined settings based upon the percentage of time students spend in general education classrooms. Thus, students who spend 80% or more of the school day are considered to have a *general education placement;* students who spend 40–79% of the day in general education classrooms are considered to be educated in *pullout settings*; students who are served less than 40% of the day in a general education classroom are considered to be educated in *separate classrooms*; and students who are served in public or private separate settings, including public or private residential facilities, homebound or hospital settings are considered to be educated in *separate schools*.

Analyses of trends in placements have revealed that since the law was implemented, students with disabilities are far more likely to be educated in general education classrooms. This is true across every disability category and age level. There continue to be substantial differences in placement across disability categories; for example, students with learning disabilities are much more likely to spend most of the school days in general education classrooms than students with emotional disabilities.

Pamela Williamson and James McLeskey

See also *Brown v. Board of Education*; Inclusion; Individuals with Disabilities Education Act; Policy Evaluation; Special Education Law

Further Readings

Crockett, J. B. (2014). Reflections on the concept of the least restrictive environment in special education. In B. G. Cook, M. Tankersley, & T. J. Landrum (Eds.), Advances in learning and behavior disabilities: Vol. 27. Special education past, present, and future: Perspectives from the field (pp. 39–61). Bingley, UK: Emerald Group.

Marx, T., Hart, J., Nelson, L., Love, J., Baxter, C., Gartin, B., & Whitby, P. (2014). Guiding IEP teams on meeting the least restrictive environment mandate. Intervention in School and Clinic, 50(1), 45–50.

McLeskey, J., Landers, E., Williamson, P., & Hoppey, D. (2012). Are we moving toward educating students with disabilities in less restrictive settings? The Journal of Special Education, 46(3), 131–140. doi:10.1177/00224669-10376670

Rozalski, M., Miller, J., & Stewart, A. (2011). Least restrictive environment. In J. M. Kauffman & D. P. Hallahan (Eds.), Handbook of special education (pp. 107–119). New York, NY: Routledge.

Yell, M. (2015). The law and special education (4th ed.). Boston, MA: Pearson Education.

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

Levels of Measurement Levels of measurement

968

969

Levels of Measurement

Also known as scales of measurement, levels of measurement describe how data variables, numbers, and associated attributes are defined and categorized. Mathematical operations and statistical techniques have requirements that must be met in order for meaningful data analysis and interpretation to be undertaken. An assessment of the data measurement level facilitates determining the appropriate statistical analysis to use based on the data parameters. Listed in order of increasing variable complexity, four major levels of measurement are commonly identified: nominal, ordinal, interval, and ratio. This entry describes each of these levels.

Nominal level is the lowest or simplest measurement level. Information is assigned to categories that are mutually exclusive (a single group) and all-inclusive (contain all cases). The categories do not have any meaningful ordering and only denote whether the information should be assigned to a particular group. For instance, nominal-level data may include information on eye color (e.g., brown, blue, and green), religious affiliation (e.g., Christian, Muslim, and Buddhism), and political orientation (e.g., democrat, republican, and libertarian), among others. As per these examples, the data provide qualitative or named information, although nominal-level data can have quantitative values. Arbitrary number codes may be assigned to groups, such as coding gender as 1 = female and 2 = male; however, the values do not denote magnitude or ordering and no mathematical operations are possible. Nominal data are often displayed in pie, bar, or line charts showing the number or percentage of cases assigned to a particular group.

Ordinal level implies an ordering or ranking relationship among the measurements. In contrast to the nominal scale, more quantitative measures can be made. Ordinal-level variables are either strongly ordered or weakly ordered. Strongly ordered data assigns a ranking to each individual value or data unit in an ordered sequence. *Consumer Reports* ranks products according to several criteria and assigns each product a number indicating those that are the best and worst performers. Each product holds a particular position in the sequence, but the ranking does not indicate how much better (or worse) the products are compared to one another.

For a weakly ordered variable, data are placed in groups and the groups themselves are ranked; thus, each group consists of frequency counts rather than individual rankings. Agreement-response surveys (e.g., Likert-scale assessments) are a good example of weakly ordered data, where responses are grouped according to relative agreement with a statement (e.g., strongly agrees, agrees, undecided, disagrees, or strongly disagrees). Regardless of whether the variable is strongly or weakly ordered, the differences and ratios between rankings are not meaningful, only the relative order.

Interval level specifies quantitative information on the exact differences or intervals between successive values on a continuous number line. Unlike ordinal-level data, the difference between values is meaningful and indicates a magnitude change. However, the magnitude of this difference is not comparable across scales with different measurement units.

Interval scales are also characterized by an arbitrary (nontrue) zero. A common example given to illustrate this point involves the Fahrenheit and Celsius temperature scales. On both scales, zero degree does not indicate an absence of heat (or average kinetic energy), only a subjective point on which higher and lower heat values are determined. Meaningful differences between points on their respective scales can be found (e.g., 40° separates 80°F and 40°F), but ratios between those points cannot be established (e.g., 80°F does not indicate twice as much heat as a 40°F). Other examples include shoe sizes, calendar years, standardized exam scores, pH levels (acidity or alkalinity measure), and intelligence tests. In the social and behavioral sciences, most measurement scales are at the interval level, whereas the ratio level is more common in economics, business, and the physical sciences.

In addition to possessing the traits of the previous scales, ratio-level data have a true zero value, meaning the amount being measured is nonexistent. Accordingly, ratios (fractions) and other mathematical operations between values are possible. Variables such as distance, weight, income, and altitude all values are possible. Variables such as distance, weight, income, and altitude all have a nonarbitrary or natural zero. If the distance between Chicago and Los Angeles is 2,000 miles and the distance between Chicago and Denver is 1,000 miles, the ratio between these two measures is readily calculated (2,000/1,000 = 2); Los Angeles will always be twice as far from Chicago as Denver, regardless of the measurement units (e.g., miles, feet, and kilometers). Zero miles signify a natural zero, in this case meaning no distance from Chicago.

Ratio (and interval) data are quantitative and represent points along continuous number lines as opposed to separate categories; hence, both descriptive and inferential statistical analysis can be undertaken. For this reason, many statistical tests and statistics analysis packages (e.g., SPSS) often do not differentiate between ratio and interval-level data.

Levels of measurements indicate a data hierarchy from the least (nominal) to the most complex (ratio). Each level includes all data characteristics and assumptions of previous levels while incorporating an additional attribute. Data can be transformed from higher levels of measurement to lower levels of measurement but not the reverse. For example, ratio-level data can be converted to ordinal level by assigning ranks to each value from lowest to highest or grouping data into low, medium, and high categories; however, the ranked data cannot be transformed into meaningful intervals or ratios between values. Consequently, the level of measurement required for data analysis is an important consideration during the data collection phase of a project.

Jill S. M. Coleman

See also Interval-Level Measurement; Likert Scaling; Nominal-Level Measurement; Ordinal-Level Measurement

Further Readings

Coolidge, F. L. (2012). Statistics: A gentle introduction. Thousand Oaks, CA: SAGE.

Tokunaga, H. T. (2015). Fundamental statistics for the social and behavioral sciences. Thousand Oaks, CA: SAGE.

Witte, R. S., & Witte, J. S. (2013). Statistics (10th ed.). Hoboken, NJ: Wiley.

Yi-Hsin Chen Yi-Hsin Chen Chen, Yi-Hsin

Yan Wang Yan Wang Wang, Yan

Jeffrey D. Kromrey Jeffrey D. Kromrey Kromrey, Jeffrey D.

Levene’s Homogeneity of Variance Test Levene’s homogeneity of variance test

969

972

Levene’s Homogeneity of Variance Test

Homogeneity of variance (HOV) is one of the assumptions of some frequently used statistical procedures for group mean comparisons, such as a one-way analysis of variance (ANOVA) or an independent-samples *t* test. Under the HOV assumption, population variances of all groups are assumed to be equal. That is, the null hypothesis (H_0) being tested for verifying the HOV assumption is that the population variances across groups are equal; that is, , where *k* denotes the number of groups compared in a study.

The examination of the HOV assumption is always an essential step before conducting a comparison of group means using ANOVAs or *t* tests. Violations of the HOV assumption may result in misleading results in terms of the test for differences in group means. Levene’s test is one of the popular tests that have been employed to assess the tenability of the HOV assumption. It was proposed by Howard Levene in 1960 as an alternative to Bartlett’s test that is sensitive to departures from normality. This means that when the underlying distributions of the data deviate from normality or approximate normality, Levene’s test is expected to perform better than Bartlett’s test. This entry introduces the formula and variations of Levene’s test, statistical software available for conducting Levene’s test, and its performance under various conditions.

Mathematical Formula and Variations of Levene’s Test

The Levene test statistic, denoted *W*, is defined as

$$W = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{Z}_{.j} - \bar{Z}_{..})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \bar{Z}_{.j})^2} \text{ and } Z_{ij} = \left|Y_{ij} - \bar{Y}_{.j}\right|,$$

where

> Y_{ij} is the value of observation *i* in group *j*,
> is the group mean of group *j*,
> Z_{ij} is the absolute value of the deviation score from the mean for observation *i* in group *j* (Y_{ij}),
> is the group mean of the deviation scores from the mean in group *j*,
> is the grand mean of the deviation scores from the mean;
> *N* = total sample size,
> n_j = sample size of group *j*.

Levene's test rejects the null hypothesis that the variances are equal across groups if the *W* statistic of Levene's test is greater than the upper critical value of the *F* distribution ($F_{\alpha,\, k-1, N-k}$) with $N - k$ and $k - 1$ as degrees of freedom in the numerator and denominator, respectively, at the significant α level (typically α = .05 or .01). If the variances across groups are unequal based on Levene's test, alternative statistic approaches rather than the traditional *F* or *t* test are desired for group mean comparisons (the Satterthwaite or Welch approximate tests are popular alternatives that do not require variance homogeneity).

Originally, Levene used the absolute values of deviations of the observations from their respective group means in an ANOVA model. Using absolute deviations as observations in ANOVA models in place of the original observations, (Y_{ij}) transforms the test of means into a test of variances. Variations of the test suggested by Levene include replacing the deviations from the group means with the squared deviations from the group means , the square root of the deviations from the group means , and the natural logarithm of the deviations from the group means .

In 1974, Morton Brown and Alan Forsythe suggested using the median or the 10% trimmed mean to replace the mean in computing the deviations to control the chances of falsely detecting unequal variances (statistically called Type I the chances of falsely detecting unequal variances (statistically called Type I error rates) under nonnormal data. The approach of using the group median to yield the deviations is also referred to as the Brown–Forsythe test.

Comparing the robustness (i.e., does not falsely detect unequal variances) and power (correctly detects unequal variances) of using the mean, median, and trimmed mean based on Levene’s test, the trimmed mean performs best when the data distributions are extremely leptokurtic (i.e., heavily tailed distributions) and the median performs best when the underlying distributions of the data are extremely skewed (e.g., many higher scores or lower scores in the distributions). As for using the mean, it yields best power when the data are symmetric, moderately tailed distributions.

Statistical Software Available for Levene’s HOV Test

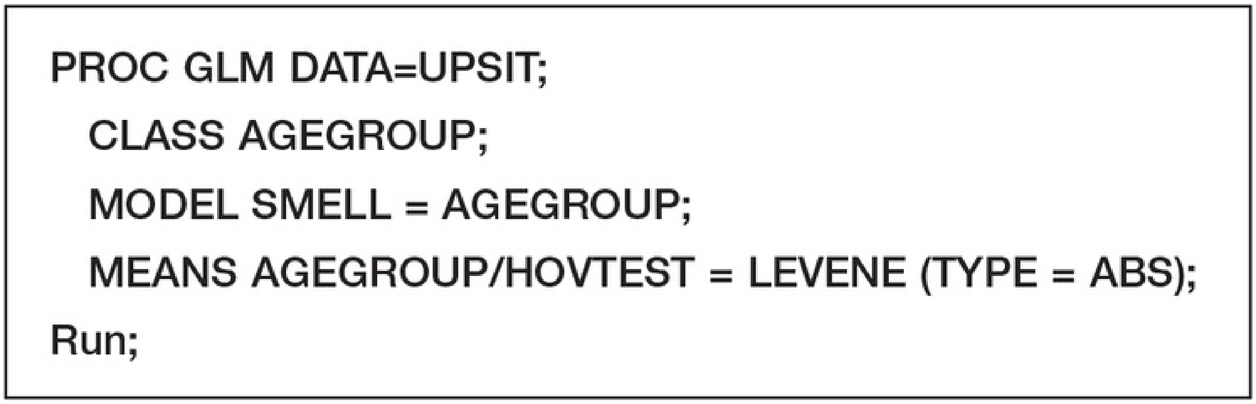
Levene’s test and its variations can be performed in a variety of statistical programs, such as Statistical Analysis System (SAS), Predictive Analytical SoftWare (commonly called SPSS), and many statistical packages in R. In SAS, Levene’s HOV test can be requested in the general linear modeling procedure (PROC GLM) by specifying the HOVTEST = LEVENE option in the MEANS statement. There are two types of deviations available in SAS: the absolute deviations, HOVTEST = LEVENE (TYPE = ABS), and the squared deviations as the default, HOVTEST = LEVENE (TYPE = SQUARE).

In SPSS, Levene’s test with the absolute deviations can be requested by checking “HOV test” in “Options” when conducting an independent-samples *t* test or a one-way ANOVA. In STATA, it can be conducted with the ROBVAR command. In the many R packages (e.g., LAWSTAT or CAR), Levene’s HOV test can be conducted with the LEVENETEST command directly, i.e., LeveneTest ( ). The LEVENETEST command with additional arguments can specify the different types of deviations. For instance, in the LAWSTAT R package, the command of LeveneTest (….., location = c(“mean”,“median”,“trim.mean”), trim.alpha=0.10,…) conducts the Levene tests with the absolute deviations from the mean, median, and 10% trimmed mean with the argument of “trim.alpha=0.10.” The Levene test is also available in BMDP, the biomedical statistical package developed at UCLA in 1965, and MINITAB statistical software.

Demonstration of Levene’s HOV Test in SAS

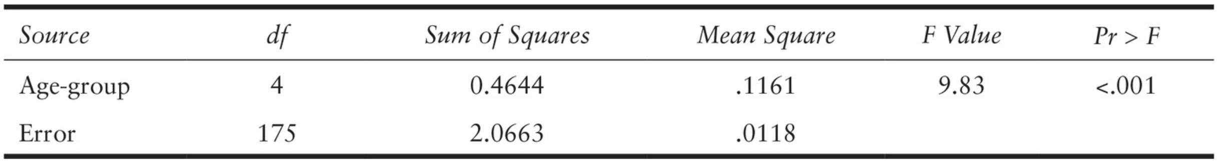
The following statements in Figure 1 illustrate the use of Levene’s HOV test in SAS to test equal variances in the sense of smell among five different agegroups, based on a sample of 180 individuals in a study Ralph G. O’Brien and M. W. Heft conducted in 1995, which can also be found at the SAS website.

Figure 1 SAS code for group means comparisons with Levene’s homogeneity of variance test



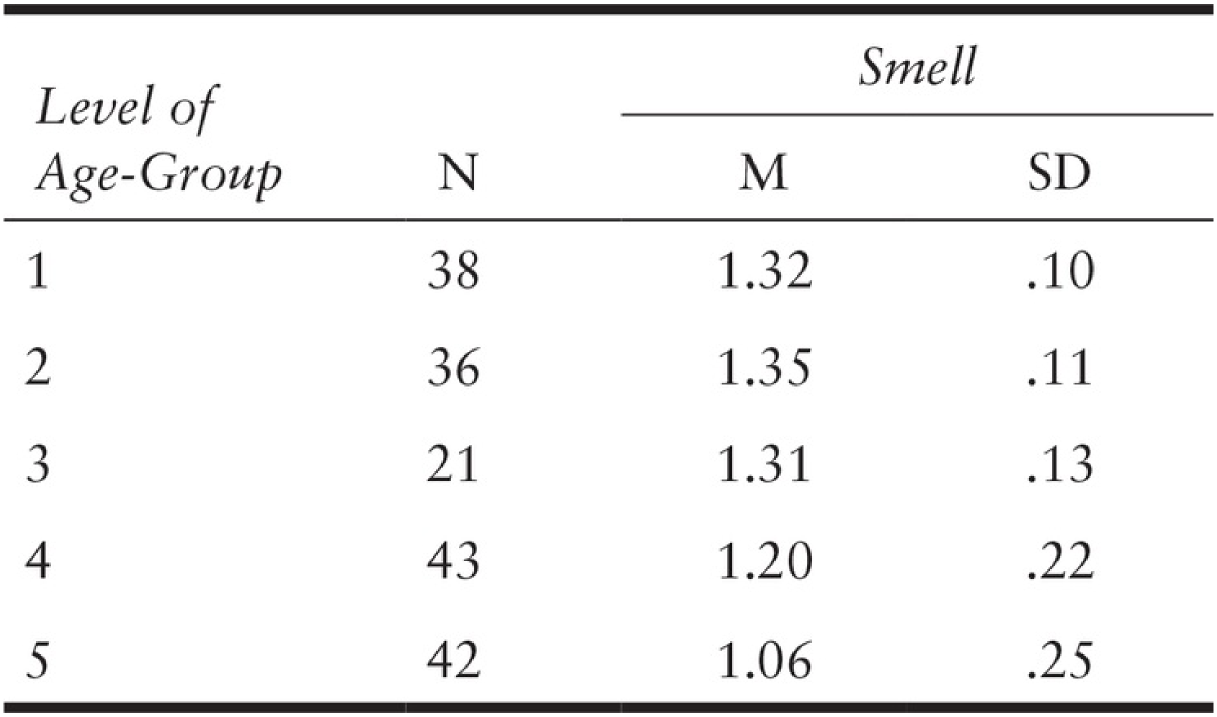
```
PROC GLM DATA=UPSIT;
   CLASS AGEGROUP;
   MODEL SMELL = AGEGROUP;
   MEANS AGEGROUP/HOVTEST = LEVENE (TYPE = ABS);
Run;
```

In Figure 1, the data set is named *upsit* and the independent variable is *agegroup*, with *smell* as the dependent variable. The HOVTEST = LEVENE option specifies the use of the Levene’s test. The absolute deviations were used by including the TYPE = ABS option. Table 1 displays the outputs of Levene’s test for homogeneity of smell variance with absolute deviations from group means for one-way ANOVA.



Source	*df*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Age-group	4	0.4644	.1161	9.83	<.001
Error	175	2.0663	.0118		

The HOVs in the sense of smell was not supported, $F(4, 175) = 9.83$ and $p < .001$. This could be confirmed by the outputs in Table 2 that show the substantial variations in the sense of smell for the fourth and fifth agegroups; that is, standard deviations are .22 and .25, respectively.



Level of Age-Group	N	Smell	
		M	SD
1	38	1.32	.10
2	36	1.35	.11
3	21	1.31	.13
4	43	1.20	.22
5	42	1.06	.25

Although Levene’s test with absolute or squared deviations, as well as Bartlett’s test, the Brown–Forsythe test, and the O’Brien test, is provided by SAS, alternative HOV tests within the one-factor ANOVA context can be conducted using an SAS macro developed by Diep Nguyen and colleagues in 2014.

Performance of Levene’s HOV Test

The performance of Levene’s test under various conditions has been examined in many simulation studies. It has been well known that Levene’s test with various types of deviations (i.e., absolute, square, or logarithm) from the group mean is less sensitive to departures from normality than Bartlett’s test but is inferior to the Brown–Forsythe test in terms of Type I error control (i.e., control of the probability of falsely detecting unequal variances) under nonnormality in both skewness and kurtosis. Levene’s test with absolute deviations has adequate Type I error control only under normal or approximately normal distributions, and it has inflated Type I error rates when the data have skewed or leptokurtic distributions.

Levene’s test with squared deviations controls Type I error adequately except for skewed distributions. Regarding the impact of average group size, the performance of Levene’s test with squared deviations in terms of Type I error control improves as average group size increases. However, increasing average group size does not improve the performance of Levene’s test with absolute deviations because it always has a poor control of Type I error regardless of average group size.

For the Levene’s test with squared deviations that has adequate Type I error control, its statistical power increases substantially when the ratio of the largest group variance to the smallest group variance increases. It also has greater power when the variance(s) of one group or several groups differ from the rest, compared with the conditions when group variances are equally spaced. Levene’s test with squared deviations has greater power under normal or platykurtic distributions than under skewed distributions. As average group size increases, the power of Levene’s test with squared deviations increases.

In 1995, Gene Glass and Kenneth Hopkins pointed out that there is a fundamental flaw in using Levene’s test and its variations to test the HOV assumption. Because Levene’s test is equivalent to an ANOVA model that examines equality of variances, it relies on the HOV assumption, that is, variances of the absolute deviations or its variations are equal across groups, similar to other ANOVA models. Therefore, results of Levene’s test are trustworthy regardless of the violations of the HOV assumption when group sizes are equal, while with unequal group sizes, Type I error rates of Levene’s test are not controlled adequately.

Yi-Hsin Chen, Yan Wang, and Jeffrey D. Kromrey

See also Analysis of Variance; *t* Tests

Further Readings

Bhandary, M., & Dai, H. (2009). An alternative test for the equality of variances for several populations when the underlying distributions are normal. Communications in Statistics: Simulation and Computation, 38, 109–117. doi:10.1080/03610910802431011

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, 69(346), 364–367. doi:10.2307/2285659

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research, 42(3), 237–288. doi:10.3102/00346543042003237

Lemeshko, B. Y., Lemeshko, S. B., & Gorbunova, A. A. (2010). Applications and power of criteria for testing the homogeneity of variances. Measurement Techniques, 53(3), 237–246. doi:10.1007/s11018–010–9489–7

Levene, H. (1960). Robust tests for the equality of variance. In I. Olkin (Ed.), Contributions to probability and statistics (pp. 278–292). Palo Alto, CA: Stanford University Press.

SAS. (n.d.). SAS/STAT(R) 9.22 User’s guide: Example 39.10 testing for equal group variances. Retrieved from https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/vie

Stefanie A. Wind Stefanie A. Wind Wind, Stefanie A.

Lexiles

Lexiles

972

973

Lexiles

A Lexile is a unit of measurement used to describe texts and readers on a common scale that represents reading comprehension. For texts, Lexile measures describe the difficulty of reading comprehension. For readers (e.g., students), Lexile measures describe reading comprehension ability. Together, Lexile measures for texts and readers make up the Lexile Framework. This entry provides an overview of the Lexile Scale, methods for obtaining Lexile measures, and current applications of the Lexile Framework.

The Lexile Framework was developed by MetaMetrics—an educational measurement and research organization based in Durham, NC. The framework is used internationally to measure student reading comprehension and text difficulty for a variety of instructional purposes, including matching readers with texts, predicting the degree to which readers will comprehend texts with known Lexile measures, and monitoring changes in reading comprehension over time. Accordingly, the topic is relevant for educational practitioners and researchers interested in the assessment of reading comprehension and in the selection or development of assessments targeted to specific levels of reading comprehension.

Units on the Lexile Scale are reported using whole numbers followed by the letter “L.” For example, a text might have a Lexile measure of 240 L. Lexile measures for texts and readers are calculated on the same linear scale that ranges from below 0 L to above 2,000 L. For texts, lower numbers indicate that less ability is needed to comprehend the text, and higher numbers indicate that more ability is needed to comprehend the text. For readers, lower numbers indicate lower reading comprehension ability and higher numbers indicate higher reading comprehension ability. Because both types of Lexile measures are on a common scale, measures for texts and readers can be compared to predict the degree to which a reader will be able to comprehend a given text.

For texts, Lexile measures are calculated using a proprietary formula developed by MetaMetrics that incorporates a variety of linguistic features, including vocabulary and sentence length. Lexile measures for texts are calculated using the Lexile Analyzer computer program. For readers, Lexile measures can be calculated using several methods. Specifically, a variety of commercially produced assessments exist that can be used to obtain a Lexile reader measure. Similarly, analyses can be conducted to determine a Lexile measure based on a reader's score on a norm-referenced or criterion-referenced assessment. Less formal methods can also be used to obtain Lexile measures for readers, including observations of students reading a text that has a known Lexile measure.

The Lexile Framework is applied internationally across a variety of contexts to describe the alignment between reading comprehension levels for texts and readers. These applications include instructional settings, test development, and the establishment of standards, including the Common Core State Standards in the United States. The Lexile Framework is available for texts in English and Spanish.

Stefanie A. Wind

See also Rasch Model; Readability; Reading Comprehension; Reading Comprehension Assessments; Scales

Further Readings

Blackburn, B. R. (2000). Best practices for using Lexiles. Popular Measurement, 3(1), 22–24.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). The Lexile Framework for reading technical report. Durham, NC: MetaMetrics.

License

See Certification

Lotta C. Larson Lotta C. Larson Larson, Lotta C.

Licensure

Licensure

973

974

Licensure

Licensure is the granting of a license to practice an act or profession that requires particular credentials. Supporters of licensure argue that regulating licensed professionals in the performance of an activity (such as performing surgery or teaching in public schools) protects and serves society at large. Critics claim that licensure provides an unfair advantage to licensed members of an occupation and reduces the available workforce. This entry discusses licensure as it relates to the teaching profession.

Obtaining a Teaching License

In the United States, teachers must meet licensing and/or certification requirements before they can teach. All 50 states require their teachers be licensed to teach in public schools. Particular licensure requirements are set by each state’s board of education. Requirements vary but commonly include obtaining a bachelor’s degree, completing a teacher preparation program, and passing standardized exams to demonstrate competency in subject matter and teaching-related skills.

In some states, teachers are required to receive certification to teach specific subjects or grade levels. Furthermore, new teachers are often expected to complete an extended student teaching experience, or internship, under the supervision of a licensed teacher. Most states also conduct criminal background checks and confirm citizenship status. Although licensure requirements are unique in each state, most engage in reciprocity agreements that recognize portions of a teaching license from other states, allowing educators a greater degree of flexibility in moving across state lines.

Types of Teaching Licenses

There are many different types of teaching licenses including early childhood, elementary education, middle level, or secondary education. Teachers can also obtain licensure or certification in a specialized field such as English as a second language, special education, or reading. Teachers new to the profession are often granted an initial or temporary licensure. By completing professional requirements (such as teaching full-time or completing additional course work), an initial license may lead to a professional license. To renew a professional teaching license, which is often valid for 5 or 10 years, most states require teachers to engage in continuing education (i.e., professional development or college courses), gain professional teaching experience, and pass a criminal background check.

States also offer alternative licensure programs. Those who enroll in these programs generally have decided to become teachers after graduating from college, sometimes after spending years in another career. They typically have a bachelor’s degree but have not taken the necessary education courses to pursue standard teacher certification. Supporters of alternative licensure programs argue that these programs can help offset shortages of teachers in certain subjects or qualify more people to teach in areas that have difficulty enticing and retaining teachers. Critics claim that alternative-certification initiatives are lowering teacher-quality standards.

Lotta C. Larson

See also Certification; Standardized Tests

Further Readings

Lowery, D. C., Roberts, J., & Roberts, J. (2012). Alternative route and traditionally-trained teachers’ perceptions of teaching preparation programs. Journal of Case Studies in Education, 3(1), 1–11.

National Association of State Boards of Education. (2016 January). State education governance matrix. Retrieved from http://www.nasbe.org/about-us/state-boards-of-education

Teach.com. (n.d.). Get your teaching credentials. Retrieved from http://teach.com/how-to-become-a-teacher/teaching-credential

TEACH.org. (n.d.). Understanding licensure and certification requirements. Retrieved from https://www.teach.org/teaching-certification

S. Jeanne Horst S. Jeanne Horst Horst, S. Jeanne

Elisabeth M. Pyburn Elisabeth M. Pyburn Pyburn, Elisabeth M.

Likert Scaling

Likert scaling

974

977

Likert Scaling

Likert scaling is a commonly used response scale format for measuring self-reported attitudes toward or beliefs about something (e.g., an idea or a product). Although statements, or “items,” assessing these attitudes or beliefs could be answered with a binary *yes/no* response scale, Likert-type scales allow for more varied levels of agreement. For example, a researcher may be interested in exploring participants’ attitudes about their own math competence. The researcher would ask participants to rate a set of related declarative statements, such as “I feel competent when completing multiplication problems,” using a rating scale typically composed of five to seven options: 1 = *disagree strongly*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, and 5 = *agree strongly*.

In this example, participants are asked to examine their feelings about their own competence and provide the rating that best reflects how they feel. Numeric scores from the participants’ ratings of the set of statements are then summed or averaged to produce a total score. For example, if the researcher asked participants to rate 5 items and a person rated each of the items a 4, then the Likert-scaled score would be 20 (4 + 4 + 4 + 4 + 4 = 20) if summed or 4 (20/5 = 4) if averaged. The use of Likert scaling is widespread throughout education, psychology, business, and other disciplines involving research examining people’s attitudes, values, beliefs, dispositions, or psychological states and traits. A few examples include the measurement of self-esteem, depression, academic motivation, and work-related attitudes.

In 1932, Rensis Likert first reported the method of Likert scaling, although Gardner Murphy is also credited with earlier work. Prior to the development of Likert scaling, a more arduous method, Thurstone scaling, was the most commonly used method of measuring attitudes. The first published examination of Likert-type scales involved items written to address racist, internationalistic, and imperialistic attitudes. The scores had high reliability and correlated with scores from other measures, supporting the use of Likert-scaled items. This entry discusses how Likert-type scales are constructed, how items are written, and the reliability and validity of Likert-type measures.

Constructing Likert-Type Scales

Although conceptually simple, Likert-type scales should be constructed while keeping several things in mind. Surveys employing Likert-type scales consist of items (i.e., most frequently declarative statements) to which respondents select a numeric response. On Likert’s original scale, respondents rated statements using the scale 1 = *strongly disapprove*, 2 = *disapprove*, 3 = *undecided*, 4 = *approve*, and 5 = *strongly approve*. However, there are multiple descriptors that can be used on Likert-type scales as long as the descriptions are equidistant on a quantitative scale. There is also debate over the optimal number of response options and whether or not a verbal descriptor is necessary for each number. For example, one variation of the Likert-type scale asks respondents to rate items on a scale of 1 = *strongly disagree* to 7 = *strongly agree*, with no descriptors between the end points. Research tends to indicate that full labeling may result in more reliable responses.

It is also important for the verbal descriptions to be evenly spaced, so that the conceptual distance between Options 1 and 2 is the same as the conceptual distance between Options 2 and 3, 3 and 4, and so forth. For example, it would be inappropriate to include a scale with the following response options: 1 = *somewhat disagree*, 2 = *sometimes disagree*, 3 = *neutral*, 4 = *occasionally agree*, and 5 = *always agree*. First, note that it may be difficult for respondents to choose between 1 = *somewhat disagree* and 2 = *sometimes disagree*. Second, note that the response options are not evenly spaced; the conceptual distance between *somewhat disagree* and *sometimes disagree* is likely smaller than the distance between *occasionally agree* and *always agree*.

One other consideration is whether to include an even or odd number of response options. If an even number of response options is provided, respondents are forced to choose one side or the other (e.g., *agree* or *disagree*)—there is no midpoint for people who are uncertain about their responses. Some Likert-type scale developers prefer an even number of response options, whereas others prefer an odd number of response options. The number of response options also has implications for the ways in which data from the scales are used in statistical analyses. When using data from items with five or fewer response options, the data can be considered categorical rather than continuous; this has implications for the choice of statistical method that can be used.

Another consideration when constructing Likert-type scales is the direction of the item wording relative to the response scale, which can take the form of "reverse-worded" and "negatively worded" items. When items are reverse-or negatively worded, the scale numbers need to be reversed when scoring. Consider the example item offered previously: "I feel competent when completing multiplication problems." A reverse-worded version might read "I feel incompetent when completing multiplication problems." When scoring this item, a researcher would need to reverse the scale to 5 = *strongly disagree*, 4 = *disagree*, 3 = *neutral*, 2 = *agree*, and 1 = *strongly agree*, so that the numeric score reflects the appropriate level of competency. If a respondent strongly agreed with the item, it means that respondent feels *incompetent* and the respondent response should accordingly be assigned a 1 to reflect this.

There are pros and cons to including reverse-worded and negatively worded items on a scale. There is consistent evidence that suggests people respond differently to negatively or reverse-worded items, which can introduce bias into the total score. On the other hand, including negatively worded items on a measure is helpful, as it can aid in identifying respondents who do not take the questions seriously. For example, if two negatively worded items were included on a 10-item measure, yet some people responded with 5 for every item, it may be an indication that those people were either not reading the items thoroughly or not taking the measure seriously.

Writing Items

In addition to considerations about Likert-type scale response options, there are factors to consider when writing items. Items employing Likert-type scales are intended to address values or desired behaviors rather than facts. For example, a

Likert-type scale item would typically not be something factual, such as “The mean is a measure of central tendency.” In contrast, a more appropriate use of Likert-type items focuses on values or attitudes, such as the statistics self-efficacy item, “I am confident that I can compute the mean of a set of numbers.”

Likert-type items should be written simply, concisely, and unambiguously. One recommendation is for the wording of items to be slightly below the reading level of the population for which the measure is intended. Another recommendation is to avoid the use of double-barreled items. An example of a double-barreled item is “The parking on campus is affordable and accessible.” A person who feels that parking is accessible, yet expensive, would not know how to best respond to this item nor would the researcher know how to interpret responses to this item. A better approach would be to offer 2 separate items: (1) “The parking on campus is affordable” and (2) “the parking on campus is accessible.” Similarly, some item wording may be unintentionally ambiguous. For example, the item “I frequently share my opinions with my friends” is ambiguous. Respondents could interpret the word “share” as (a) having the same opinions as their friends or (b) telling their opinions to their friends.

In addition to avoiding double-barreled and ambiguously worded items, it is best to avoid items that include double negatives. For example, the item “Students should not avoid using counseling services if they need them” would be confusing and cognitively taxing for many participants. Perhaps better wording might be, “When appropriate, students should seek counseling services.” Items that are clear, concise, and straightforward are particularly important for certain populations, such as children or respondents who are not native speakers of the language.

Social desirability is also an important consideration when writing Likert-type items. For example, people may respond positively to the item “It is important to be welcoming to people of all nationalities” simply because they realize that it is socially appropriate to do so. For this reason, it is important to pilot test items on a representative sample of respondents and to explore qualitative approaches, such as focus groups. One clue that socially desirable responding may have occurred is when data points primarily fall at one end of the scale. For example, perhaps only 1% of respondents selected option 1 = *strongly disagree* and 99% selected 5 = *strongly agree* for the item “I desire close personal relationships.”

Reliability and Validity

Given that Likert-type measures typically address attitudes or values, which are not tangible—they can’t be seen or touched—it is crucial to examine both the reliability and validity of inferences drawn from the scores. Internal consistency of scores, such as coefficient α, is a measure of the extent to which respondents used the scale consistently and the extent to which scores consist of random measurement error. When measures are used at multiple time points, it is important to consider whether the scores are consistent across those time points; this is referred to as test–retest reliability. If multiple forms of the measure are used, it is also important to evaluate whether scores across the forms are consistent.

In addition to evaluating the reliability (consistency) of Likert-type scale scores, it is also important to evaluate the validity of inferences that are drawn from those scores. That is, researchers should gather a body of evidence that supports or refutes the interpretations of the scores. When evaluating validity, researchers frequently investigate whether the scores from the measure correlate with scores from other measures hypothesized to be related to (or not related to) the measure of interest. It is also important to investigate whether the scores from the measure distinguish between groups as would be expected. For example, when examining the validity of scores from a depression measure, a researcher may hypothesize that people seeking treatment for depression should score higher on the measure than the general population. Evidence supporting this hypothesis would serve as an indication that the scores actually represent respondents’ levels of depression.

When gathering validity evidence, researchers also want to consider the dimensionality of scores. For example, a researcher may develop a Likert-type measure of test anxiety for use in a college setting that includes the items “My palms feel sweaty when I take a test” or “I do not enjoy taking tests.” Although these items may both address test anxiety, it is feasible that the 2 items represent two different dimensions of test anxiety: (1) physical response and (2) attitudes toward tests. Researchers typically use factor analysis to examine the dimensionality of scores from Likert-type measures.

S. Jeanne Horst and Elisabeth M. Pyburn

See also Internal Consistency; Reliability; Scales; Self-Report Inventories; Social Desirability; Surveys; Thurstone Scaling; Validity

Further Readings

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Arnold, W. E., McCroskey, J. C., & Prichard, S. V. O. (1967). The Likert-type scale. Research Methods in Speech, IV, 15.

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education, 38, 1212–1218.

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. Journal of Psychoeducational Assessment, 32, 663–673.

Likert, R. (1932). A technique for the measurement of attitudes. In R. S. Woodworth (Ed.), Archives of Psychology: Vol. 22 (pp. 5–55). New York.

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point Likert-type response scales. Educational and Psychological Measurement, 47, 27–35.

Dmitriy Poznyak Dmitriy Poznyak Poznyak, Dmitriy

Lisrel

Lisrel

977

980

Lisrel

LISREL—an abbreviation of linear structural relationships—is a statistical software package primarily dedicated to estimating structural equation models (SEMs) although it can handle a variety of other statistical models. SEM unifies several estimation methods into one analytic framework. The methodology has received considerable attention in education research where it has been used to develop and validate measurement instruments, estimate the relationship between students' development over time and various outcomes of interest, and assess the simultaneous independent effect of a program on students' achievements.

LISREL's target uses are students, applied researchers, and practitioners interested in SEM. It is commonly used in the areas of education, psychology, and other social science disciplines. This entry discusses the development of LISREL and then provides an overview of the LISREL statistical package and its modeling capabilities.

Designed in the early 1970s by Karl Jöreskog and Dag Sörbom, LISREL was the first dedicated software package developed for SEM. Since then, it has set the standard in SEM software and served as a prototype for many other SEM programs, such as Amos, EQS, and Mplus. LISREL's name, notation, and modeling approach have become synonymous with SEM methodology itself. Indeed, SEMs are often referred to as LISREL models, and LISREL notation using Greek characters is often used to specify SEMs in a text.

The LISREL model bridged two distinct statistical traditions: psychometrics and econometrics. The psychometric tradition is reflected in the *measurement* component of SEM, in which a matrix of observed variables are used to estimate a set of latent, or unobserved, variables. These latent variables can also be referred to as factors. Exploratory factor analysis and confirmatory factor analysis are two common measurement techniques in SEM. The econometric tradition is manifested in the *structural* component of SEM, in which simultaneous regression equations are used to test hypothesized relationships between a set of latent and/or observed variables. Observed variables can also be referred to as items, indicators, or manifest variables. A researcher then determines how tenable the estimated relationships are, given the observed correlations or covariances between the variables.

The most general SEM is defined by three matrix equations specified using LISREL notation: the measurement model for the latent exogenous variables $x = \wedge x\eta + \delta$, the measurement model for the latent endogenous variables $y = \wedge y\eta + \varepsilon$, and the structural model $\eta = B\eta + \Gamma\xi + \zeta$, where Λx, Λy B, and Γ are coefficient matrices and δ, ε, and ζ are vectors of latent variables.

The version of the LISREL software package released in 2015 is 9.20. The package is distributed as a 32-bit application for Windows computers. The core of the package is written in FORTRAN, but its vendor—Scientific Software International Inc. (Skokie, IL)—developed a visual interface in C/C++.

Applications

Originally, LISREL was developed as a dedicated program for estimating SEMs, but it is no longer limited to just SEM. LISREL 9.20 is packaged as a suite of five advanced statistical programs for multivariate analysis. The first program, LISREL, provides the ability to estimate both standard and multilevel SEM. The second program, PRELIS, was developed as a preprocessor for LISREL. It can be used for data import, preparing correlation and covariance matrices to be read by LISREL, data manipulation and transformation, data imputation, conducting basic statistical tests, and multivariate statistical analyses. The third program, MULTILEV, fits multilevel linear and nonlinear models to hierarchical data using both continuous and categorical outcome variables. The fourth program, SURVEYGLIM, extends LISREL’s capabilities to fit generalized linear models to data from simple random and complex surveys. This program supports a range of sampling distributions, including Gaussian, inverse Gaussian, multinomial, binomial, negative binomial, Bernoulli, Poisson, and γ. Finally, the fifth program, MAPGLIM, can be used to fit GLM models to multilevel data. Each of these programs can function together or independently.

To estimate a model with missing data, LISREL by default uses full information maximum likelihood. Users may also opt to impute missing values with either expectation maximization or Markov chain Monte Carlo algorithms. Beginning with Version 9.10, LISREL provides robust estimation of standard errors and X^2 statistics if users input a raw data file. The default estimation method in LISREL is maximum likelihood, but users may override this default when setting up their model.

In sum, the LISREL 9.20 statistical package makes it possible to estimate a wide range of statistical models used in social science research. These models include exploratory factor analysis and confirmatory factor analysis (CFA) with continuous and ordinal variables, multiple-group analysis, multilevel linear and nonlinear models, latent growth curve models, and GLMs with simple and complex survey data.

LISREL Interface, Data Format, and Model Specification

In the most general form, analyzing data in LISREL involves three files: a raw data file to be read into LISREL; a Syntax file that contains the commands used to read the data, estimate the model, and produce necessary output; and an Output file that contains the model results. Only raw data in ASCII (.dat) format can be read directly by LISREL. By default, the data will be read in free format (rows are separated by blanks). All other formats need to be imported to PRELIS, which will then convert the data to a covariance matrix format. Covariance and correlation matrices may be imported into the program in a .txt format.

There are several different ways to generate a Syntax file in LISREL. The first approach is to specify the model using LISREL command language. This can be done by either writing a text file with LISREL commands (*.ls8) or by generating a text file with LISREL commands using a program menu (*.lpj). LISREL command language is based on key words where only the first two characters will be recognized by the program.

The structure of a LISREL Syntax file will depend on the type of data to be

The structure of a LISREL Syntax file will depend on the type of data to be processed and the type of model to be estimated. If the data are imported as a PRELIS file (.psf), the LISREL Syntax file for an SEM will have the following general structure:

```
TITLE <string>
RA=<Raw data from file name>.psf
MODEL <model specification>
LK <label for an independent latent variable>
LE <label for a dependent latent variable>
PATH DIAGRAM <specified without further options>
OUTPUT <output specifications>
```

If the data are inputted as a covariance or correlation matrix, additional commands are required to create a Syntax file. These commands are necessary to input the matrices, specify their type, record the number of observations, and define observed variables.

LISREL’s introduction of a dialogue box that facilitates writing commands has made model specification much easier compared to previous versions of the program. However, creating a Syntax file still requires a basic understanding of matrix algebra and the matrices involved in model specification.

The second way to create a Syntax file is to use SIMPLIS command language. In 1993, SIMPLIS language was introduced to simplify the coding of LISREL input files. Users can specify and estimate an SEM by writing a text file with SIMPLIS commands (*.spl) or by generating a text file using the menu (*.spj). This approach requires users to specify the data file, the names of the observed and latent variables, and the estimated regression paths. These paths can be further specified as symbolic relations between the variables, so that matrix specification is no longer necessary.

Finally, the third way to generate a Syntax file is to utilize SIMPLIS or LISREL commands through the graphical interface. This involves simply drawing a path diagram (*.pth) on the screen for which the program identifies regression paths to be estimated. This approach to model specification may be particularly appealing to researchers who are new to the program.

LISREL Output

The LISREL Syntax file is used to produce an Output file (*.out). The Output file consists of several parts and provides information about model results and the fit of the model. The structure of the Output file varies depending on the type of model being estimated and the options specified in the Output command.

For a basic SEM with both observed and latent variables, the LISREL Output file will contain the following results: a copy of the Syntax, a detailed specification of all model parameters, observed and fitted covariance matrices between the variables in the model, unstandardized and standardized parameter estimates with *t* values and standard errors, squared multiple correlations for structural equations (R^2), goodness-of-fit statistics, and finally modification indices. If there are problems with the model, warnings and errors appear in the appropriate sections.

The Output section called “goodness-of-fit statistics” indicates how well the hypothesized model fits the data. For a basic SEM like the one described earlier, LISREL provides relative fit indices to compare the fit of nested models. These indices include the log likelihood, X^2, and Akaike information criteria. LISREL also provides absolute fit statistics to compare the fit of the model to an established cutoff criteria. These indices include the root mean square error of approximation, comparative fit index, goodness-of-fit index, adjusted goodness-of-fit index, normed fit index, root mean residual, and expected cross-validation index.

Finally, the LISREL Output file provides model modification indices to highlight potential issues with the model and changes that could improve its fit to the data.

Evaluation of a Statistical Model

Statistical adequacy of the LISREL model can be determined by evaluating the global-and local fit statistics provided by the program. The so-called *global fit statistics* show how well the overall model fits the data. For factor SEM in LISREL, global fit statistics include the relative and absolute fit indices described earlier. The *local fit indices* indicate whether a model yields reasonable point estimates and standard errors. For instance, excessively large or small standard errors or negative error variances may suggest issues with model fit. For a variety of models, LISREL also provides *model modification indices* that need to be reviewed to improve statistical fit of the model.

Availability: Download and Materials

LISREL 9.20 can be purchased or rented for 6 or 12 months from Scientific Software International’s website. A discount is available for those purchasing multiple licenses. A free student edition is also available, but it is limited to 12 variables. All purchased licenses are permanent and require no maintenance or renewal fees. The vendor will respond to brief technical questions received via e-mail from end users with active licenses. Manuals for LISREL and accompanying statistical programs are available for download from its website.

Dmitriy Poznyak

See also Amos; Confirmatory Factor Analysis; EQS; Exploratory Factor Analysis; Mplus; Path Analysis; Structural Equation Modeling

Further Readings

Jöreskog, K. G., & Sörbom, D. (2015). LISREL 9.20 for Windows [Computer software]. Skokie, IL: Scientific Software International.

Teo, T., & Khine, M. S. (Eds.). (2009). Structural equation modeling in educational research: Concepts and applications. Rotterdam, the Netherlands: Sense.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), Handbook of structural equation modeling (pp. 209–246). New York, NY: Guilford Press.

Websites

Scientific Software International, Inc.: http://www.ssicentral.com/

Lotta C. Larson Lotta C. Larson Larson, Lotta C.

Literacy

Literacy

980

983

Literacy

The term *literacy* is commonly defined as the ability to read, write, and understand print language. In reality, literacy is more encompassing than such a simple definition. The International Literacy Association (n.d., n.p.) views literacy as “the ability to identify, understand, interpret, create, compute, and communicate using visual, audible, and digital materials across disciplines and in any context.” Literacy is a complex, multifaceted process that requires a wide variety of skills and approaches. This entry first discusses the evolution in how literacy is viewed and the different aspects of literacy and then describes effective literary instruction and the stages of literacy development.

Although there is no single definition of literacy, it is commonly acknowledged that literacy changes as society and technology evolve. In the 1950s, a person who could both read and write a short statement was considered literate. Twenty years later, an emphasis was placed on functionality and specific social context. In other words, literate persons were expected to successfully engage in activities in which literacy is required for effective functioning within their social group or community.

In the early 2000s, the concept of functional literacy involved “varying contexts,” which recognize that literate people need different literacy skills depending on varying situations. In addition, it was generally accepted that literacy entails a “continuum of learning” because people constantly develop literacy skills. At this time, literacy was also recognized as a social practice, meaning that literacy seldom develops or takes place in isolation. Today, the meaning of literacy has further evolved to reflect the increased use of information and communication technologies, along with diverse practices in information and communication technologies, along with diverse practices in relation to political, socioeconomic, cultural, and linguistic circumstances and contexts.

People who are not functionally literate, often due to lack of education, are considered *illiterate*. Illiteracy rates are highest among developing countries. However, in certain regions or communities in the United States, illiteracy is a concern, often due to economic factors. Children raised in homes with at least one illiterate parent are more likely to be illiterate.

A literate person who has the skills to read and write, but chooses not do so, is considered *aliterate*. Although illiteracy is primarily a problem in the developing world, aliteracy is a growing trend in developed countries, including the United States. Lack of motivation is often cited as a reason for aliteracy. It is important that educators understand the difference between illiteracy and aliteracy to provide students with appropriate literacy instruction or intervention.

Aspects of Literacy

Literacy is a collection of communicative and cultural practices shared among and within societal groups and contexts. To be considered literate, students need to develop skills in multiple areas of language arts, including reading, writing, listening, speaking, viewing, and visual representation to effectively receive and communicate information. Receptive language consists of messages taken in, received, comprehended, or interpreted through processes of reading, listening, or viewing. Communicative language involves messages formulated, symbolized, constructed, and relayed to others through writing, speaking, or visual representation.

All areas of literacy are interrelated and cannot effectively be taught in isolation. In almost all instructional scenarios, the language arts blend together as interactive communication and reception in the larger context of authentic learning. Although not an exhaustive list, what follows is a brief overview of critical areas of literacy.

Reading

Reading includes the skills, strategies, and meaning making required to make sense of printed text. Students engage in the reading process to develop skills in phonemic awareness, phonics, fluency, vocabulary, and comprehension. Phonemic awareness is the understanding that words are composed of sounds. Phonics instruction helps students learn the relationship between letters, or groups of letters, of the written language and sounds of the spoken language. Fluency is the ability to read words and text passages smoothly and correctly with understanding. Vocabulary knowledge is critical to reading development and goes beyond decoding. Vocabulary is understanding the meaning of a word. Comprehension is the ultimate goal of reading instruction and involves the process of constructing meaning through interaction with a text.

Writing

Writing describes the process of recording language graphically to communicate or preserve ideas through print while using conventional spelling, grammar, and punctuation to express those ideas. Students engage in the writing process to write narrative texts, informative texts, and arguments to clearly convey ideas, communicate information, and support opinions or viewpoints.

Listening

Listening includes the processing and interpretation of sounds and language for information and pleasure. Students apply and integrate information presented while evaluating information, reasoning, and a speaker’s point of view.

Speaking

Speaking involves the transmission of a message through the use of oral language. To speak effectively requires the knowledge of words, articulation of sounds, and the ability to structure the message in a decodable fashion for the audience. Students need to be able to adapt speech to a variety of contexts, tasks, and audiences and be able to use digital media and visual displays to enhance oral presentations.

Viewing

Viewing includes the visual observation and interpretation that results in meaning making. This ability has been defined as being able to “interpret, recognize, appreciate, and understand information presented through visible actions, objects, and symbols" (Institute of Museum and Library Services, n.d., n.p.). This can include the ability to understand information presented in illustrations, multimedia, and figures.

Visual Representation

Visual representation involves the transmission of a student's thought or message to be viewed by another person. Students communicate thoughts, messages, and interpretations of information through visual media, symbols, graphics, and art.

Technological Skills

An additional aspect of literacy involves proficiency with a wide variety of technologies; the ability to collaborate and share information in global communities for a variety of purposes; and the ability to create, analyze, and evaluate multimedia texts. These skills are often referred to as *new literacies, digital literacies*, or *multimodal literacies*. Although a singular definition is not possible, most educators and researchers agree that these literacies are multiple, dynamic, and malleable. They change over time and in response to transformations in information and communication technologies, digital and print texts, and global contexts.

Effective Literacy Instruction

Effective literacy instruction requires that teachers model the skill, strategy, or process for students and allow students time to practice and apply what has been taught. Because all students are unique individuals who develop literacy in various stages (see below), teachers must differentiate literacy instruction to meet individual needs of all learners. Research indicates that effective literacy instruction challenges all students and holds them to high standards. Furthermore, effective literacy instruction involves creating quality lessons that are based on standards and are designed with clarity and purpose. In the United States, each state generally has its own set of literacy or language arts standards. Most states' standards are aligned with the Common Core State Standards for English language arts/literacy.

Comprehensive balanced literacy instruction involves both teacher-directed instruction and student-centered activities. Teacher-directed instruction involves explicitly and systematically modeling how to use a strategy, skill, or process. In student-centered instruction, students engage in an assigned task or activity from which they are expected to develop certain skills. Some students learn best through direct instruction, whereas others benefit from a student-centered approach. Hence, in balanced literacy classrooms, teachers integrate explicit instruction with authentic literacy experiences.

Literacy Development

Children and young adults develop literacy in various stages. There is much overlap between the stages and children do not completely finish one stage and then move to the next. There are various names for these stages and some disagreement over the approximate age ranges associated with them. The following examples use the names that are commonly applied in schools.

Early Emergent Literacy (Birth to Age 3)

Generally, children develop the foundations of literacy before they enter school. During this stage, they develop oral language, scribble and imitate writing, handle books, and enjoy being read to.

Emergent Literacy (Ages 2–5)

During this stage, children begin to show greater interest in literacy. They develop concepts about print, begin to recognize that print and images carry meaning, and use more standard oral language. Most children remain in this stage until kindergarten or the beginning of first grade.

Beginning Reading and Writing (Ages 4–8)

Beginning readers and writers are developing phonological awareness and are able to hear, count, and manipulate sounds and syllables within words. As their oral language expands, they also begin to read and write in conventional ways. They figure out how to pronounce words and develop fluency. This stage usually lasts through first, second, or even third grade.

Almost Fluent Reading and Writing (Ages 7–11)

During this stage, children grow more sophisticated in all aspects of literacy. In particular, they develop greater vocabulary, fluency, and reading comprehension; written work becomes more sophisticated. For most children, this stage lasts between second grade and fourth or fifth grade.

Fluent Reading and Writing (Ages 10 and Up)

Here, students use reading, writing, and oral language for a variety of purposes. They continue to build reading fluency and word identification strategies. Written work becomes more organized and coherent across multiple genres. For some students, this stage begins in fourth grade and continues into the upper elementary grades and into middle school and high school.

Lotta C. Larson

See also Common Core State Standards; Learning Styles; Reading Comprehension; State Standards

Further Readings

Cooper, J. D., Robinson, M. D., Slansky, J. A., & Kiger, N. D. (2015). Literacy: Helping students construct meaning (9th ed.). Stamford, CT: Cengage Learning.

Gambrell, L. B., & Morrow, L. M. (2015). Best practices in literacy instruction (5th ed.). New York, NY: Guilford Press.

Institute of Museum and Library Services. (n.d.). Museums, libraries, and 21st century skills: Definitions. Retrieved from https://www.imls.gov/impact-imls/national-initiatives/museums-libraries-and-21st-century-skills/museums-libraries-and-21st-century-skills-definitions

International Literacy Association. (n.d.). Why literacy? Retrieved from http://www.literacyworldwide.org/why-literacy

International Reading Association. (2009). New literacies and 21st-century technologies: A position statement of the International Reading Association. Retrieved from http://www.literacyworldwide.org/docs/default-source/where-we-stand/new-literacies-21st-century-position-statement.pdf?sfvrsn=6

National Council of Teachers of English. (2013). NCTE position statement. The NCTE definition of 21st century literacies. Retrieved from http://www.ncte.org/positions/statements/21stcentdefinition (Original work published 2008) National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). Common Core State Standards for the English Language Arts. Retrieved from http://www.corestandards.org/ELA-Literacy/

National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Robinson, R. D., McKenna, M., & Conradi, K. (2012). Issues and trends in literacy education (5th ed.). Upper Saddle River, NJ: Pearson.

United Nations Educational, Scientific and Cultural Organization. (2004). The plurality of literacy and its implications for policies and programmes (Position paper). Paris, France: Author.

Jackie Waterfield Jackie Waterfield Waterfield, Jackie

Literature Review

Literature review

983

987

Literature Review

The term *literature review* can be viewed as both what is read and the process that has been undertaken to produce the work in question. Broadly, it denotes the bringing together and summary or synthesis of previous published work. In the academic literature, the terms *review, literature review, descriptive review, systematic review*, and *narrative review* are often used. At times, such terminology denotes specific distinctions between different types of review, but on other occasions some of the terms are used interchangeably.

The term *literature review* covers a range of approaches. This entry first discusses literature reviews in education. It then describes narrative reviews and systematic reviews, before explaining two approaches commonly used in systematic reviews, meta-analysis, and metasynthesis.

Literature Reviews in Education

Literature reviews are seen in academic writing, both in students' essays and theses and in professional peer-reviewed journals. The purpose in the former is most commonly either as a very circumscribed piece of written work for an assessment or to create a rationale for an individual's research. In the latter, it may well be a resource to influence policy, to guide practice, or to identify gaps in knowledge and influence future research.

In an educational thesis, students demonstrate through the literature review their knowledge and understanding of the research and theories in their field of study and, by critiquing others' work, position themselves and the focus of their research. The review therefore gives both a theoretical framework and a methodological rationale for the student's research. These reviews are of the traditional type. There is a logic in how this type of review is undertaken and, depending on the level of study, the expectation that the review will demonstrate a certain level of synthesis and theoretical thinking. These literature reviews can be criticized for being too broad, and the process of gathering information often lacks an explicit structure; this makes them susceptible to personal bias, both in what is included and how it is interpreted. Yet the literature reviews serve a specific function in an individual's learning.

In peer-reviewed publications, a traditional literature review is presented as a way of synthesizing more than one piece of research or theoretical work and is often a means of bringing the reader up to date with current thinking on the subject or a vehicle for influencing policy and/or practice. It usually seeks to demonstrate a more standardized and transparent process than in a piece of student work; this is part of the move to evidence-based policy and practice. In addition to peer-reviewed journals, the literature reviews in education are published by organizations such as the Campbell Collaboration. The important aspect of any such review is that, in terms of both its conception and execution, it is the most appropriate for the purpose. This normally requires a review to meet the following criteria:

- *Comprehensive*—the main sources relevant to the topic or question should be included.
- *Relevant*—as well as being comprehensive, the review should at the same time be discriminating and exclude sources that have little or no direct bearing on the topic.
- *Up-to-date*—sources should represent contemporary thinking or research in the area concerned, though it is important to note that some literature written many years ago retains its relevance to the present day.
- *Unbiased*—sources should not be included in a tendentious way so as to advance one particular viewpoint to the exclusion of others.

In the light of these requirements, in published reviews, there has increasingly been a move from traditional reviews to more systematic reviews, including such features as inclusion/exclusion criteria, an explicit and reproducible search strategy, specific means of assessing the quality of included items, and clear mechanisms to reduce bias. The following two forms of review reflect this move.

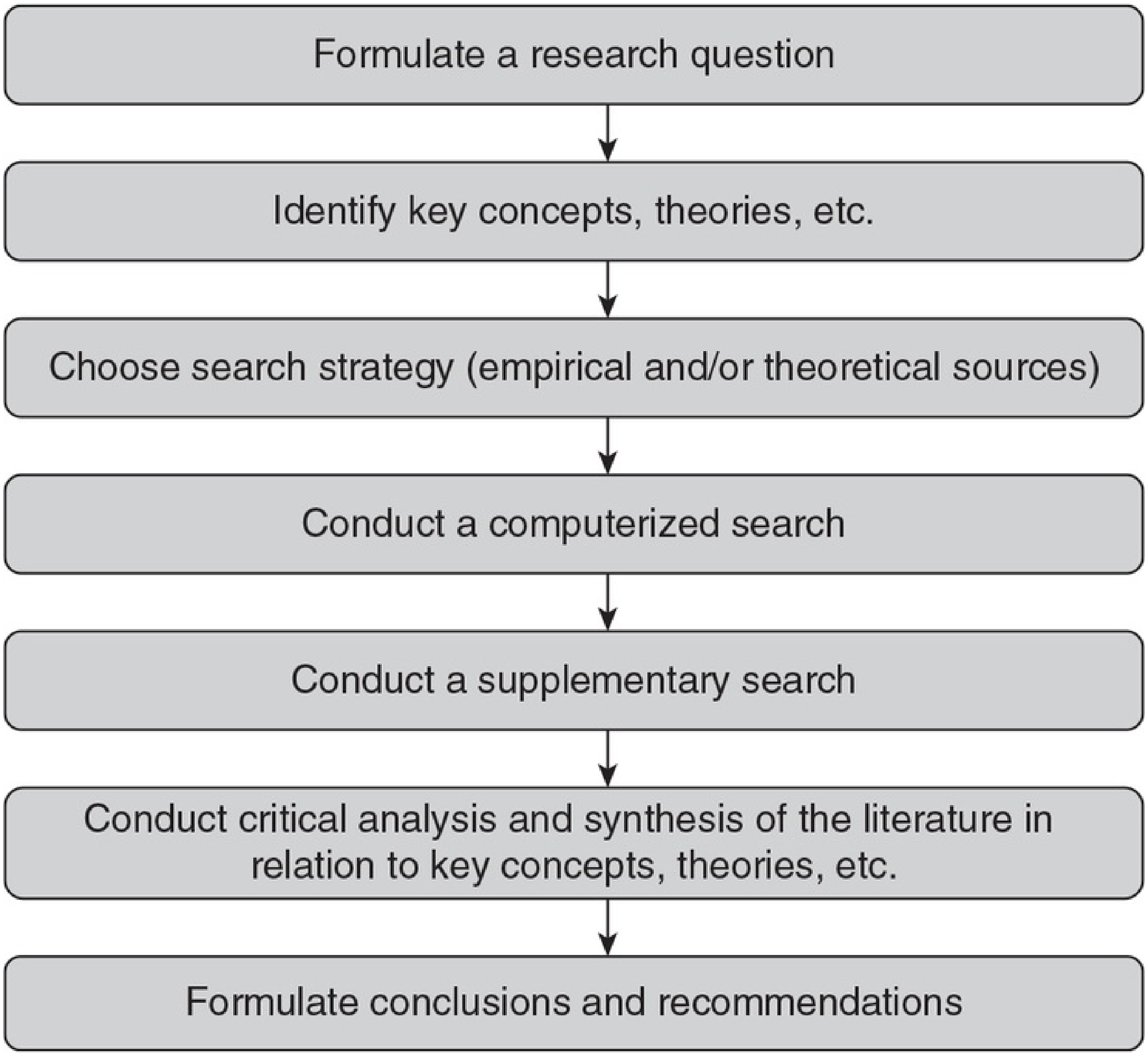
Narrative Reviews

Although traditional reviews have also been labeled as narrative reviews, what we now understand as a narrative review differs from the traditional review. More specifically, the purpose of, and the processes undertaken within, narrative reviews has changed, with a resulting increase in their quality.

The purpose of a narrative review can be considered different from that of a systematic review. Narrative reviews can fulfill several purposes, such as: to describe the current state of both art and science (theory and practice); add dimensions of insight or application that are not available in existing literature; and provide critical evaluation of accepted theory or practice, such as the use of learning records in undergraduate education. Frequently, narrative reviews do not seek to answer a single focused question, such as determining what the *best* intervention is in a practical situation, but they nonetheless have a clear focus on a research question and the results are presented in the form of one or more propositions, which may lead to new theories or research or summarize current practice, often presenting controversies and emerging issues that may not have presented themselves in individual works.

Increasingly, narrative reviews present an approach to the finding and interpretation of literature that is more systematic than that of the traditional review (hence the term *systematic literature review* is sometimes used, possibly to avoid the stigma sometimes attached to the traditional review). They have a much more focused question, and they demonstrate a search strategy, including clarity over the inclusion and exclusion of literature, and a process by which findings are analyzed and pulled together. These later stages of the review process often mirror what happens in empirical studies that collect qualitative data. The literature used may be drawn from empirical research, theoretical papers, electronic sources such as websites, and media articles; much will depend on the focus of the review. As such, the process is replicable, but the theoretical propositions and inferences drawn from the data may be considered more subjective than those of systematic reviews. Typically, the stages shown in Figure 1 are followed in a narrative review.

Figure 1 Steps in a narrative review


Formulate a research question
Identify key concepts, theories, etc.
Choose search strategy (empirical and/or theoretical sources)
Conduct a computerized search
Conduct a supplementary search
Conduct critical analysis and synthesis of the literature in relation to key concepts, theories, etc.
Formulate conclusions and recommendations

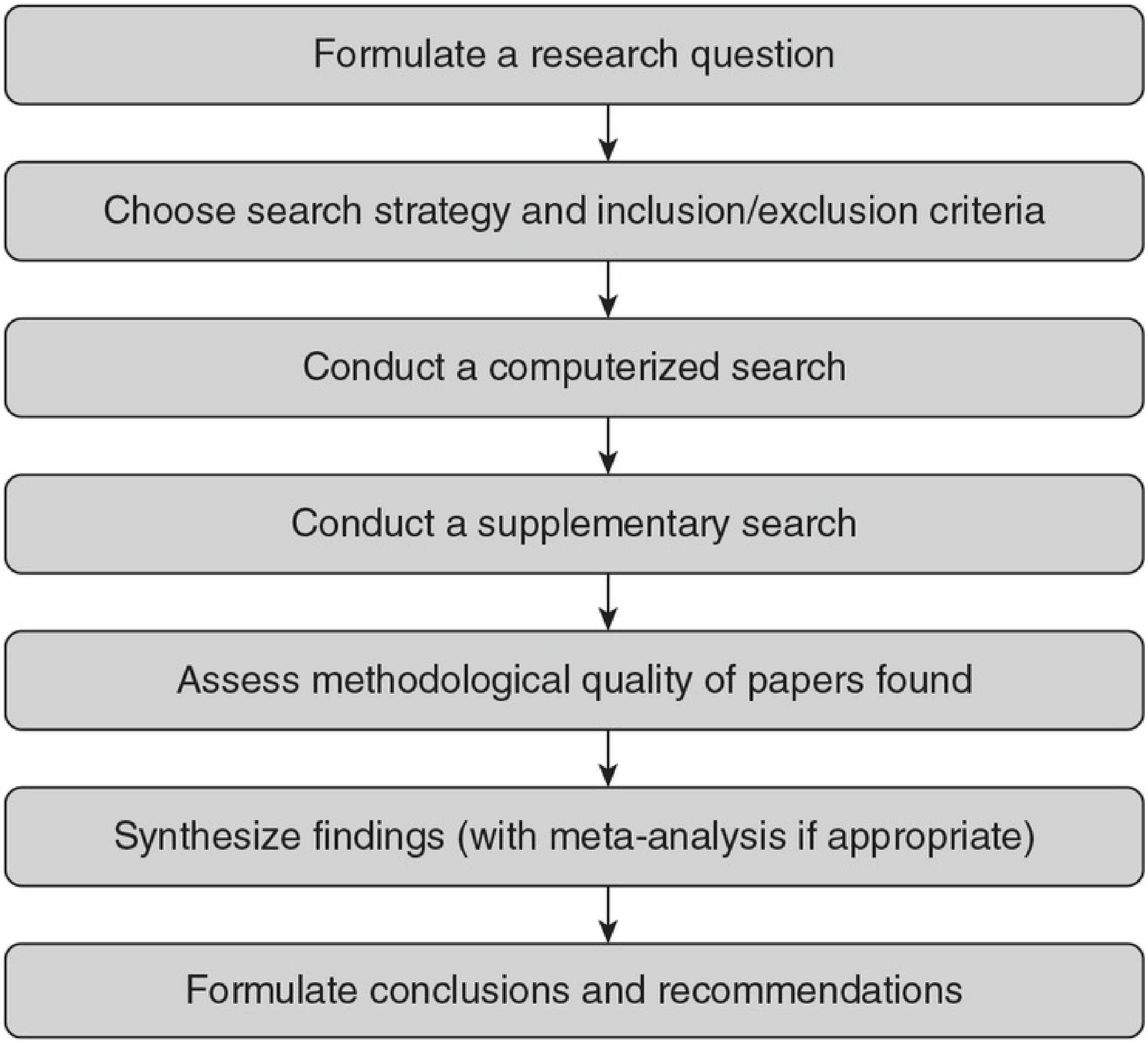
Systematic Reviews

Although the narrative review follows a systematic process, the systematic review, as is reflected in its name, lays an even greater emphasis on explicit, structured, and reproducible processes, which are principally aimed at maximizing the criteria for reviews outlined earlier.

Education has a history of combining educational experiments within systematic reviews, and many such reviews to be found in the discipline look at the question as to whether one method of learning or educational management is more effective than another. Although experimental studies seem most commonly to be associated with systematic reviews, other forms of empirical research can be used (e.g., cross-sectional studies looking at statistical associations or longitudinal panel studies). The type of study included very much depends on the question the reviewers are trying to answer. This is an important issue, as it will affect the search strategy—in terms of inclusion and exclusion criteria, databases, and literature generally—and the way the results from the search are assessed for quality.

This form of review shares many of the same process characteristics as the narrative review. The typical stages in a systematic review, shown in Figure 2, are similar to those illustrated earlier in Figure 1 for a narrative review. An important difference is that a systematic review will tend to focus on a very specific question and will rely solely on empirical studies (usually quantitative) to answer this question. In addition, conscious efforts are made to locate relevant studies, whether published or not, owing to the tendency for the findings of published papers to differ from those of unpublished papers (including those never submitted for publication in the first instance) or papers published in less accessible sources; this is referred to as the file drawer problem.

Figure 2 Steps in a systematic review


Formulate a research question
Choose search strategy and inclusion/exclusion criteria
Conduct a computerized search
Conduct a supplementary search
Assess methodological quality of papers found
Synthesize findings (with meta-analysis if appropriate)
Formulate conclusions and recommendations

Another distinctive feature of systematic reviews is that they normally use a specific methodological tool to appraise the studies included. These tools look at whether certain aspects of design and methods can be found in the write-up of the research. However, they are different from tools designed simply to assist critical appraisal, as they have specifically been formulated to attach a numerical value on the basis of the fulfillment of specific criteria. So, the paper can be given a total score, or the score ascribed to a specific aspect, such as inclusion criteria or sample size, can be examined. Sometimes, the tool has a threshold score, whereby the reader is advised that those papers whose score falls below this threshold should be considered methodologically poor and the findings within them should be viewed with caution. Recently, there has been a move away from the numerical scoring of studies to an approach that creates a profile of those specific aspects of methodology that make a study prone to bias.

Within a systematic review, the findings from individual studies are presented, often in numerical form, and summary conclusions are drawn across these studies (or sometimes across homogeneous categories of studies). The way in which the findings in a systematic review are synthesized is often referred to as narrative synthesis, to distinguish it from the statistical aggregation of data that occurs in a meta-analysis. It is important to note, however, that this is a rather different use of the word *narrative* than in the term *narrative review*.

Meta-Analysis and Metasynthesis

Systematic reviews may include a meta-analysis or metasynthesis as ways of bringing empirical findings—quantitative and qualitative, respectively—together so as to increase their explanatory or interpretive power.

Meta-Analysis

A meta-analysis often occurs within the context of a systematic review and seeks to pool the numerical estimates from these studies into a single summary estimate; it is normally undertaken on the *summary* data from the papers (e.g., mean, standard deviation, and sample size in each group). In a meta-analysis, the outcome from each study should relate to the same variable (e.g., a particular outcome in educational testing or perhaps a particular category of such outcomes) and should normally be in the same form—so, all studies should generate a mean difference or they should all generate an odds ratio. Under these circumstances, a summary estimate can be calculated across all of the studies. The estimates from each study and the overall (or pooled) estimate can be listed —and normally presented graphically—along with their 95% confidence intervals. The overall estimate not only summarizes the individual studies but also provides a more precise estimate of the effect or relationship of interest as it is based on the combined sample sizes of the included studies.

Metasynthesis

Increasingly, researchers in the qualitative tradition have been looking at ways of synthesizing the findings of papers reporting qualitative data. There is as yet no real agreement on the “best” way to synthesize qualitative findings, and indeed, there is a range of terms used to describe this general process (for example, *metasynthesis, meta-ethnography, metatheory*, and *metastudy*). Metasyntheses offer the opportunity to draw theoretical inferences from several related research reports by synthesizing the findings presented within these studies. This takes the form of an interpretive integration of the insights from individual studies, in the context of an existing or developing body of theory. Metasyntheses often include a methodological evaluation of the included papers, but as there is less consensus on methodological criteria for qualitative than for quantitative research, this evaluation tends to be more tentative than in a quantitative systematic review or meta-analysis.

Jackie Waterfield

See also Abstracts; APA Format; File Drawer Problem; Journal Articles; Meta-Analysis; Methods Section; Results Section

Further Readings

Aveyard, H. (2014). Doing a literature review in health and social care: A practical guide (3rd ed.). Maidenhead, UK: Open University Press.

Cooper, H. M. (2009). Research synthesis and meta-analysis: A step-by-step approach (4th ed.). Thousand Oaks, CA: Sage.

Fink, A. (2013). Conducting research literature reviews: From the Internet to paper (4th ed.). Thousand Oaks, CA: Sage.

Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Oxford, UK: Blackwell.

Ridley, D. (2012). The literature review: A step-by-step guide for students (2nd ed.). Thousand Oaks, CA: Sage.

Sandelowski, M., & Barroso, J. (2007). Handbook for synthesizing qualitative research. New York, NY: Springer.

Torgerson, C. (2003). Systematic reviews. London, UK: Continuum.

Websites

Campbell Collaboration: http://www.campbellcollaboration.org

Yi-Hsin Chen Yi-Hsin Chen Chen, Yi-Hsin

Local Independence Local independence

987

988

Local Independence

Local independence or *local item independence* is an important assumption for latent variable models such as latent class models, factor analytical models, and item response theory (IRT) models. The basic concept of local independence is that the response to an item (or a question) is independent of that to any other items conditional on the latent variable(s) being measured. Thus, local independence is also known as *conditional independence*. For instance, a mathematics achievement test is purported to measure a general mathematical ability. After removing the measured general mathematical ability, there are no relationships between any pairs of test items, indicating that the mathematics test items meet the assumption of local independence. If local independence is not met by test items, these items are considered local dependence or local item dependence (LID). This entry elaborates the basic concepts and importance of local independence as well as the potential sources of LID and some popular statistical methods for testing LID.

Basic Concepts and Importance of Local Independence

Classical true score theory assumes that the observed scores are equal to the true scores plus the error scores (O = T + E). Local independence in classical true score theory means that the error scores are uncorrelated to each other given the examinee’s true score, also referred to as local independence of item scores. IRT models formulate the probability of a response to an item as a function of an examinee’s latent trait and an item’s features (i.e., item difficulty, discrimination, and guessing).

In IRT models, local independence assumes that the probability of a response

In IRT models, local independence assumes that the probability of a response pattern of all items is a product of the probabilities of individual items, given the examinee’s ability level. Mathematical expression of local independence is presented as

$$p(X_1, X_2, \cdots, X_n \mid \theta) = p(X_1 \mid \theta) \times p(X_2 \mid \theta) \times \cdots \times p(X_n \mid \theta),$$

where p is the probability of a response pattern or an individual response $(X_1, X_2, \ldots, X_n)$ is a vector of a response pattern for all items, $X_1, X_2, \ldots, X_n$ are individual responses to Item 1, Item 2,…, Item n, and θ is the ability level. Local independence is essential in IRT because many IRT models are formulated based on the local independence assumption using this mathematical equation. Local independence also fits the multidimensional IRT models. In sum, the assumption behind the local independence is that the latent variable(s) being measured by test items is the only factor that affects students’ performance on the test. Thus, local independence is also related to the dimensionality assumption.

Potential Sources and Statistical Detection Methods of Violations of Local Independence

Potential sources of violations of local independence have been proposed for several decades. Wendy M. Yen broadly discussed some of the sources for LID in her journal article in 1993. Some are related to examinees, such as external assistance, speededness, fatigue, and practice. Some are related to the test or test items, such as item or response format, passage dependence, item chaining, explanation of previous answer, scoring rubrics or raters, and exposure in the curriculum of testing content, knowledge, and abilities. The key idea of these additional effects causing violations of local independence is that they consistently disturb the performance of some students on some test items to a great degree. It would not lead to LID if they have equal effects to all examinees and/or to all test items. The statistics, Q_3 proposed by Yen as well as Pearson’s chi-square and the likelihood rate G^2 developed by Wen-Hung Chen and David Thissen, are commonly used to detect LID. The testlet IRT models are also developed to fit the data that inevitably occur LID, such as reading comprehension items with the same passage.

Yi-Hsin Chen

See also Item Response Theory; Latent Class Analysis

Further Readings

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22, 265–289. doi:10.3102/10769986022003265

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

S. Jeanne Horst S. Jeanne Horst Horst, S. Jeanne

Jessica N. Jacovidis Jessica N. Jacovidis Jacovidis, Jessica N.

Locus of Control Locus of control

988

989

Locus of Control

Locus of control refers to the extent to which individuals believe they have control over events in their lives. Those who believe they can personally influence the events that happen to them exhibit an *internal locus of control*. In contrast, those who believe events in their lives are outside of their personal control (e.g., due to powerful others, fate, luck, or chance) exhibit *external locus of control*. This entry provides an overview of locus of control, a brief description of group and cultural differences, and a summary of the measurement of locus of control.

Although locus of control is sometimes seen as a stable personality trait, theory and research indicate that it is largely a learned response. Locus of control is often included as a variable in educational and psychological research and is related to personal adjustment and success. Locus of control tends to correlate with variables such as academic achievement, interpersonal and familial relationships, job performance, physical and mental health, and the ability to cope with stress. Further, locus of control has contributed to other psychological theories such as learned helplessness and self-efficacy.

The concept of locus of control was studied throughout the late 1950s but is often attributed to the work of Julian Rotter (rhymes with “voter”) in 1966. Locus of control is the most well-known feature of a larger theory proposed by Rotter. This theory was one of the first to bridge behavioral and cognitive psychology, stating that human behavior depends on how strongly we expect a positive outcome (outcome expectancy) and how much we value the reinforcement (reinforcement value). Behavior is guided by reinforcements (i.e., rewards and punishments), and we develop beliefs about whether we can control rewards and punishments), and we develop beliefs about whether we can control those reinforcements (i.e., locus of control).

As an example, say we have a student who earns an “A” on an exam. If the student has high internal locus of control, the student may be more likely to attribute the high grade to effort put forth in studying for the exam. Consequently, the belief that the student has control over the grades the student makes is reinforced. In contrast, if the student has high external locus of control, the student may be more likely to attribute the high grade to an external force (e.g., the teacher selecting easy items). Thus, the belief that the student does not have control over the grades the student makes is reinforced, and the student may be unsure of how to achieve this outcome in the future.

General Versus Specific Locus of Control

The examples thus far describe general locus of control. That is, a person’s general approach to explaining life events may be internal or external. However, locus of control may also consist of dimensions specific to certain contexts, such as health-or work-specific locus of control, which refers to attributions people make about health or work events. For example, those with high blood pressure may feel that their blood pressure can be controlled through diet, exercise, and weight management (i.e., high internal locus of control). Alternatively, those with high external locus of control may attribute their high blood pressure to genetics or fate. Research findings suggest that health-specific locus of control relates to overall better health outcomes. Similarly, work-specific locus of control relates to overall job satisfaction, commitment to the job, burnout, absenteeism, social support, and other areas of work adjustment.

Group and Cultural Differences

Like other personality concepts, some group and cultural differences in locus of control are relevant. There appears to be some variation in locus of control across the lifespan. Generally, locus of control becomes more internal as an individual approaches middle age and becomes more external thereafter. Additionally, males tend to report higher internal locus of control than females.

Locus of control also differs by culture. For example, individuals from Asian countries tend to report higher external locus of control than individuals from the United States. One explanation for these cultural differences lies in whether the society is collectivist or individualist. Differences between individuals from the United States and European countries tend to be small. Within the United States, African Americans report higher external locus of control than non-Hispanic White Americans, even when taking into consideration factors such as socioeconomic status.

Measuring Locus of Control

Locus of control is generally conceptualized as a unidimensional continuum, ranging from external to internal. Locus of control is typically measured via self-report scales, where people respond to written statements describing themselves in light of attitudes related to locus of control. Most widely used is the Rotter Internal-External Scale, which consists of 23 pairs of locus of control items and 6 pairs of filler items. Respondents choose between two options—one internal and one external. Scores are based on the number of internal and external statements endorsed. Two other common measures are the Adult Nowicki-Strickland Internal-External Locus of Control Scale and the Duttweiler Locus of Control Scale. The full Adult Nowicki-Strickland Internal-External Locus of Control Scale consists of 40 items, answered yes or no, and includes versions appropriate for use with children. The Duttweiler Locus of Control Scale consists of 28 items, answered on a 5-point Likert scale (rarely, occasionally, sometimes, frequently, or usually).

S. Jeanne Horst and Jessica N. Jacovidis

See also Behaviorism; Learned Helplessness; Personality Assessment; Reinforcement; Self-Efficacy; Self-Report Inventories

Further Readings

Cheng, C., Cheung, S., Chio, J. H., & Chan, M. S. (2013). Cultural meaning of perceived control: A meta-analysis of locus of control and psychological symptoms across 18 cultural regions. Psychological Bulletin, 139, 152–188.

Furnham, A., & Steele, H. (1993). Measuring locus of control: A critique of general, children’s, health-and work-related locus of control questionnaires. British Journal of Psychology, 84, 443–479.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs: General and Applied, 80, 1–28.

Wang, Q., Bowling, N. A., & Eschleman, K. J. (2010). A meta-analytic examination of work and general locus of control. Journal of Applied Psychology, 95, 761–768.

Jennifer C. Greene Jennifer C. Greene Greene, Jennifer C.

Logic Models

Logic models

989

994

Logic Models

A logic model is an organized graphic display of the major components of a social or educational program, either an existing program or a proposed new intervention. A logic model is sometimes accompanied by a narrative description. The program components typically included in a logic model are inputs or resources (time, expertise, and financing); activities (the services, trainings, experiences, or other pursuits offered to program participants); outputs (short-term benefits for participants, e.g., steady access to job training); and short-and long-term outcomes (sustained changes in participant knowledge, skills, health, status, opportunities, and well-being). Beginning in the 1980s, the construct of a logic model was iteratively developed and implemented by the community of program evaluators, primarily in the United States.

For evaluators, a logic model provides an appropriate and useful framework or structure for generating contextually relevant evaluation questions and prioritizing evaluation efforts. The utility of logic modeling has been extended to program planners and developers, as a logic model not only offers a framework for establishing relevant evaluation questions and priorities but simultaneously provides a framework for thoughtful and well-substantiated program development. This entry discusses the historical roots of logic models, looks at the contributions of logic model constructs to program development and program evaluation, and provides an example of the use of a set of logic models to address childhood obesity.

Historical Roots

The connection between the clear and defensible articulation of the underlying logic of a program on the one hand and the development of well-targeted and contextually relevant evaluation priorities has a relatively long history within the contemporary development of the field of evaluation. As early as 1980, Joseph Wholey, who worked in an evaluation capacity for the U.S. federal government, developed the concept of *evaluability assessment*. Using the lenses of market analysis, an evaluability assessment yields a judgment of a program's readiness to benefit from an evaluation. Readiness is enhanced when (a) program goals and priority information needs are well-defined, (b) program goals are clearly articulated and attainable, (c) relevant performance data can be collected at reasonable cost, and (d) intended users of the evaluation results have agreed how they will use this information. Absent these conditions, evaluation results are not likely to be useful or used.

With some prescience, more than a decade earlier, Daniel Stufflebeam developed a CIPP (Context, Input, Process, Product) framework for evaluating many federal programs that did not meet the requirements for randomized experimental evaluations. A formative CIPP evaluation was designed to provide guidance for program development, based on contextual and conceptual analyses. A summative CIPP evaluation intended to compare data on program design, implementation, and outcomes to known participant needs and to accomplishments of results for critical competitors. Stufflebeam's CIPP thinking importantly contributed to the later development and widespread popularity of the logic model as a framework useful for both program planning and program evaluation.

In the 1980s and 1990s, another construct in the logic model family appeared in the work of evaluation scholars Carol Weiss and Huey-Tsyh Chen. This is the construct of *program theory*. A program theory articulates the causal change model that *underlies* a social or educational program and that is specifically designed to redirect or refocus behaviors or encourage movement in a new direction by the intended participants. A logic model is a descriptive account of the planned building blocks of change. A program theory is an explanatory account of the intended change process. The Aspen Roundtable for Community Change popularized the program theory construct and extended it to practitioners by promoting a theory of change approach in their early 2000s community development initiatives. The theory of change construct quickly spread to multiple sectors around the globe.

One additional member of the logic model family is the logical framework

One additional member of the logic model family is the logical framework analysis (LFA) and the related logical framework matrix. The LFA is a tool for project planning, management, and evaluation, originally developed for military planning and now popular in the international development field. Like the logic model, the LFA presents a description of the target project, including the logic model components of inputs, activities, outputs, and outcomes. The LFA also includes statements of the external (pre)conditions needed for project success and the indicators to be used to measure project progress. The logical framework matrix formalizes this planning in a structured matrix that also includes sources of verification and further details of the planning needed to ensure or address the external conditions required for project success.

In contemporary evaluation practice, some form of attention to the design and logic of the program being evaluated is common; in fact, it is almost considered standard practice. In parallel, in contemporary program planning and program administration—especially within the domains of social and educational programming—logic models or other forms of representing the core elements of a program are also commonly featured as components of standard good or even best practice.

Contributions of Logic Model Constructs to Program Development

Logic models in various forms have become quite versatile tools, with multiple potential contributions to educational and social program development and to subsequent evaluations thereof. Several of these contributions are presented in this and the next sections for program development and evaluation, respectively. This discussion is intended to apply to multiple forms of logic models, including those that focus on program components (inputs-activities-outputs-outcomes), those that also include attention to key contextual factors, and those that further include attention to underlying mechanisms of causal change. The latter is elsewhere commonly referred to as a program theory precisely because it includes the mechanisms of the change process.

Stronger Program Designs

The W. K. Kellogg Foundation has promoted the use of logic models in program development using the following arguments, among others. The development of a logic model requires clear thinking about just how planned resources and activities will lead to desired outcomes. Logic models can thereby cultivate critical reflections on a program’s design and aspirations. Notably, high aspirations require powerful interventions. For example, revitalizing the economic health and independence of a poor community likely would require more than the development of community gardens. Even if these gardens are themselves bountiful and contribute measurably to a healthier and less costly diet for participating families, this intervention may not be powerful enough to transform the economics of the whole neighborhood.

The development of a logic model, or a similar representation of a planned social or educational intervention, can also surface underlying assumptions, especially among the program planning team. These assumptions can be about the character of the problem being engaged, the nature of the population to be served, the feasibility of a planned action, the availability of requisite resources, and more. Examination of these assumptions—through discussion, research, and consultation with experts and community members—can provide insight into their justifiable warrant and practical feasibility. With more grounded and warranted assumptions in hand, a stronger logic model and program design can be developed.

Stronger Program Designs in Context

In addition to strengthening the internal coherence of a program’s design, logic model thinking can strengthen the contextual fit, and thus the likely success, of a planned intervention. As noted earlier, the logical framework matrix work emphasized attention to the external or contextual conditions needed for project success.

Thoughtful, logic-oriented program planning addresses the question of what needs to be in place or to happen in the context for the program to be implemented as designed and to have a good chance of being successful. Meaningful and useful responses to this question may well be procured from long-time community leaders, residents, and activists; local politicians and clergy; and local media; in addition to substantive experts. Questions of contextual fit and responsiveness are critical components of good program planning. Their importance is underscored via critical reflections on a program’s logic and its likely success in the particular contexts at hand.

Meaningful Participation by Diverse Constituencies

Logic model thinking can serve as an opportunity for using inclusive program development processes, specifically including practitioners. In logic, modeling practitioners, that is, those who will be responsible for delivering services in the program under development, have an important voice in conversations about program planning, implementation, evaluation, and, more broadly, knowledge generation. Practitioners have direct interactions with the intended beneficiaries of the program and often have unique understandings of beneficiaries' life circumstances, assets, and challenges, and how likely they are to respond to the program design and activities being developed.

Logic modeling as a program development process can also include as participants local and community leaders, as well as representatives of the intended beneficiaries of the program. This more fully inclusive program planning and development process can lead to (a) enhanced understanding of what kind of program is most likely to successfully provide the intended services to the targeted audiences, and (b) opportunities for participation and voice provided for selected beneficiaries. These dual benefits—both to the contextual meaningfulness of the program's design and to the sense of self among the local participants in the process—are similar to those advanced by participatory approaches to program development, planning, and evaluation.

Contributions to Organizational Capacity

Beyond specific enhancements to the internal logic, contextual fit, beneficiary relevance, and thereby overall quality of a program's design, logic model thinking can also contribute to organizational health and capacity.

Specifically, participating in conversations about the logic of a program's design, resources, and intended outcomes can cultivate stronger evaluative thinking among staff in an organization. Determining what constitutes a worthy and defensible program design is, indeed, an act of evaluative thinking. Deliberating with colleagues about the kind of program most likely to succeed in the targeted contexts is, similarly, partaking in evaluative thinking. So, a thoughtful logic modeling initiative in an organization can catalyze stronger and more reflective evaluative thinking among participants.

Similarly, an additional contribution of a thoughtful and inclusive approach to

Similarly, an additional contribution of a thoughtful and inclusive approach to program development—like constructing a logic model—is a benefit for the group or the team involved. Specifically, the process of thoughtfully developing a logic model for a planned program initiative can generate shared understandings of the intent of this initiative and can strengthen team cohesion.

Strong, critical, reflective thinking about the character and quality of a program’s design and intended effects can also contribute importantly to the character and quality of an evaluation of this program. These contributions are presented next.

Contributions of Logic Model Constructs to Program Evaluation

A strong, well-considered logic model or program theory can provide a contextually relevant and practically useful road map for developing and implementing an evaluation plan. This includes the overall purpose, audience, and intended uses of the evaluation; the key questions to be addressed; the criteria to be used to judge program quality; the design of the study; and the character and role of communications with and reporting to key program stakeholders during the evaluation process.

Clearly, an evaluation is importantly shaped by the policy and decision context for which it is commissioned. But, often, the evaluation request or commission remains quite general, leaving multiple key decisions to be worked out by the evaluator, in consultation with key program decision makers and stakeholders. A thoughtful logic model or program theory can contribute directly to an evaluation plan that is appropriately tailored to the context, to stakeholder information needs, to particular program values (e.g., inclusion or equity), and to pending decisions. Examples of these contributions are explained below.

Appropriate Evaluation Purpose, Audience, and Key Questions

Often, evaluators are asked to conduct an outcomes evaluation to assess how well a program has reached its intended outcomes. This request is made for well-established programs and relatively new programs alike. A thoughtful logic model of the program can clearly convey the readiness of the program for an outcomes evaluation and can catalyze a redirection of the evaluation for a program that is still being piloted and refined. This redirection could usefully focus on the quality of the program design and implementation, rather than on outcomes. Further, this thoughtful logic model can be used by the evaluator, and supportive program stakeholders, to make a convincing case for this redirection to the evaluation commissioners and funders.

In counterpoint to this example, a well-developed logic model can advance outcome accountability in the context at hand. The logic model for a mature program will have a set of clearly identified and well-defined outcomes. An evaluation focused on the magnitude and character of these outcomes, as specified in the logic model, can enhance program accountability for outcome attainment.

Appropriate Criteria to Be Used to Judge Program Quality

A well-developed logic model can also contribute to the articulation of the key criteria to be used to judge program quality. In evaluation, these criteria are typically established for the quality of the program experience and the quality and magnitude of program outcomes. That is, what constitutes meaningful and likely consequential program participation? And what important changes in participants' lives can be expected from program participation? Criteria can also be established for other dimensions of a program, both as designed and especially as implemented. For example, many programs are targeted for particular audiences or participants; thus, criteria for participant recruitment and selection could be established, aided by a thoughtful logic model.

Paradoxically, criteria for judging quality are often not specified in an evaluation plan, although they constitute the very heart of the evaluation enterprise. More deliberate use of logic model thinking may well advance this underdeveloped component of evaluation practice.

Meaningful and Consequential Evaluation Use

The sine qua non of evaluation is use. Evaluators rarely, if ever, initiate an evaluation study. Rather, they respond to requests for proposals from other individuals and organizations who wish to have an evaluation conducted of a particular program—either voluntarily or because it is required by a funder, a board, or existing policy or legislation. So, the merit and worth of evaluation studies are significantly assessed by their usefulness for various stakeholders and actions.

A thoughtful logic model can contribute positively to evaluation use in several ways. First, as discussed earlier, logic models can help identify critical, high-priority questions for the evaluation and important quality criteria for evaluative judgments. Aided by a logic model, the evaluation questions and quality criteria identified for the evaluation can be both central to the program's theory and contextually important at the time the evaluation is being conducted. Second, the process of developing and discussing a logic model, and the resulting graphic program portrayal, can help position evaluation as an opportunity for learning and not only a means for accountability. Third, like the use of logic models for program development, using the nontechnical language and accessible process of logic modeling to help plan an evaluation can be a welcoming and inclusive process that itself contributes to enhanced evaluation use.

Example: Preventing Childhood Obesity

In 2007, the U.S. Institute of Medicine (IOM, now called the National Academy of Medicine) issued a report by an IOM-convened Committee on Progress in Preventing Childhood Obesity. The charge of this committee was to assess the progress made since the 2004 IOM publication of a comprehensive action plan to combat childhood obesity. The committee focused its work on developing a framework and agenda for evaluating childhood obesity prevention efforts. In service of this goal, the committee held three regional hearings and reviewed numerous evaluation reports and other related resources.

The framework generated by this committee is a logic model, or rather a set of logic models, one for each of five identified sectors: government, industry, communities, school, and home. Each model has three primary columns. The first column is "resources and inputs," which include leadership, political commitment, and adequate funding. The second column is "strategies and actions," for example, research, education, partnerships, collaboration, and new technologies. And the third column is a set of "outcomes" representing different sectors in the prevention of childhood obesity. These outcomes include structural, institutional, and systemic outcomes; cognitive and social outcomes;

environmental outcomes; behavioral outcomes; and, ultimately, health outcomes. A box below the whole logic model identifies “crosscutting factors that influence the evaluation of policies and interventions,” for example, demographic, contextual, and cultural characteristics. In the 2007 IOM report, this basic logic model is then elaborated differentially for each of the five identified sectors of relevance.

This example illustrates the power and potential contributions of logic model thinking to substantial and critical societal challenges. There are few more critical challenges than the well-being of a nation’s children. More generally, this example demonstrates that a logic model can provide consequential framing for both a program and its evaluation. A logic model can foster critical thinking about (a) the enormously complex task of changing—constructively and for some duration—the knotty dysfunctions in human behavior, both in individuals and in our collective and intertwined systems, *and* about (b) how best to assess these changes.

Jennifer C. Greene

See also CIPP Evaluation Model; Program Evaluation; Program Theory of Change; Utilization-Focused Evaluation

Further Readings

Funnel, S., & Rogers, P. J. (2011). Purposeful program theory: Effective use of theories change and logic models. San Francisco, CA: Jossey-Bass.

Koplan, J. P., Liverman, C. T., Kraak, V. I., & Wisham, S. L. (Eds.). (2007). Progress in preventing childhood obesity: How do we measure up? Washington DC: Institute of Medicine, National Academies Press.

McLaughlin, J. A., & Jordan, G. B. (2004). Using logic models. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.)., Handbook of practical program evaluation (2nd ed., pp. 7–32). San Francisco, CA: Jossey-Bass.

Rogers, P. J., Hacsi, T. A., Petrosino, A., & Huebner, T. A. (Eds.). (2000). Program theory in evaluation: Challenges and opportunities. New Directions for Evaluation (Vol. 87). San Francisco, CA: Jossey-Bass.

Websites

Centers for Disease Control and Prevention. Program Performance and Evaluation Office—Program Evaluation. Other Evaluation Resources (including logic models). http://www.cdc.gov/eval/resources/

University of Wisconsin Extension, Program Development and Evaluation, Logic Models. http://www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html

W. K. Kellogg Foundation Logic Model Development Guide. https://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide

Thorlene Egerton Thorlene Egerton Egerton, Thorlene

Logistic Regression

Logistic regression

994

999

Logistic Regression

Logistic regression is a statistical method to test for associations, or relationships, between variables. Like all regression analyses, logistic regression is a predictive analysis where a model is tested to find out whether the value of one variable, or the combination of values of multiple variables, can predict the value of another variable. The distinguishing feature of logistic regression is that the dependent (also called outcome or response) variable is categorical. This entry first describes the method and the concepts of causal inference and biological plausibility. It then discusses positive and negative associations and the odds ratio and provides an example of the use of logistic regression analysis to determine whether depression increases the risk of older people needing home help. The entry concludes by reviewing some assumptions and sources of error in logistic regression.

In binary logistic regression, which is the most common type of logistic regression, the dependent variable is binary or dichotomous. That means that there can only be two options for its value. For example, yes/no, pass/fail, alive/dead, satisfied/unsatisfied, and so on. In logistic regressions where there are more than two categories for the dependent variable, a less common multinomial logistic regression test is needed.

The dependent variable is the thing you are trying to explain or predict. There can be one or multiple independent (also called predictor or explanatory) variables tested in your model, and these can be either discrete variables (including dichotomous or ordinal), or they can be continuous (interval) variables. The term *dependent* suggests that this variable is dependent upon the status of the independent or predictor variable(s). As with all regression analyses, when there are multiple independent variables in a model, you are testing the predictive ability of each independent variable while controlling for the effects of other predictors. In logistic regression, the results lead to an estimation of the change in probability or odds of the outcome event occurring with a change in the value of the independent variable(s) relative to the probability or odds of the outcome event occurring given no change in the predictor variables. The results are not as easily interpreted as the results of a linear regression analysis, where the level of the outcome can be predicted from the predictor variables.

In logistic regression, the odds of the outcome of interest occurring for one unit change in the predictor variables is given in relation to the null hypothesis or equal odds. Equal odds is represented by an odds ratio value of 1.0. An increase in odds of the outcome occurring is indicated by an odds ratio value of greater than 1.0, and a decrease in the odds of the outcome occurring is indicated by an odds ratio value of less than 1.0. Statistically significant odds ratios are an indication of an association existing between the variables. The further the odds ratio number is from 1.0, the greater or stronger the association.

An example of a logistic regression inquiry can be: Does the value of x (independent variable) change the likelihood of y (dependent variable) being “yes” (rather than “no”)? For example, does eating bread crusts increase the likelihood of having curly hair (rather than straight hair)? In this case, a statistically significant odds value of greater than 1.0 would indicate that eating bread crusts does increase the chance of hair being curly.

Logistic regression can also indicate the strength of this predictive relationship by providing a value for the increased or decreased odds of the outcome occurring for a given change in the predictor variable. In our example, if the odds ratio is only a little bit greater than 1.0, then eating crusts only slightly increases the likelihood of having curly hair, and other factors are probably more important. However, if the odds ratio is a lot greater than 1.0, then eating bread crusts makes a really big difference to your chances.

In another example, the odds of being obese among children watching 11–20 hours of TV per week compared with children watching ≤10 hours of TV per week is around 1.4. That is, children are 1.4 times more likely to be obese in the 11–20 hours group than in the ≤10 hours group—a rather modest 40% increase. However, for children watching more than 30 hours of TV per week, the odds of

However, for children watching more than 30 hours of TV per week, the odds of being obese is around 3.6 or 3.6 times the odds than for children watching TV for ≤10 hours. Note that the dependent variable is obese versus not obese, and the independent variable is TV watching per week in hours categorized into ≤10 hours, 11–20 hours, 21–30 hours, and >30 hours. The reference group in the model is the ≤10 hours per week group, and therefore the odds of being obese among children watching ≤10 hours of TV per week is assumed to be 1.0. The odds of obesity in the other groups is given relative to the odds for the reference group and hence called an odds ratio. So far the examples only have one independent variable. With multiple independent variables, you can see the relative importance of the predictors. That is, which of the variables in the model is the strongest predictor of the outcome?

Note that logistic regression does not tell you the actual likelihood or odds of an outcome in an individual. The results give the probability of the outcome occurring with 1 unit value higher of the predictor variable compared with the probability given the original value of the variable. Nor can logistic regression be used to determine whether a variable causes an increased or decreased probability of the outcome.

Causal Inference

The term *dependent* does not suggest that the independent variable(s) cause the outcome. This is a very important concept to understand when interpreting the results of regression analyses. In our example, if you found an association between eating bread crusts and curly hair, you cannot conclude that eating bread crusts causes curly hair. Similarly, it cannot be known from the TV watching data whether it is the increased TV watching that causes the increased likelihood of obesity. In fact, in this example, the causal relationship is likely to be complex, multifactorial, and possibly bidirectional. That is, there may be an element of higher body mass index (BMI) causing children to choose more sedentary behaviors. The possible reasons for a finding of increased odds include

A direct causal relationship exists. Eating crusts does in fact make your hair grow curly.

A reverse causal relation exists. Having curly hair makes you eat more bread crusts.

An indirect causal pathway. People who eat more bread crusts are more likely to have curly hair, but the causal pathway is more complex. For example, eating more bread crusts makes you drink more water and drinking more water makes your hair go curly.

A third factor is associated with both predictor and outcome variables. For example, bread crust eating tends to be higher in people who eat more bread, and it is bread that causes hair to be curly.

There is no relationship between eating bread crusts and having curly hair and the finding was purely coincidence. This is a false positive or a Type I error.

The point is that finding an association does not tell you which of these possible reasons for the association is true. If you find an association between two variables, you cannot assume that the predictor variable caused the increased odds of the outcome occurring. From the results, you may be able to suggest an explanation, but you need to test your new hypothesis with another type of experimental design. A significant association in a regression analysis does not necessarily indicate a causal relationship.

Biological Plausibility

This leads on to the concept of biological plausibility, or “Does this explanation make reasonable or logical sense?” Of course, in reality you should not find an association between eating bread crusts and curly hair because there is no biologically plausible rationale why eating bread crusts would make your hair curly. If you use a *p* value cutoff of <.05 for statistical significance in your logistic regression analyses, then 5 in every 100 relationships tested, where no relationship exists, will be statistically significant and simply a random chance finding. For this reason, it is important to use logistic regression to test only biologically plausible theories rather than to analyze all the combinations available in the data and then try to subsequently explain the significant relationships found. In other words, it is important to have a research question with a stated hypothesis or expectation before any analyses are carried out.

Positive and Negative Associations and Increased or Decreased Odds

Whenever a logistic regression analysis identifies an association, the association may be either positive or negative. This tells you about the direction of the association or whether the factor increases or decreases the likelihood of the association or whether the factor increases or decreases the likelihood of the outcome of interest. In terms of odds, a positive association produces an odds ratio of greater than 1.0, and a negative association produces an odds ratio of less than 1.0. For this reason, it is important to be aware of how you code your outcome variable for the analysis.

Typically, statistical software packages will provide the odds for the outcome coded with the higher value compared with the outcome coded with the lower value. Thus, if you coded obesity with “one” and normal weight with “zero,” the odds will be for the probability of having obesity. In the example of higher grades at school increasing the odds of the student going on to tertiary education, if enrolling in tertiary education is coded “one” and not enrolling in tertiary education is coded “zero,” the association would be positive and the odds would be >1.0. However, if enrolling in tertiary education is coded “one” and not enrolling in tertiary education is coded “two,” the association would be negative and the odds would be <1.0. The results have the same interpretation. The odds ratio values are simply the inverse of each other. The odds of enrolling in tertiary education are better for students with higher grades and worse for students with lower grades. The difference in direction of the association and value of the odds ratio is simply due to the coding.

Example of Logistic Regression Analysis

Let’s use the following example to help explain the results of a logistic regression analysis. An analysis tested the hypothesis that depression increases the risk of older people needing home help. The model has one independent (predictor) variable, depression, and a dichotomous dependent (outcome) variable, home help. Depression scores can range from 0 (no depression) to 21. Home help can either be 1 (yes) or 0 (no).

Unstandardized Coefficient or B Value

In this model, the unstandardized coefficient (B value) was .15. The B value is similar to the B value in a linear regression analysis and can be used in a predictive equation. However, in logistic regression, the equation predicts the probability of a case falling into the desired category rather than the value for the outcome variable. In this case, the *B* value is positive; therefore, higher depression scores (if significant) are associated with greater likelihood of needing home help.

Standardized Odds Ratio, Exp(B), or β Value

The β value is the exponential of the *B* value, or the odds ratio, and because it is standardized, its magnitude can be considered relative to the magnitude of the β value(s) for other variable(s) in the model or for variable(s) in other models. The β value is the point estimate of the strength of the association. The further away the β value is from 1.0, the stronger the association. In the depression versus home help example, the β value is 1.16 for depression. That means the odds of needing home help are 1.16 times higher for someone reporting one point more on the depression scale than for a person with a depression score one point lower.

Significance

A *p* value of .05 is most commonly selected as the cutoff level to signify the statistical significance of an odds ratio. The cutoff value doesn’t have to be .05, and there may be reasons why you choose a cutoff value that is more (e.g., <.01) or less (e.g., <.1) stringent. A *p* value of <.05 means that there is a 5% chance of the association not being a true association and purely down to chance or coincidence or that there is 95% confidence of a true association existing between the two variables. Thus, if there is an association between two variables with a *p* value of .08, there is an 8% chance that a true association does not exist, which is generally considered unacceptably high.

The *p* value for the depression versus home help example was <.001. Therefore, we can be more than 99.9% sure that there is a true association between depression and home help (although we cannot assume that depression *causes* people to need home help).

Confidence Interval

The confidence interval is another way of expressing likelihood an association truly exists. The β value is the point estimate of the odds ratio, whereby odds of 1.0 means that a one increment change in the independent variable does not increase or decrease the probability that the dependent variable will be in the category of interest. If the 95% confidence interval includes 1.0, there is a greater than 5% chance that a true relationship between the variables does not exist. For example, a 95% confidence interval of [0.93, 3.76] includes an odds ratio estimate of 1.0 and therefore we cannot say with confidence that a true association exists. However, confidence intervals give more information than just statistical significance and therefore more information than *p* values.

The 95% confidence interval for the β values is the range within which we can be 95% confident that the true β value lies for your population of interest based on the information from your sample. Thus, while the β value gives the point estimate of the odds ratio and therefore an indication of how *much* greater or lesser the odds of the outcome is, the 95% confidence interval provides an estimation of the precision of your point estimate. In the example, the true odds is likely to be somewhere between 0.93 and 3.76. This is a wide range of possible values, so the estimation of the odds is considered imprecise. And while the data do not support there being an association, it would be foolish to conclude that no association exists.

The 95% confidence interval for the odds ratio for home help with an increase in depression score was [1.12, 1.19]. That means we can be 95% confident that the true value for the population is between 1.12 and 1.19. This is only a small increase in odds, but a very precise finding thanks to the large sample size available.

The *p* value and width of the confidence interval is highly influenced by the sample size and homogeneity of the sample. In other words, if you have a very large sample with a wide spread of values, then your *p* value is more likely to be smaller, your confidence interval narrow, and your point estimate is likely to be closer to the true value for the population. In the depression and home help study, there were data available from over 6,000 people which enabled such a precise estimate of the odds ratio.

The sample size can influence the *p* value and the precision estimate (confidence interval) but does not influence the strength of the association (point estimate) apart from the possibility of it being closer to the true population value with greater sample sizes.

Assumptions and Sources of Error

There are a number of assumptions and sources of error in a logistic regression analysis that should be considered. Logistic regression can handle ordinal and nominal data as independent variables as well as continuous (interval or ratio scaled) data. Binary logistic regression requires the dependent variable to be binary. Ordinal or interval data can be reduced to a dichotomous level but doing this loses a lot of information, which may make this test inferior compared to ordinal logistic regression or linear regression in these cases.

In regression analyses, it is good to have a wide range of values of the independent variable(s) in the analysis sample. If the sample includes only a small portion of the range of possible values for one or more of the independent variables, you might not get a very accurate indication of their relationship with the dependent variable. Certainly you will have limited generalizability of the results.

Models do not need to have linear relationships between the dependent and independent variables. Logistic regression can handle all sorts of relationships because it applies a nonlinear log transformation to the predicted odds ratio. The independent variables do not need to be normally distributed—although multivariate normality yields a more stable solution. Also the error terms (the residuals) do not need to be normally distributed.

As explained earlier, because logistic regression assumes that the odds ratio is the probability of the event occurring given a change in the independent variable, it is necessary that the dependent variable is coded accordingly for the event of interest. That is, for a binary regression, the higher factor level of the dependent variable should represent the desired outcome or outcome of interest.

Adding independent variables to a logistic regression model will always increase its statistical validity because it will always explain a bit more variance of the outcome. However, adding more and more variables to the model makes it inefficient and over fitting can occur. Only include as many variables as needed for your research question/hypothesis. That is, only the meaningful variables should be included. But you should try and include *all* meaningful variables, and this requires a good knowledge of the field of inquiry and deep consideration of the research question and hypothesis and is likely to be the most challenging part of a logistic regression analysis.

Logistic regression requires each observation to be independent, that is, that the data points should not be from any dependent samples design, such as before-after measurements, or matched pairings. The model should have little or no multicollinearity. That is, the independent variables should be pretty much independent from each other. As long as correlation coefficients among independent variables are less than .90, the assumption can be considered met. There is, however, the option to include interaction effects of categorical variables in the analysis.

Logistic regression assumes linearity of independent variables and log odds. Although it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. Otherwise, the test underestimates the strength of the relationship and rejects the relationship too easily (i.e., indicating there are not significant results or not rejecting the null hypothesis) when the relationship is significant. A possible solution to this problem is the categorization of the independent variables. That is transforming interval variables to ordinal level and then including them in the model. An example of this is to transform BMI values into ordinal categories of underweight (BMI < 20), normal weight (BMI = 20–25), overweight (BMI >25 but ≤30), and obese (BMI >30).

Large sample sizes are important. Maximum likelihood estimates are less powerful than ordinary least squares (used for simple and multivariable linear regression). Ordinary least squares analysis needs at least five cases per independent variable in the analysis; however, maximum likelihood estimates need at least 10 cases per independent variable, and some statisticians recommend at least 30 cases for each parameter to be estimated. Odds ratios are most accurate if the outcome rate in the sample closely approximates the outcome rate in the population. There should be no outliers in the data. The presence of outliers can be assessed by converting the continuous predictors to standardized, or *z* scores, and removing values below −3.29 or greater than 3.29.

Thorlene Egerton

See also Multiple Linear Regression; Odds Ratio

Further Readings

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Hoboken, NJ: Wiley.

Kulkarni, V. G. (2016). Modeling and analysis of stochastic systems. Boca Raton, FL: CRC Press.

Lemeshow, S., Sturdivant, R. X., & Hosmer, D. W. (2013). Applied logistic regression (Wiley series in probability and statistics). Hoboken, NJ: Wiley.

Daniel Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Rosemary Luyin Liang Rosemary Luyin Liang Liang, Rosemary Luyin

Longitudinal Data Analysis Longitudinal data analysis

999

1002

Longitudinal Data Analysis

Longitudinal studies utilize a research design that measures the same variables of interest repeatedly over a period of time for the same group of participants. This design allows researchers to examine change within individuals and contextual factors that account for interindividual differences. The analysis of data from such designs is common in educational, psychological, and sociological research. Examples of studies that made use of longitudinal data analysis are research on problem behavior and psychosocial development in youth published by Richard Jessor and Shirley Jessor in 1977, Michael Resnick and colleagues’ 1997 analysis of data from the National Longitudinal Study of Adolescent Health, and a 2012 study of the impact of a positive youth development program on the development of adolescents’ risk behavior by Daniel Shek and Lu Yu. This entry discusses the differences between longitudinal studies and cross-sectional studies, the forms and characteristics of longitudinal study designs and analysis, common models for quantitative longitudinal data analysis, and limitations of longitudinal data analysis.

Differences Between Longitudinal Studies and Cross-Sectional Studies

Longitudinal designs focus on tracking change in behavior seen by observing subjects over a period of time. Providing observations on the subject beyond one point in time, this design allows researchers to track the variations or development of characteristics of a target population in both individual and group levels. A significant advantage of longitudinal data collection is that it is able to distinguish a time-varying effect (i.e., variability of particular able to distinguish a time-varying effect (i.e., variability of particular characteristics occurred within an individual) from a cohort effect (i.e., difference between individuals in different age-groups).

Unlike longitudinal design, cross-sectional design does not provide repeated measurements on the data over time. It aims to describe the particular characteristics of the subjects at a single point in time. Cross-sectional studies are sometimes carried out to examine the links between different predictors and the outcome of interests.

Cross-sectional design is less effective in explaining cause-and-effect relationships than longitudinal designs because it is unable to show any indication of the order of the measured events. It also fails to provide definite information for distinguishing between cohort and time-varying effects. For instance, when investigating the age effect on social competence among adolescents, the finding of a cross-sectional design might indicate that older adolescents tend to have a higher level of social competence. In contrast, longitudinal research design can distinguish the effect that is due to increased age from the effect that is caused by individual differences. It also generates a trajectory of social competence to show the pattern of change.

Forms and Characteristics of Longitudinal Study Designs and Analysis

Longitudinal designs and data analysis has undergone a period of rapid development over the past 2 decades. There are four common forms of longitudinal design, namely, repeated cross-sectional study, panel design, event-oriented design (event history data), and qualitative longitudinal studies.

Repeated Cross-Sectional Study

Cross-sectional study is not suitable for describing and analyzing social change due to its one-off nature. It is common for cross-sectional data to be collected at two or more points in time so that the trend of development can be detected. Within the repeated cross-sectional design, the same questionnaire is applied to all data collection occasions based on different samples. These samples may either contain completely new cases or involve a very small number of cases that could be considered as insignificant.

There are some advantages in this type of longitudinal approach. First, the data are easier to gather and it could be analyzed through utilizing fairly simple statistical techniques. Second, repeated cross-sectional design is able to provide a long-term prediction of social change. Third, this approach does not require repeated measurement on the same research participants, so there is less worry about the risk of having a biased sample because of sample attrition (mortality).

Panel Design

Panel study is commonly regarded as the “true” longitudinal research by regularly collecting data from the same sample over a certain period of time. Measuring fixed subjects (e.g., students, nation-states) at fixed duration is the signature feature of the panel design. There are several types of data collection designs for panel study. For example, household panel study aims at following up all members who come from the chosen households over the course of time. Rotating panel study combines the features of cross-sectional and panel studies. It periodically replaces portions of samples by new yet comparable samples. This type of design has the strength of minimizing the occurrence of “survey boredom,” which may result in the loss of subjects. It is also an effective way to maintain the original features of the sample.

Cohort panel design is a special form of panel study that is conducted to track a sample of births in a given time period. As a further type of prospective panel study, linked panel study is mainly based on census or administrative data. Using this method, unique personal identifiers are linked together, although they were not initially collected for longitudinal purposes. Cross-sequential design is a method that provides repeated measurement of individuals from two or more cohorts. It addresses the inherited shortcomings of both longitudinal and cross-sectional design.

Event-Oriented Design

Event-oriented design is also known as event history data. It is another longitudinal method of recording the variations of the occurrence of one or more types of events. Unlike other longitudinal designs, which collect duration data at regular points in time, event-oriented design requires the respondents to recall the events based on their occurrence time or the reversed chronological sequences.

The gathered data of event-oriented design usually involves three essential parts: the duration of the event, its initial stage, and its terminal state. The first time when an event shows its significance in an individual's life is often determined as the start of the observation. Event-oriented design does not require continuous investigation of every discrete event. Instead, it only tracks the milestone events and therefore effectively minimizes the record errors. This methodological design permits the observation of an individual's life course. It is able to offer an understanding with reference to how all significant events change the developmental trend of the trajectories of the individual's life.

Qualitative Longitudinal Studies

There is often a misconception that only quantitative data (e.g., survey data, administrative records) are suitable for longitudinal analysis. In fact, qualitative methods, such as in-depth or structured interviews that focus on biographical information, could also be used in longitudinal study. This data collection technique is called biographical interview.

Generally speaking, biographical data require the interview to dig deep through the construction of sophisticated questions in order to make sure sufficient information is gathered and to reduce data distortion. Biographical analysis can be applied to detect the paths of academic development, mobility and career events, transitions, and variations in status, especially as they relate to age. It can also be applied to shifts in roles, particularly in relation to gender features. Semistructured or unstructured interviews, life stories, biographical interviews, biograms, letters, personal and day-to-day diaries, and life history calendars are the common forms of this type of design.

Common Models for Quantitative Longitudinal Data Analysis

Traditional Analyses

General linear model is a traditional longitudinal method that utilizes several types of analysis of variance-based analytical approaches. For example, multivariate analysis of variance and multivariate analysis of covariance are two widely applied models that extend the analysis of variance and the analysis of covariance and regression models. Basically, repeated measures of multivariate analysis of variance and multivariate analysis of covariance are designed for repeatedly measuring whether multiple response variables are simultaneously determined by grouping independent variables. Both approaches are made up of two-step procedures, the significant test and the post hoc test.

Although these models could offer observations on the pattern and the variability over time, they also have some shortcomings that limit their applicability. First, they cannot be applied to estimate unbalanced design over time. Second, the assumption of independence of observations intrinsic to these models could not be easily met due to the duplication of the data in longitudinal analysis. Third, these repeated measures models are unable to handle missing data. Thus, listwise deletion of missing data is typically used for addressing the problem, which often results in the significant decrease of sample size. Furthermore, as these models focus on measuring the group mean trend, they cannot provide information on how specific individuals change over time.

Linear Mixed-Effects Model (LMM)

LMM is also known as hierarchical linear model. It has been developed on the basis of the conventional multivariate linear models. LMM is commonly used to describe how a set of independent variables determines a continuous response under the longitudinal data setting. This linear approach involves a two-stage measurement. In the first stage, ordinary least squares regression is applied for estimating subject-specific regression coefficients. Then, the nonparametric (or standard parametric) methods are further performed to measure the estimated regression coefficients in the second stage. It is important to remember that LMM only applies to continuous response. Discrete longitudinal data such as binary response or categorical response could not be modeled within LMM. Another major limitation of LMM is that it requires the longitudinal linear data to be approximately normally distributed.

Generalized LMM (GLMM)

GLMM is an extension to the LMM. It has been frequently used in the analyses of longitudinal data. GLMM is a flexible approach that is particularly suitable for estimating nonnormal data that involve random effects and making sure that more realistic models can be applied to the data. GLMM allows repeated measurements of different types of responses, including continuous responses, binary responses, and counts. This is a significant feature that makes GLMM differ from LMM.

It is also important to note that GLMM has the technical advantage of dealing with unequal and small cluster sizes because data from all clusters are used for examining fixed regression coefficients as well as their standard errors. There are several software packages available for performing GLMM, such as R, SAS, and S-PLUS. Although GLMM presents various advanced functions on analyzing different types of data, as the development of this approach is still in its infancy, there are very few clear guidelines available for its operating procedure. Therefore, GLMM is a statistical approach that is simple at the theoretical level yet not easy to carry out.

Latent Growth Curve Model (LGM)

LGM could be considered as a special case of standard error of the mean. It provides examination of within-person difference over time as well as longitudinal between-person variability. In LGM, growth trajectory could be modeled for the observation of the change of outcome variables. LGM is able to not just investigate the antecedents and consequence of change but also detect the fit of model to data. It is viewed as one of the most flexible and useful techniques for longitudinal analysis. Unlike other conventional statistical techniques (e.g., multivariate analysis of variance and multivariate analysis of covariance), LGM could address the limitations of missing data or unequal sample size. Besides, it does not require cases to be measured on the same occasions. LGM could be applied for modeling complex nonlinear growth as well as the univariate and multivariate linear models.

Limitations of Longitudinal Data Analysis

Longitudinal designs and analyses have some inherent limitations. First, missing data and dropout are significant problems with these designs. In real life, it is challenging to follow the same respondents for a long period of time. Second, the occurrence of measurement errors in some variables is very common, and that problem can be confounded across multiple observation times. Because longitudinal design involves continuity and change, it is also more time-consuming and expensive compared with cross-sectional study. Huge manpower and financial resources are needed. Furthermore, the complexity of the and financial resources are needed. Furthermore, the complexity of the trajectories of longitudinal data may post increased challenges in the applications of statistical techniques.

Daniel Tan-lei Shek and Rosemary Luyin Liang

See also Analysis of Covariance; Analysis of Variance; Applied Research; Educational Psychology; Growth Curve Modeling; Hierarchical Linear Modeling; Mixed Methods Research; Multivariate Analysis of Variance; Time Series Analysis

Further Readings

Diggle, P., Heagerty, P., Liang, K. Y., & Zeger, S. (2013). Analysis of longitudinal data. Oxford, UK: Oxford University Press.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). An introduction to latent variable growth curve modeling: Concepts, issues, and applications (2nd ed.). Mahwah, NJ: Erlbaum.

Fitzmaurice, G. M., & Molenberghs, G. (2009). Advances in longitudinal data analysis: An historical perspective. Longitudinal Data Analysis, 3–30.

Hser, Y. I., Shen, H., Chou, C. P., Messer, S. C., & Anglin, M. D. (2001). Analytic approaches for assessing long-term treatment effects examples of empirical applications and findings. Evaluation Review, 25(2), 233–262.

Ruspini, E. (2002). Introduction to longitudinal research. New York, NY: Psychology Press.

Shek, D. T., & Ma, C. M. (2011). Longitudinal data analyses using linear mixed models in SPSS: Concepts, procedures and illustrations. The Scientific World Journal, 11, 42–76.

Shek, D. T., & Ma, C. M. (2012). Impact of the Project P.A.T.H.S. In the junior secondary school years: Objective outcome evaluation based on eight waves of longitudinal data. The Scientific World Journal, 2012. doi:10.1100/2012/170345

Shek, D., Sun, R. C., & Ma, C. (Eds.). (2014). Chinese adolescents in Hong Kong: Family life, psychological well-being and risk behavior: Vol. 5. Singapore: Springer.

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. Oxford, UK: Oxford University Press.

Sun, R. C., & Shek, D. T. (2013). Longitudinal influences of positive youth development and life satisfaction on problem behaviour among adolescents in Hong Kong. Social Indicators Research, 114(3), 1171–1197.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. Biometrics, 44, 1049–1060.

Carrie La Voy Carrie La Voy Voy, Carrie La

Long-Term Memory

Long-term memory

1002

1004

Long-Term Memory

Memory is the mental power, process, or capacity to remember, retain, or recall what has been learned. Research tells us that our brains are actively at work both storing and recalling information and that this process of creating memories does not occur in one single area of the brain.

Humans have the capacity to store both short-term and long-term memories. Long-term memory is the storage of information over a long period of time. Some information stored in lasting long-term memory may be lost, but long-term memory has an unlimited capacity and the information stored can last a lifetime. This entry first discusses the differences between long-term memory and short-term memory and those between implicit memory and explicit memory. It then discusses research on long-term memory.

Over the course of an individual's lifetime, memories from experiences and events are initially received as sensory input and briefly stored in short-term memory. Short-term memory is termed working memory, because it is temporary yet active. Some believe the capacity of short-term, working memory is between 15 and 30 seconds, while long-term memory is anything that lasts for more than 24 hours.

In 1956, George Miller described the capacity of short-term memory as "the magical number seven, plus or minus two." Miller suggested that short-term memory has a limited number of storage spots and adults can only store between five and nine items at one time. For various reasons, some of the information and experiences that are briefly stored in short-term memory are lost. However, with practice and under certain circumstances, initial experiences stay long enough to be moved to long-term memory with the potential for permanent storage.

Different regions of the brain are devoted to different functions related to storing memories. Short-term and long-term memories are housed in different places. Short-term memories of events and experiences are stored in a region of the brain called the hippocampus. The hippocampus is an important part of the limbic system of our brains that plays a vital role in learning and memory formation. In the hippocampus, short-term memories begin to be converted to long-term memories. When an event is experienced for the first time, a link is formed in the hippocampus. Other links are created and connected to the initial link. Memories are stored when connections between neurons are active and strengthened over time. The stronger the link, the quicker we are able to retrieve stored memories.

The conversion from short-term to long-term memory takes time. Memories moved to long-term memory become resistant to competing stimuli in the brain and other outside influences. Eventually, our experiences maintain a permanent place in our memory.

As experiences stored in short-term memory are collected and organized, immediate changes happen in our brains. Neurons, or nerve cells, play a part in creating or enhancing synapses in the brain. For long-term memories to survive for a long period of time, we must recall the memories from time to time. When we recall or remember something, we strengthen connections between neurons, or nerve cells, in our brain. These nerve cells exchange information and help us remember events and experiences.

In general, long-term memory can be categorized as implicit or explicit. Implicit, or nondeclarative memory, is recalled unconsciously. Little effort is required to recall these memories. Implicit memory includes cognitive processes, emotional memories, and motor skills and habits. These are memories that we recall automatically and without thinking. Examples include remembering how to brush our teeth, ride a bike, or drive a car. These are all things we learned to do at one time, but we soon do automatically, without thinking about the steps involved.

Declarative, or explicit, memory is the conscious recall of facts and general knowledge or other specific personal experiences and events. We must work to recall these memories. Explicit memory can be episodic or related to episodes or events. An example might be remembering when you purchased your first car or events. An example might be remembering when you purchased your first car or recalling a recent trip you took. Explicit memory can also be categorized as semantic, related to memories of factual information, and independent of human experience. An example of this is remembering what a car is, as opposed to remembering the experience of purchasing or driving a car.

Scientists and researchers have studied memory for many years and found there are ways to improve long-term memory, even as we age. For example, scientists have looked into the relationship between sleep and memory. Some findings suggest that experiences must be replayed in our minds, during sleep, in order to then be transformed and stored as long-term memories. Memories are strengthened or weakened with each new stimulus and experience. Scientists also believe repetition of experiences or events creates new synapses between nerve cells or stronger synapses between neurons in the cells of the brain.

Carrie La Voy

See also Behaviorism; Cognitive Development, Theory of; Cognitive Neuroscience; Learning Theories; Short-Term Memory; Working Memory

Further Readings

Bailey, C. H., Bartsch, D., & Kandel, E. (1996). Toward a molecular definition of long-term memory storage. Proceedings of the National Academy of Sciences of the United States of America, 93, 13445–13452.

Committee on Developments in the Science of Learning, Bransford, J. D., Brown, A., & Cocking, R. R. (Eds.). (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Darling-Hammond, L., & Bransford, J. (Eds.). (2005). Preparing teachers for a changing world. San Francisco, CA: Jossey-Bass.

Halber, D. (2009). Sleep helps build long-term memories. Cambridge, MA: MIT News Office. Retrieved from http://news.mit.edu/2009/memories-0624

Preston, A. (2016). How does short-term memory work in relation to long-term memory? Are short-term daily memories somehow transferred to long-term storage while we sleep? Scientific American. Retrieved from https://www.scientificamerican.com/article/experts-short-term-memory-to-long-term

LRE

LRE

1004

1004

LRE

See Least Restrictive Environment

M

John T. E. Richardson John T. E. Richardson Richardson, John T. E.

Mann-Whitney Test

Mann-whitney test

1005

1008

Mann-Whitney Test

A very simple design in quantitative research involves the random allocation of a sample of *N* individuals to two different groups. The groups are exposed to different treatments, and the research question examines whether there is any difference between the two groups on some criterion variable. Classically, this question is addressed using Student's *t* test for independent groups (or the equivalent one-way between-subjects analysis of variance). However, this procedure assumes that the criterion variable in question (a) is measured on an interval or ratio scale, (b) is normally distributed, and (c) has the same variance in both of the groups. The Mann-Whitney test (also known as the Mann-Whitney *U* test, the Wilcoxon-Mann-Whitney test, the Mann-Whitney-Wilcoxon test, or even simply the Wilcoxon test) was devised for use in situations in which one or more of these assumptions is not met. This entry describes the original derivation of the Mann-Whitney test, provides a simple worked example, discusses exactly what the test is measuring, and concludes by discussing the test's power and power efficiency.

Analysis by Ranks

In 1946, an American statistician, Frank Wilcoxon, suggested that the use of ranks would be helpful in situations in which the assumptions underlying Student's *t* test and other conventional procedures were not met. In a design in which the observations in two conditions are paired (such as a repeated-measures design), Wilcoxon argued that researchers could compute the difference scores between the two conditions and then rank the magnitude of the difference scores across the different cases; in this procedure, the direction of the difference was ignored in the ranking, but the total of the ranks for the positive differences was compared with the total of the ranks for the negative differences. The smaller of the two totals constituted a test statistic, which is usually denoted by the symbol T. Wilcoxon reported the probability of obtaining particular values of this statistic under the null hypothesis. This procedure became known as the Wilcoxon matched pairs signed-rank test.

In a design in which the observations in two conditions were unpaired (i.e., an independent-groups design), Wilcoxon proposed that researchers should rank the observations themselves, so that 1 refers to the smallest observation and N refers to the largest observation across *both* of the groups. The total of the ranks in each of the two groups is then calculated, and the smaller of the two totals is used as a test statistic, which is usually denoted by the symbol W. Wilcoxon once again reported the probability of obtaining particular values of this statistic under the null hypothesis of no difference between the groups. This is sometimes known as the Wilcoxon rank-sum test. Nevertheless, Wilcoxon's account was limited in two ways: First, he assumed that the two groups were of equal size and, second, he only provided a few points of the distribution of his statistic W under the null hypothesis.

Two other American statisticians, Henry B. Mann and D. Ransom Whitney, presented a more general account of this situation. Suppose that the numbers of cases in the two groups are n_1 and n_2 and that the totals of the ranks in the two groups are R_1 and R_2, respectively. A statistic U_1 is defined as $[n_1\, n_2 + n_1\, (n_1 + 1)/2 - R_1]$ and a statistic U_2 is defined as $[n_1\, n_2 + n_2\, (n_2 + 1)/2 - R_2]$. Each of these statistics measures the separation of the two distributions of scores (i.e., the extent to which the cases in Group 1 tend to score less than the cases in Group 2 and vice versa). It is easy to show that $U_1 + U_2 = n_1 n_2$, and hence, strictly speaking, one of the statistics is redundant. Mann and Whitney defined the statistic U as the smaller of U_1 and U_2. They calculated the exact probabilities of obtaining particular values of U or less for values of n_1 and n_2 between 3 and 8.

Mann and Whitney demonstrated that under the null hypothesis, the sampling distribution of U would approach the normal distribution as n_1 and n_2 increased. This distribution would have a mean of $(n_1 n_2 / 2)$ and a variance of $[n_1\, n_2\, (n_1 + n_2 + 1)/12]$. Consequently, for larger samples, the obtained value of U could be standardized and compared with the standard normal distribution. Exact critical values of U were subsequently published for values of n_1 and n_2 between 9 and 20, but these are largely redundant with the use of modern statistical packages. If $n_1 = n_2 = n$, Mann and Whitney's statistic U is related to Wilcoxon's statistic W by the formula $W = U + n(n+1)/2$.

The Mann-Whitney test assumes that the original observations have been measured on at least an ordinal scale, that they are independent of one another, and that those within each group come from the same population. Even so, it does not make any assumptions about the parameters of the populations from which the data are drawn, and hence it is an example of a nonparametric statistical test. Mann and Whitney recognized Wilcoxon's contribution to their work. However, William H. Kruskal subsequently identified six other independent accounts of procedures akin to Mann and Whitney's, going back to the work of a German statistician, Gustav Deuchler, in 1914.

A Worked Example

Suppose that two groups of four participants have undergone different treatments and have then received a standard test. The four participants in Group A obtained scores of 8, 9, 10, and 12, and the four participants in Group B obtained scores of 11, 13, 14, and 15. For samples as small as these, the overlap between the two distributions can be obtained by direct counting.

First, consider the number of participants in Group A who obtained lower scores than each participant in Group B. Three participants in Group A scored less than the first participant in Group B (11), all four participants in Group A scored less than the second participant in Group B (13), all four participants in Group A scored less than the third participant in Group B (14), and all four participants in Group A scored less than the fourth participant in Group B (15). Summing these figures, $U_1 = 3 + 4 + 4 + 4 = 15$.

Next, consider the number of participants in Group B who obtained lower scores than each participant in Group A. No participant in Group B scored less than the first participant in Group A (8), no participant in Group B scored less than the second participant in Group A (9), no participant in Group B scored less than the third participant in Group A (10), and just one participant in Group B scored less than the fourth participant in Group A (12). Summing these figures, $U_2 = 0 + 0 +$

0 + 1 = 1.

Note that $n_1n_2 = 4 \times 4 = 16$ and that $U_1 + U_2 = 15 + 1 = 16$. U is the smaller of U_1 and U_2, which is 1. According to the tables provided by Mann and Whitney, the probability of obtaining a value of 1 or less under the null hypothesis is .029. However, this result applies for a one-tailed test. For a two-tailed test, the result would be .058 and therefore not statistically significant.

Alternatively, the overlap between the two distributions can be obtained by ranking the eight participants, so the score of 8 is ranked as 1 and the score of 15 is ranked as 8. The ranks for the four participants in Group A are 1, 2, 3, and 5, so $R_1 = 11$. The ranks for the four participants in Group B are 4, 6, 7, and 8, so $R_2 = 25$. For this example, $n_1 = n_2 = 4$, and Wilcoxon's statistic W is the smaller of R_1 and R_2 (i.e., 11). Accordingly, $U_1 = [n_1\mathrm{n}_2 + \mathrm{n}_1 (\mathrm{n}_1 + 1) / 2 - R_1] = (4 \times 4) + (4 \times 5) / 2 - 11 = 16 + 10 - 11 = 15$ and $U_2 = [n_1\mathrm{n}_2 + \mathrm{n}_2 (\mathrm{n}_2 + 1) / 2 - R_2] = (4 \times 4) + (4 \times 5) / 2 - 25 = 16 + 10 - 25 = 1$, yielding the same results as were obtained by counting. Finally, note that $W = U + n(n + 1)$ *2 = 1 + (4 × 5)* 2 = 1 +10 = 11.

This example was deliberately chosen to avoid tied observations. Wilcoxon proposed that tied values should be assigned the mean of the relevant ranks but did not suggest any amendment of the calculations. However, the outcome is affected if ties occur between participants in both of the groups. A test for comparing k groups proposed by Kruskal and W. Allen Wallis is formally equivalent to the Mann-Whitney test when $k = 2$, and Kruskal and Wallis described a procedure that could adjust the outcome for tied values in this situation. The correction seems to make little practical difference, but it is employed routinely in modern statistical packages.

What Is the Mann-Whitney Test Measuring?

The statistic U measures the tendency for observations in one of the two populations to be larger (or smaller) when paired randomly with observations in the other population. András Vargha and Harold D. Delaney called this *stochastic heterogeneity*. In contrast, stochastic *homogeneity* is where the probability of an observation in one of the populations being larger or smaller when paired randomly with observations in the other population is exactly .5. In this case, Vargha and Delaney showed that the expected values of the mean ranks R_i / n_i were the same in both groups.

On the basis of results that they had obtained in the case of the Kruskal-Wallis test, Vargha and Delaney argued that the Mann-Whitney test was a valid test of the hypothesis of stochastic homogeneity only if the variance of the ranks was the same between the two groups. If this assumption were violated, Vargha and Delaney recommended the use of a robust parametric test on the ranks instead. They also showed that, if the relevant distributions were symmetrical, then stochastic homogeneity would imply equality of the group *medians*. However, it would only imply equality of the group *means* if the samples were actually drawn from identical populations. Conversely, when the relevant distributions are skewed, stochastic heterogeneity can result from differences in the variance or shape of the populations rather than differences in their means.

Power and Power Efficiency

The *power* of a statistical test is the probability of rejecting the null hypothesis when it is false. (Its complement is the probability of *not* rejecting the null hypothesis when it is false, in other words the probability of making a Type II error.) In general, nonparametric tests tend to be less powerful than the corresponding parametric test because they use less of the information that is contained in the data. (For instance, the Mann-Whitney test only employs the ranks of the observations, whereas Student’s *t* test employs the actual values of the observations.) Even so, Mann and Whitney stated that assessing their test’s power would present formidable difficulties.

The power of two different statistical tests in the same research design can be compared using the notion of *power efficiency*. This notion relies upon the fact that the power of a test in a particular situation depends (other things being equal) on the sample size. Suppose that Test 1 is the most powerful statistical test when used in a particular research design with data that meet its underlying assumptions. Test 2 is a less powerful test in the same design, in that it would need to be used with a sample of N_2 cases to match the power that is achieved by Test 1 with N_1 cases (where $N_2 \geq N_1$). The power efficiency of Test 2 is N_1 / N_2, often expressed as a percentage. In 1954, Alexander M. Mood demonstrated that the power efficiency of the Mann-Whitney test in comparison with Student’s *t* test for independent samples approached a value of 3 / π or 95.5% as the overall sample size increased. Accordingly, the Mann-Whitney test can be recommended as a powerful distribution-free test.

John T. E. Richardson

See also Analysis of Variance; Kruskal-Wallis Test; Power; *t* Tests; Type II Error; Wilcoxon Signed Ranks Test

Further Readings

Kruskal, W. H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. Journal of the American Statistical Association, 52, 356–360. doi:10.2307/2280906

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583–621. doi:10.2307/2280779

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50–60. doi:10.1214/aoms/1177730491

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. Annals of Mathematical Statistics, 25, 514–522. doi:10.1214/aoms/1177728719

Randles, R. H., & Wolfe, D. A. (1979). Introduction to the theory of nonparametric statistics. New York, NY: Wiley.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York, NY: McGraw-Hill.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics, 23, 170–192. doi:10.3102/10769986023002170

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1, 80–83. doi:10.2307/3001968

Juana Gómez-Benito Juana Gómez-Benito Gómez-Benito, Juana

M. Dolores Hidalgo M. Dolores Hidalgo Hidalgo, M. Dolores

Mantel-Haenszel Test

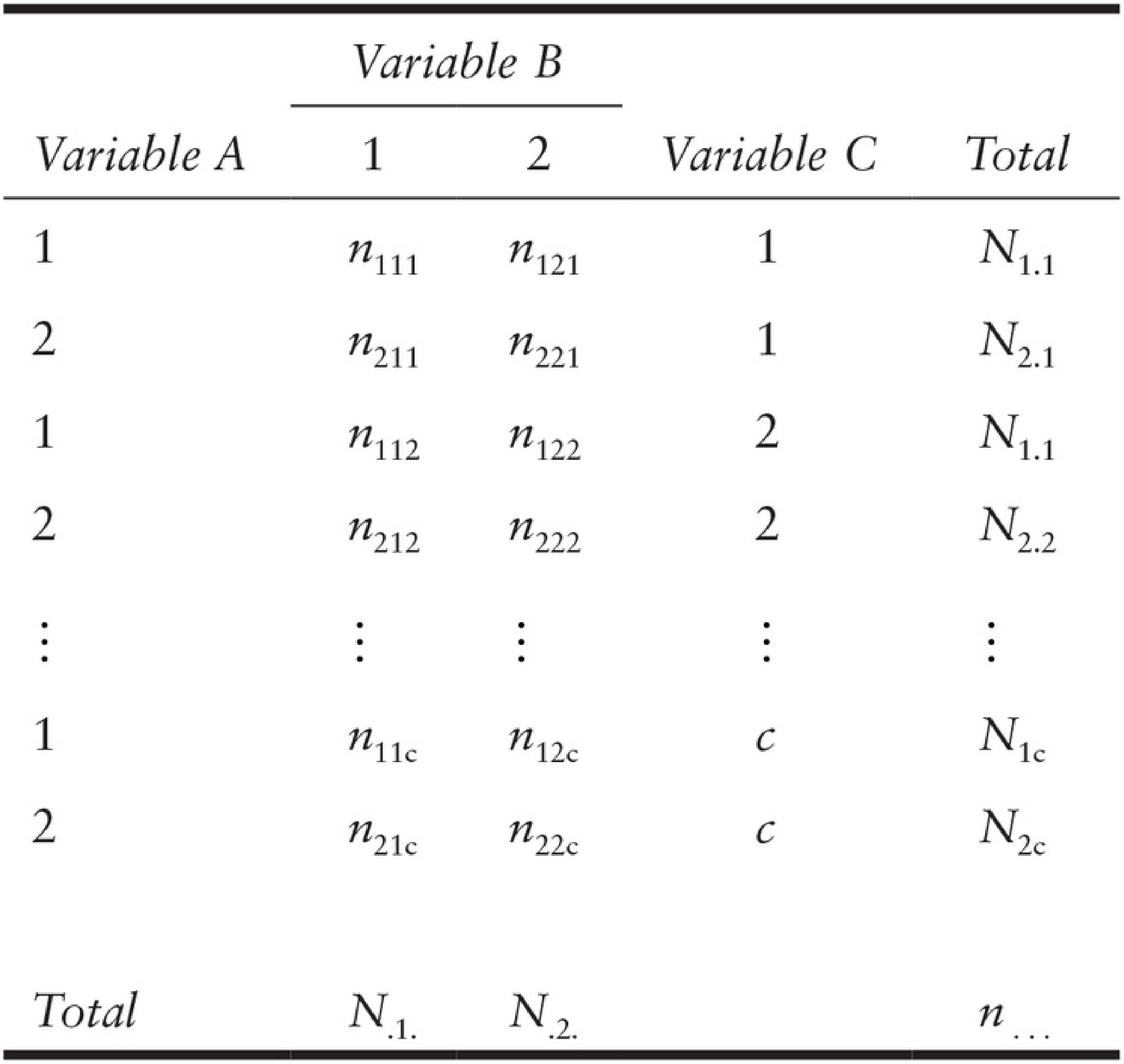
Mantel-haenszel test

1008

1011

Mantel-Haenszel Test

In the field of educational research, one often encounters situations in which the goal is to study the association between two variables, by means of a 2 × 2 contingency table, in the presence of an additional variable that, when taken into account, may lead to a better understanding of the relationship between these variables. Situations of this kind are most commonly found in quasi-experimental studies, in which the researcher has a classification variable and wishes to study the association between two other variables for each level of the classification variable as well as in meta-analytic studies and, especially, in the analysis of differential item functioning (DIF). In these three contexts, there are always *k* levels, studies, or groups, respectively, and in each case, the researcher analyzes the association between two variables, generating a $k \times 2 \times 2$ contingency table (see Table 1).



	Variable B			
Variable A	1	2	*Variable C*	*Total*
1	n_{111}	n_{121}	1	$N_{1.1}$
2	n_{211}	n_{221}	1	$N_{2.1}$
1	n_{112}	n_{122}	2	$N_{1.1}$
2	n_{212}	n_{222}	2	$N_{2.2}$
⋮	⋮	⋮	⋮	⋮
1	n_{11c}	n_{12c}	c	N_{1c}
2	n_{21c}	n_{22c}	c	N_{2c}
Total	$N_{.1.}$	$N_{.2.}$		$n_{...}$

In 1959, Nathan Mantel and William Haenszel proposed the Mantel-Haenszel (MH) statistic that can be used for calculating the degree of association between variables. It is given by the following expression:

$$MH = \frac{\left[\left|\sum_{j=1}^{c} n_{11_j} - \sum_{j=1}^{c} E(n_{11_j})\right| - 0.5\right]^2}{\sum_{j=1}^{c} \mathrm{Var}(n_{11_j})},$$

where $E(n_{11j})$, the expected value of n_{11j}, is given by

$$E(n_{11_j}) = \frac{N_{1.j}N_{.1j}}{N_{..j}}$$

and its corresponding variance by

$$\mathrm{Var}(n_{11_j}) = \frac{N_{1.j}N_{2.j}N_{.1j}N_{.2j}}{(N_{..j})^2(N_{..j}-1)}.$$

This statistic follows a χ^2 distribution with one degree of freedom. If, for a given level of confidence, the value of the MH is greater than the theoretical value of , we can state that there is a relationship between Variables A and B when the effects of Variable C are controlled for. These hypotheses are normally operationalized in terms of the odds ratio (*OR*), which for a given level of variable C is defined as follows:

$$OR_{MH} = \frac{n_{11_j}n_{22_j}}{n_{12_j}n_{21_j}}$$

and which may take values from 0 to infinity.

After describing what the MH test provides, this entry reviews the detection of DIF by way of the MH test, provides an example of this analysis method, and offers some concluding remarks.

The MH analysis provides (a) a statistical test of whether the *OR*s are equal (homogeneous) or unequal (heterogeneous) across levels of the C variable and (b) an estimation of the *OR* that represents the odds that an outcome (i.e., the B variable) will occur given a particular level of the A variable, as compared with the odds of this outcome occurring in the absence of this level of the A variable (i.e., in the presence of the other level of the A variable).

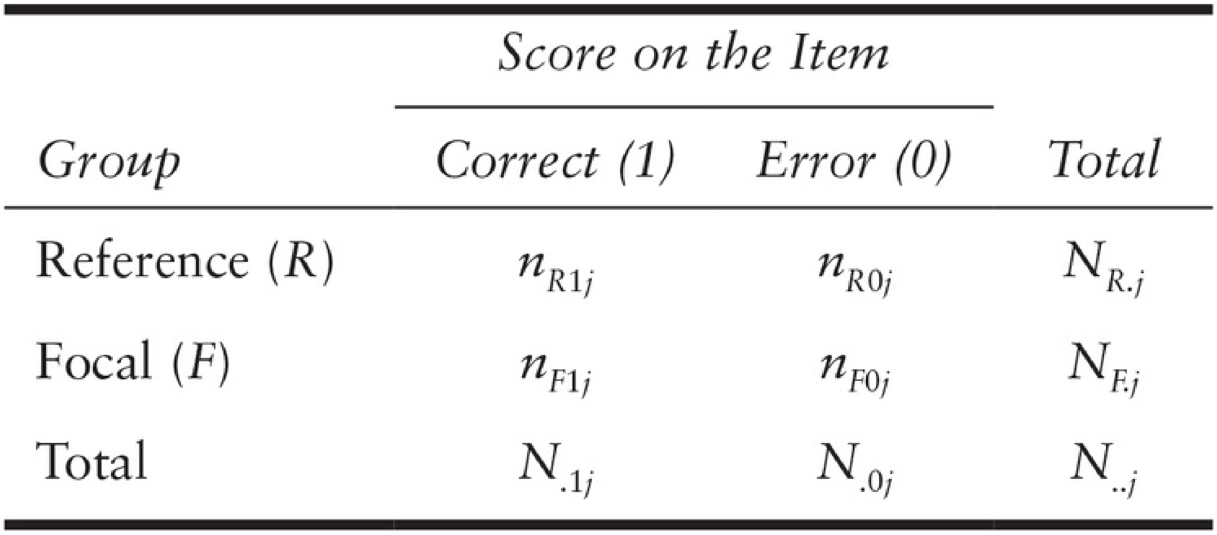
When using this procedure, it is necessary to bear in mind two assumptions: Observations are independent of each other, and all observations are identically distributed.

In 1963, Mantel proposed an extension of the statistic for use with ordinal categorical variables and $k \times m \times 2$ tables, and in 1986, Grant W. Somes proposed the generalized MH statistic that can be used in the case of variables with more than two response levels but that are not necessarily ordinal.

Detection of DIF by Means of the MH

Currently, one of the most common applications of the MH statistic is in the analysis of DIF in educational or psychological tests. DIF analysis indicates whether each item of a test behaves the same or differently across different groups or, in other words, whether the probability of responding in a certain direction or correctly to a given item depends on the group to which the respondent belongs. The identification of DIF indicates that the item (or test) functions differently across comparable groups of examinees, that is, groups that have been matched on the trait or construct that the test is designed to measure.

Although the MH statistic was developed toward the end of the 1950s, almost three decades passed before Paul W. Holland and Dorothy T. Thayer adapted the procedure as a technique for detecting DIF in dichotomous items. Since that time it has become one of the most widely used techniques in the DIF field. Its success is due not only to its simplicity and computational economy but also to the fact that it can efficiently manage the different ability levels as a control variable, thus allowing the researcher to assess and describe how the relationship between the group variables and item responses is modified by the presence of another categorical variable with *j* levels (i.e., levels of the trait or characteristic that is measured by the test). Thus, this technique compares the performance of an item in both the reference (*R*) group and the focal (*F*) group by considering the different trait or ability levels of the matching variable. It is assumed that on each one of these levels, the individuals of both groups are comparable, and thus if an item shows no DIF, their performance on it would be equivalent; in other words, they would have the same probability of a correct response for all trait levels. When applying the MH as a DIF technique, the observed score on the test is usually established as the matching variable. To analyze differential functioning in an item by means of the MH, examinees are divided into *j* ability levels, typically based on their observed test scores. The standard MH procedure then requires the construction of as many 2×2 contingency tables, that is, (groups to be compared, *R* and *F*) × (item response levels, 0 and 1), as the number of *j* levels into which the matching variable has been divided. The 2×2 contingency table for an item at ability level *j* is shown in Table 2, where n_{R1j} and n_{R0j} denote the number of examinees at the *j*th ability level who belong to the *R* group and who answered the item *correctly* (1) and *incorrectly* (0), respectively. Similarly, n_{F1j} and n_{F0j} refer to the same information but in the F group. $N_{..j}$ is the total number of subjects at the *j*th level of the matching variable, and $N_{R.j}$ and $N_{F.j}$ indicate the total number of subjects at the *j*th level for the *R* and *F* group, respectively. Finally, $N_{.1j}$ and $N_{.0j}$ refer to the number of subjects who answered correctly and incorrectly, respectively, at the *j*th level of the matching variable.



	Score on the Item		
Group	Correct (1)	Error (0)	Total
Reference (*R*)	n_{R1j}	n_{R0j}	$N_{R.j}$
Focal (*F*)	n_{F1j}	n_{F0j}	$N_{F.j}$
Total	$N_{.1j}$	$N_{.0j}$	$N_{..j}$

The null hypothesis of no DIF postulates that the probability of responding correctly to a given item at ability level *j* is the same for both the reference and focal group, whereas the alternative hypothesis (DIF present) states that this probability is different. These hypotheses are normally operationalized in terms of the *OR* (α_{MH}), which is defined as follows:

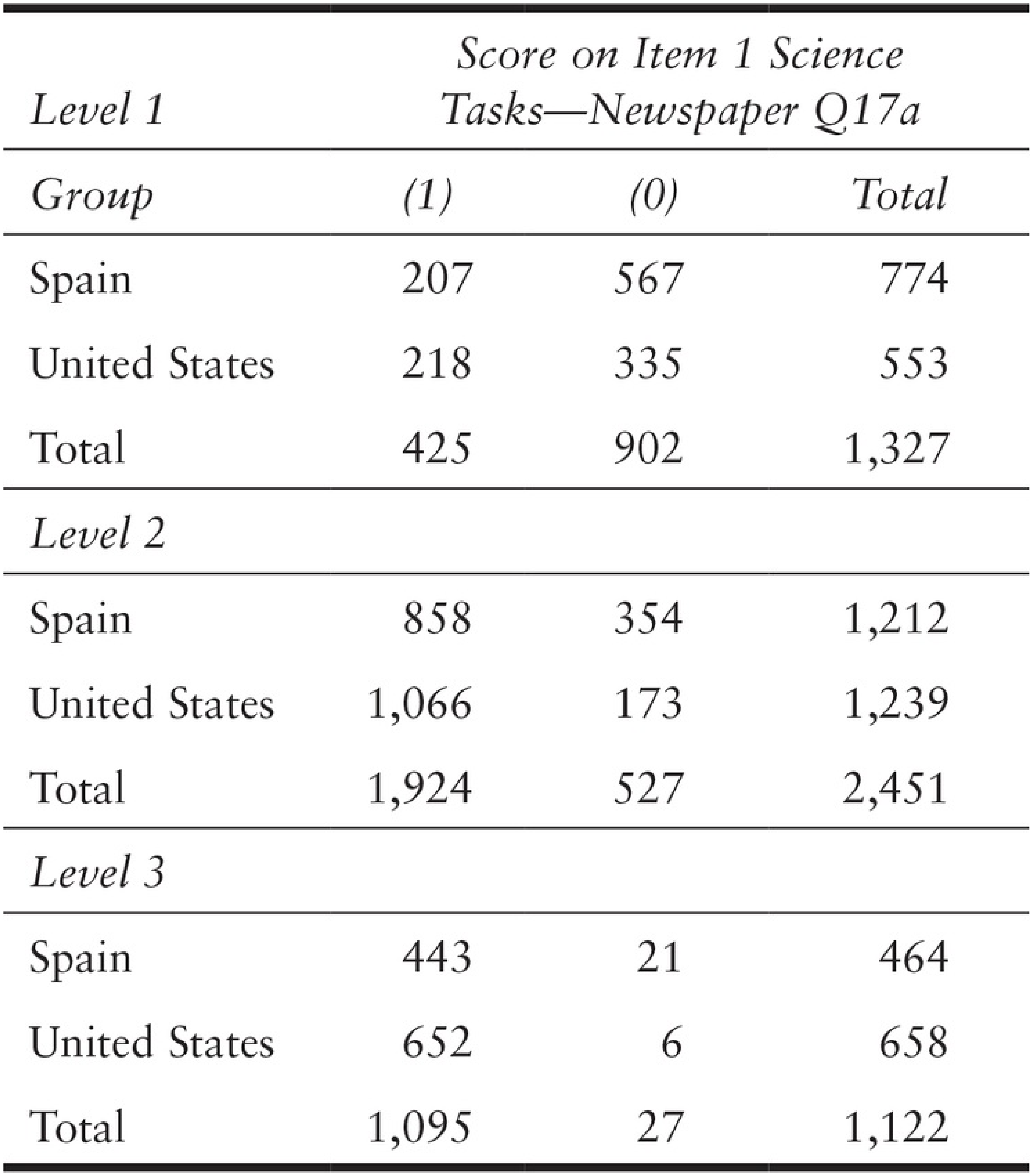
$$\alpha_{MH} = \frac{\sum n_{R1j} n_{F0j} / N_{..j}}{\sum n_{R0j} n_{F1j} / N_{..j}}$$

and which may take values from 0 to infinity.

In terms of the *OR*, the null hypothesis would be represented by an α_{MH} value of 1, whereas the alternative hypothesis would yield an α_{MH} different from 1. If α_{MH} is greater than 1, this would indicate that the *R* group is more likely to respond correctly to the item than is the *F* group; by contrast, a value below 1 indicates that the *F* group has an advantage over the *R* group.

An Example

In order to illustrate how the MH works for DIF detection, we will use data obtained from the PISA database. Specifically, we will analyze data corresponding to the science self-efficacy scale from the 2006 PISA Student Questionnaire. This scale contained 8 items with a 4-point Likert-type item response. Subjects were selected from samples pertaining to the United States and Spain, 2,450 participants from each country. The participants from Spain were all 16 years old, while those from the United States were either 15 or 16 years old. Due to the polytomous nature of the 2006 PISA Student Questionnaire items, the items were recoded into two categories: Thus, *do easily* and *with some effort* responses were coded as 1, and *struggle on own* and *couldn't do it* as 0. DIF analysis was then carried out using country as the group variable, with Spain being the reference group and the United States the focal group. Three levels of the matching variable were considered: Level 1, scores at or above 21 in the science self-efficacy scale; Level 2, scores between 14 and 20; and Level 3, scores at or below 13. Table 3 shows the 2 × 2 contingency table for Item 1 of this scale at each score level.



Level 1	*Score on Item 1 Science Tasks—Newspaper Q17a*		
Group	*(1)*	*(0)*	*Total*
Spain	207	567	774
United States	218	335	553
Total	425	902	1,327
Level 2			
Spain	858	354	1,212
United States	1,066	173	1,239
Total	1,924	527	2,451
Level 3			
Spain	443	21	464
United States	652	6	658
Total	1,095	27	1,122

For the data shown in this table, MH = 114.53, which is a statistically significant result, thus indicating that we should reject the null hypothesis of no DIF. Furthermore, α_{MH} = 2.25, a value that, by being greater than 1, indicates that the Spain group is more likely than the United States group to respond with “1” to this item.

Final Remarks

Final Remarks

The MH statistic has become widely used as a test of independence or as a measure of association, and this entry has described its formulation and main applications. In particular, we have focused on its use as a technique for detecting DIF, as this is undoubtedly the aspect of educational research to which the MH statistic is most applicable. Studies of DIF have become a routine and obligatory procedure when the aim is to develop a new measurement instrument, to revise an existing one, or to adapt a test into other languages and/or cultures. In fact, the educational field is one of the areas that is showing a growing interest in issues related to DIF. This is understandable if one bears in mind the negative repercussions that a lack of measurement equivalence can have in terms of equal opportunities and social justice. The MH statistic has a key role to play in this context, and it has become the technique of choice for detecting DIF that is used by the Educational Testing Service. It is also still used as the gold standard in comparisons of statistical efficacy (in terms of power and classification errors) when other more recent techniques are used to detect DIF.

Finally, it should be noted that the MH statistic can be calculated using easily accessible statistical software, because its implementation is included in packages such as SAS, SPSS, and R.

Juana Gómez-Benito and M. Dolores Hidalgo

See also Categorical Data Analysis; Differential Item Functioning; Effect Size; Hypothesis Testing; Odds Ratio; Programme for International Student Assessment; Two-Way Chi-Square

Further Readings

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential item functioning. Mahwah, NJ: Erlbaum.

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of Mantel-Haenszel procedure for detecting DIF: A meta-analysis. Psychological Methods, 18(4), 553–571.

Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), International encyclopedia of education (3rd ed.). Oxford, UK: Elsevier—Science & Technology.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129–145). Hilldale, NJ: Erlbaum.

Mantel, N. (1963). Chi-square tests with one degree of freedom, extensions of the Mantel-Haenszel procedure. American Statistical Association Journal, 58, 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of National Cancer Institution, 22, 719–748.

Organisation for Economic Co-operation and Development. (2006). PISA 2006 database. Retrieved July 20, 2010, from http://pisa2006.acer.edu.au/downloads.php

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. The American Statistician, 40, 106–108.

W. Holmes Finch W. Holmes Finch Finch, W. Holmes

Marginal Maximum Likelihood Estimation Marginal maximum likelihood estimation

1011

1013

Marginal Maximum Likelihood Estimation

Maximum likelihood estimation is one of the backbones of statistical analysis. It is used to obtain parameter estimates for a wide variety of models, including regression, factor analysis, and item response theory (IRT) analyses, among many others. When these estimates are based on data that are only marginally or partially observed, the procedure is called marginal maximum likelihood estimation (MMLE).

This article uses IRT analyses, a context in which the MMLE strategy is most common, to describe the assumptions, mathematics, and procedures of MMLE. The various models contained within the IRT family allow researchers to obtain estimates of an individual’s level of a latent trait of interest, typically referred to as θ, as well as information about the items, including their location on the θ scale (difficulty), their ability to differentiate individuals with different levels of θ (discrimination), and the likelihood that an individual would respond to an item correctly due solely to chance (pseudo-guessing). A distinct advantage of using IRT to estimate these quantities is that it is able to do so in such a way that item parameter estimates are independent of the individuals actually responding to the items, and person ability estimates are independent of the items that they complete. This property, along with the fact that item location and person ability are on the same scale, makes IRT a particularly attractive modeling tool for researchers working with item response data. It is the use of MMLE that supports these properties.

Maximum Likelihood Estimation

Maximum likelihood estimation procedures rest on assumptions about the data distribution of the population from which the sample was drawn. For example, in IRT, it is generally assumed that the latent trait of interest, θ, is normally distributed, although it is sometimes possible to relax this assumption. Consider a reading test made up of 20 multiple-choice items, each of which has a specific level of difficulty. Furthermore, let’s assume that the difficulty values are known. Given this information, one can use the item response pattern to estimate an examinee’s reading ability. If Examinees A and B both answer 8 of the 20 items correctly, each will have a raw score of 8. However, if the 8 items answered correctly by Examinee A were more difficult than the 8 items answered correctly by Examinee B, then the value of ability level, θ, for Examinee A should be estimated as higher than that of Examinee B. The ability to account for differences in the difficulty of items answered correctly by respondents is one of the great strengths of using maximum likelihood estimation in the context of IRT parameter estimation.

Maximum likelihood estimation works by finding the values of the model parameters that maximize a likelihood function for the item responses. For the reading test example, the item response pattern is simply the combination of 1s and 0s that reflect the correct and incorrect responses to the items. Thus, for the 20-item reading assessment, one possible response pattern would be 11010011101110001110. In short, the maximum likelihood approach can be thought of as a sophisticated search algorithm that seeks to find the combination of item and person parameters that as closely as possible reproduces the various item response patterns in the data.

MMLE

In the context of maximum likelihood estimation, we must know θ in order to estimate the item difficulty values, and we must know the item difficulties in order to estimate θ. A better algorithm would allow for the simultaneous estimation of the item and person parameters. One approach is *joint maximum likelihood estimation*, which can provide person and difficulty estimates independently of one another. However, it has been found that joint maximum likelihood estimation can produce biased parameter estimates and lacks computational efficiency. A better solution to this problem is MMLE. This technique has been shown to produce more accurate person and item parameter estimates than joint maximum likelihood estimation under most cases and also yields accurate estimates θ for individuals with extreme scores (e.g., all correct or none correct). Unlike joint maximum likelihood estimation techniques, which treat each of the individual-by-item responses as separate observations, MMLE is based primarily on only the individuals and assumes that likelihoods are random effects sampled from some larger distribution.

The MMLE algorithm begins with the assumption of a known distribution for θ, typically the standard normal, although others could be used. Given this latent trait distribution, item difficulty is estimated using the expectation-maximization algorithm with the following steps: 1. In the expectation (E) step, one calculates the expected number of examinees providing correct responses to the item at a given level of θ for a given item response pattern (e.g., 11000). One also calculates the expected number of examinees at each level of θ.

2. In the maximization (M) step, one attempts to maximize the likelihood of the function in order to obtain an estimate of *b* (difficulty) for the item.

Once there are difficulty estimates for each item, one can then obtain the θ estimate for each person in one of two ways: (1) Use the expectation-maximization algorithm as just described but find the θ values that minimize the likelihood function or (2) use a Bayesian estimator. In this latter approach, a posterior distribution for θ is derived given the item difficulty estimates and prior information about the distribution of the latent trait (e.g., it is normally distributed with a mean of 0 and a standard deviation of 1). Both expectation-maximization and Bayesian estimates of θ are used in practice, with research supporting the use of the Bayesian approach, particularly when there are extreme scores in the sample.

W. Holmes Finch

See also Bayesian Statistics; Item Response Theory; Maximum Likelihood Estimation

Further Readings

de Ayala, R. J. (2009). The theory and practice of item response theory. New York, NY: Guilford.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum and Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Thousand Oaks, CA: SAGE.

Magdalena Bielenia-Grajewska Magdalena Bielenia-Grajewska Bielenia-Grajewska, Magdalena

Market Research

Market research

1013

1018

Market Research

Market research can be defined as a set of activities and approaches aimed at gathering information about companies as well as current and potential customers and their needs, possibilities, and expectations. Many of the research methods and tools central to educational research such as surveys, attitude assessment, interviews, and focus groups are also central to market research, and market research is used to investigate aspects of educational institutions and products. After further breaking down the definition of market and market research, this entry investigates various types of markets, stages of market research, and the two main sources used in market research. Additionally, the entry explores the quality/quantity dichotomy, perspectives on the place, time, and online/off-line aspects of market research, as well as discipline-oriented approaches and a model approach to market research.

The way market research is conducted depends on various determinants related to the market itself and the broadly understood environment. The multidisciplinary character of market research is also connected with the diversity of functions played by markets in the modern economy and other areas of life. First, a market is the place where goods and services are offered, prices are negotiated, and information is shared. It is the place where relations are established and different people meet to sell and purchase goods. In addition, a market does not exist in a vacuum; it responds to the changes taking place in the close and far environment, being determined by different factors. The term *market* thus fails easy categorization. It can be subcategorized by taking into account subject (what is offered) and locale (where the market takes place). The second criteria (where) was crucial in the past centuries, with a market anchored in a geographical sense to a given location; however, it does not bear the same importance in the 21st century. With the advent and development of the Internet, a market is also situated online, with no anchorage in a particular setting, and with broad access to it, regardless of geographical and time differences. Moreover, the approach to market can encompass functional perspectives. There are different functions of market research. One of them is to gain or sustain competitive advantage, offering products or services that are more often selected than the ones made available by competitors. Market also determines other spheres of life, supplying merchandise indispensable for the existence of a given industry. Taking into account the complexity of modern markets, market research is often induced by the need of companies wanting to know how to make their products and services more competitive on the market. Thus, specialized marketing agencies are hired to observe the performance of a company or the market to predict the future behaviors of customers. In their book *Market Research in Practice: An Introduction to Gaining Greater Market Insight*, authors Matthew Harrison, Julia Cupman, Oliver Truman, and Paul Hague elaborate on the possible applications of market research, which include segmenting markets; improving a brand position, customer satisfaction, and loyalty; achieving optimum pricing; operating on a new market; checking advertising effectiveness; introducing a new product; and reporting.

Types of Markets

Market research may involve studying the type of market, such as monopoly, oligopoly, duopoly, or perfect competition. Another distinction is taking into account the characteristics of a given market and the type of products offered on the market. *Homogeneous markets* focus on offering one merchandise, whereas *heterogeneous markets* provide a palette of products. Markets may also be studied through the perspective of its scope, distinguishing international and domestic markets. Because market research focuses on studying facts (data), relations, causes, participants, and results, it encompasses different methods and approaches. As Harrison and colleagues elaborate in their book, market research may also be conducted by taking into account the type of customer. A *consumer market* includes fast-moving consumer goods, financial, banking, and leisure markets. Taking into account the scope of consumer market, it is often researched by using quantitative methods focused on precise sampling and qualitative approaches aimed at gaining information on customer motivation.

Business-to-business market research, on the other hand, encompasses a relatively smaller number of respondents (calculated in hundreds or thousands, in comparison with millions as in the case of consumer studies). In addition, business-to-business markets are diversified, consisting of companies operating in different industries and varying in size.

Stage Perspective

Markets can be researched through the prism of stages. The first stage is coining experiments and selecting approaches and fields of analysis. An important stage in preparing the study is sampling. In market research, sampling is often used to show the tendencies for a studied group or population by observing a selected part of it. Types of sampling include snowball sampling, convenience sampling, quota sampling, and chunk sampling. The advantages of sampling include saving time and money, whereas the possible disadvantages include selecting improper samples and not being a representative of a given population. The second stage is conducting the experiment itself and gathering information. Data are acquired by observing people and their activities (e.g., how they choose products, their verbal and nonverbal behavior, how often they select products, and how much they spend on services) and monitoring market trends. The third stage is devoted to working on the gathered data. This stage is connected with distributing the results of research (e.g., publishing papers and presenting results at conferences). This stage also includes classifying exploratory and explanatory market research. *Exploratory market research* is often conducted when the research topic is not well established. Exploratory research facilitates the selection of efficient methods, approaches, and sampling as well as enhances the familiarity of the research subject. Explanatory studies, on the other hand, search for the relations between analyzed variables, trying to find an answer for a given phenomenon.

Source Perspective

There are two main types of sources that can be used in market research. *Primary market research* is devoted to gathering original data. Types of primary research may include, but are not limited to, questionnaires and interviews as well as experiments. On the other hand, *secondary market research* focuses on using data that have been gathered and is often distributed by somebody. An example may be desktop research, devoted to analyzing, for example, published papers on a studied topic or statistical analyses available in statistical yearbooks.

Secondary market research is used when the costs of primary research are too high for researchers or when the access to sources is limited for some reason. In many cases, both types of research are used, with primary market research used to forgo secondary market research.

Quality/Quantity Dichotomy

The approach to market research can be investigated through the perspective of quality versus quantity. In qualitative market research, a focus group is conducted, as the name suggests, utilizing a group of people who are asked about their opinions, beliefs, and feelings about a particular product. Their ideas are recorded by the person who conducts the research. In the case of market research focused on education, the members of a group can be asked about their feelings on a given educational offer. The advantage of this method is the ability to discuss a given idea or problem in a properly selected group. The disadvantages may range from the wrong selection of groups and their participants, artificial flows of interaction (e.g., connected with the setting participants are not familiar with), and the way the researcher perceives the group and its interactions. A similar technique is participant observation, with different types of participation exercised by researchers. Another research technique is a case study, concentrating on, for example, one selected company or product and investigating it in greater detail. Researchers also use questionnaires and interviews, conducted both online and off-line, encompassing such techniques as dichotomous questions and open questions, and so on.

As far as quantitative market research is concerned, such methods as mathematical models, factor analysis, regression analysis, and break-even analysis may be used in quantitative studies. Factor analysis is efficient in investigating different variables, narrowing the choice to a small number of variables that can be further investigated to observe some features of consumer behavior, such as their purchasing preferences. Another method, regression analysis, can be applied to measure and predict consumer behaviors, calculating the predicted responses by taking into account the current ones.

Place Perspective

Another classification includes the taxonomy of research depending on where it is conducted. *Desktop research* includes the type of studies conducted mainly by analyzing data and materials that have already been collected, whereas *field research* involves direct contact with participants, entities, and environments, usually in their natural setting. The ones used in situ are mainly observational techniques directed at looking at people, things, and phenomena in their natural locale. In the literature, overt and covert observation are described. When the observation is covert, those who are observed are not aware of the study. In overt observation, participants are aware that they are being observed. Another technique of observational character is mystery shopping. Mystery shopping is directed at measuring the quality of service offered to customers or checking if standards are met in everyday service. In most cases, the person who is supposed to check the given company/shop pretends to be a real customer and observes how products are displayed and offered and how communication with customers is conducted. The mystery shopper later inserts data into questionnaire forms to be analyzed by companies or marketing agencies.

Time Perspective

Markets can be studied through the prism of time. *Diachronic market studies* focus on observing the development and changes connected with markets through history. The historical perspective offers data on how markets changed during a given period of time. It provides the possibility to compare ages or shorter periods of history. On the other hand, *synchronic market studies* concentrate on markets in a given period of time, without taking into account how they evolved throughout years, decades, or centuries.

Online/Off-Line Perspective

Regarding online research, the growing role of the Internet in all spheres of life has also influenced the tools used in researching modern markets as well as the channels responsible for informing people about markets. Adapting the latter perspective, namely the one focusing on the ways information is distributed, such data dissemination outlets can include social online networking tools, websites, and so on. *Off-line market research* encompasses methods that do not require the Internet in conducting the research itself.

Discipline-Oriented Approaches

Taking into account the discipline perspective, the closest domain to market

Taking into account the discipline perspective, the closest domain to market research is marketing. It should be mentioned, however, that market research may benefit from such disciplines as neuroscience and linguistics.

Neuroscience

Neuroscientific tools can be used in researching markets. *Functional magnetic resonance imaging* (fMRI) is one of the most advanced techniques used to observe brain activity. During the experiment, a subject is asked to lie still in an MRI scanner for about 1 hour. During the first minutes of scanning, the anatomical scans of the brain are made. In the next stage, the subject performs some tasks and the machine records how given parts of the brain change during these activities. In this stage, the blood-oxygen-level-dependent signal is measured and later analyzed. In the case of educational research, fMRI may help facilitate more effective methods of teaching by showing how individuals react to a given educational stimuli. The discussed methods may also find application in the postlearning phase, checking, for example, the effectiveness of the learned structures and words. Although this method is very effective in different types of research, one of its key disadvantages is the high cost of the machine itself that makes the access to fMRI rather limited for research purposes. Because the equipment is of relatively large size, it cannot be used for in situ experiments; rather, subjects have to be asked to come to a laboratory where fMRI is available.

Other neuroscientific methods can be used in market research. One of them is *facial electromyography*. This technique offers data on one’s feelings, emotions, and attitudes as well as features of personality. In the *facial electromyography* technique, by attaching small electrodes to the subject’s face, the subject’s features and reactions are observed by measuring electrical impulses generated by facial muscles. This technique is used in market research, for example, in advertising, selecting the options that catch the attention of potential customers. A similar procedure is *galvanic skin response*. Also known as electrodermal activity or skin conductance, this technique is used to observe the psychological arousal of the subject, visible in, for example, the change of sweat glands on the skin. Similar to *electromyography, galvanic skin response* may observe emotions and reactions to a given stimuli. The growing needs of market research in different settings have resulted in technologically advanced and portable equipment that can be used by researchers in almost all types of conditions. One portable technique used in neuromarketing is eye tracking. In an experiment using eye tracking, a subject wears special equipment on the head (in most cases similar to glasses) that monitors and records a subject's gaze and eye movements. This device is used by researchers in shops to observe where a customer is most likely to concentrate one's attention and, consequently, which products the subject is likely to purchase.

It should be mentioned that neuroscience not only offers tools for market research but also provides approaches to look at the ways economies can be studied. An example of such approaches is the set of growth drivers for market research opportunities proposed by Ian Lewis and Simon Chadwick that stress new marketing research opportunities. *Left brain–oriented* research includes mega databases, advanced analytics, and marketing accountability. *Right brain–oriented* research encompasses consumer listening, online communities, as well as body and brain measurement. *Dual* research incorporates synthesis of information, shopper insights, and global and multicultural measures.

Linguistics

Language is a crucial element of the modern market. The performance of business entities in the 21st century is determined by the language itself and this relation can be investigated in different ways. One of the key notions in the discussion on the linguistic side of companies is treating language as the crucial determinant for companies' performance on the internal and external level. This dimension encompasses the increasing complexity in terms of languages used in organizations, as well as professional sublanguages, dialects, and genres. The mentioned multiplicity on the linguistic level results in organizational strategies that aim at respecting corporate linguistic rights, exercising effective linguistic policy, and facilitating organizational performance. Language is also an important factor of cooperation between the company and the broadly understood environment, including customers. Because the modern global market is not restricted in terms of geographical distance, companies may also operate in very distant countries that differ, among others, in terms of languages and codes of communication used by diversified stakeholders. Thus, one of the key notions shaping the external side of corporate performance is providing high-quality translation and localization of products and services.

Apart from the organizational dimension, the customer perspective is important in understanding the nuances connected with the modern market. The linguistic level of customer performance is manifold. First, one’s mother tongue and dialects play an important role in shaping one’s attitude toward products and services. Second, one’s linguistic competence (being especially crucial in multilingual environments) determines one’s selection of market offers. Third, different nonlinguistic factors shape one’s attitude toward the offered merchandise. Thus, the linguistic dimension of consumers is shaped by the performance of organizations and individuals, and how the dialogue between customers and companies is affected by the selection of proper linguistic tools. Taking the mentioned complex relation between language and the market, the approaches used in linguistics can also prove useful in researching the communicative dimension of modern markets. One option is to apply one of discourse theories. For example, critical discourse analysis studies the functions of text understood in the broad sense, taking into account not only words but also sentences, texts, pictures, drawings, and other elements that shape the text. Other approaches used in linguistics that can be incorporated into market research are such theories as ethnography of communication, studying communication in a broader, and social context.

Marketing

Marketing, being an interdisciplinary field itself, offers different methods and approaches that can be used in investigating modern markets. For example, such tools as an SWOT analysis or the marketing mix can be applied in market research. Taking the example of an SWOT matrix, strengths, weaknesses, opportunities, and threats of a given business venture are estimated that facilitate the strategy building for a given enterprise. Harrison and colleagues approach the notion of market research by using the Ansoff matrix, analyzing such notions as new products and existing products as well as new markets and existing markets. Taking into account existing markets, market research may help evaluate the possibility of adopting new products. As far as existing products are concerned, market research is used to estimate customer satisfaction and increase competitive advantage. By analyzing new markets, market research may help explain unmet needs and unknown markets. Moreover, Harrison and colleagues discuss the application of market research in the product or service life cycle.

Taking into account the axis of time and demand, market research is used in the youth stage to elaborate on the unsatisfied needs for new products and facilitate the understanding of the predicted demand. As far as practical applications are the understanding of the predicted demand. As far as practical applications are concerned, market research can be used to estimate prices for products and services and draw specifications for products. In maturity, market research is used to build a brand. Customer satisfaction is elaborated by looking at such notions as building strengths and limiting weaknesses. By analyzing the old age of products, market research shows how to “breathe new life” into products and find new markets. Harrison and colleagues also elaborate on the notion of the four Ps in market research (product, price, place, and promotion). As far as product is concerned, market research can be used to check attitudes to products by, for example, displaying them in a focus group or a test hall. Regarding price, market research is used to measure the value individuals attach to products and, consequently, estimate the optimal price for merchandise. With regard to place, the distribution of products is important, thus market research facilitates finding the best venues for offering merchandise. Promotion benefits from market research by stimulating effective advertising and testing products.

Model Approach

Magdalena Bielenia-Grajewska developed the 6S Model of Methods in Consumer Studies, with such notions as stage, setting, sense, stimulus, stakeholders, and scope, to offer a complex approach of investigating consumer studies. The proposed model can also be used to investigate market research, thus its elements are described as follows. Stage encompasses, for example, the pre-and postpurchasing behaviors, including the actual purchase as well. Market research may also be investigated through the phase perspective, paying attention to the preparation stage, conducting experiments, gathering data, and writing publications. As far as setting is concerned, market research takes into account varied situations and locations where this type of research can be conducted. The loci can also be studied through the perspective of online and off-line markets, studying, for example, the characteristics of a given place. Analyzing stakeholders is another key aspect of market research. Such topics as the relation between customers and products may be studied, as well as the efficient ways of reaching potential clients (e.g., by using social online networking tools). In modern times, characteristics of many stimuli—sense—is important in market research. Thus, the tactile, auditory, or visual aspects of markets shape the way products and companies are perceived by stakeholders. It helps understand how new customers can be reached and how competitive advantage may be gained, by, for example, focusing on multimodal communication or simultaneously engaging different senses.

Stimulus may be understood as the inducement used in research. For example, verbal, auditory, olfactory, or pictorial notions are applied to check a given market hypothesis. The example may be the mentioned neuroscientific market research, involving fMRI scanning with verbal stimuli used to trigger brain reactions and provide knowledge about market behaviors. Scope of market research is connected using micro, meso, or macro perspectives. For example, market research may focus on a local, regional, and national level.

Magdalena Bielenia-Grajewska

See also Discourse Analysis; Focus Groups; Interviews; Participant Observation; Snowball Sampling

Further Readings

Bielenia-Grajewska, M. (2013). International neuromanagement: Deconstructing international management education with neuroscience. In D. Tsang, H. H. Kazeroony, & G. Ellis (Eds.), The Routledge companion to international management education (pp. 358–373). Abingdon, UK: Routledge.

Harrison, M., Cupman, J., Truman, O., & Hague, P. N. (2016). Market research in practice: An introduction to gaining greater market insight. London, UK: Kogan Page.

Lewis, I., & Chadwick, S. (2012). New roles for marketing researchers. In R. J. Kaden, G. Linda, & M. Prince (Eds.), Leading edge marketing research (pp. 2–22). Thousand Oaks, CA: SAGE.

Adam Michael Johansen Adam Michael Johansen Johansen, Adam Michael

Markov Chain Monte Carlo Methods Markov chain monte carlo methods

1018

1021

Markov Chain Monte Carlo Methods

Markov chain Monte Carlo (MCMC) methods use a carefully constructed sequence of dependent random variables to approximate expectations with respect to probability distributions of interest (i.e., appropriately weighted averages). Although first developed in the 1940s by physicists at Los Alamos, NM, the interest in these methods within the field of educational research is a consequence of their broad applicability to statistical inference. In many modern statistical problems, it is necessary to either find the maximum of a complicated function (as in maximum likelihood estimation) or compute high dimensional integrals (as in many forms of Bayesian inference). Monte Carlo methods provide approximation schemes suitable for addressing both of these problems when analytical or simpler numerical schemes fail; MCMC methods are perhaps the most broadly used class of Monte Carlo methods for dealing with the complex probability distributions that arise in the context of Bayesian inference for complicated models. This entry provides a short review of the basics of Monte Carlo methods, Markov chains, and common MCMC algorithms; a short discussion of some important practical considerations; and concludes with some applications of these methods within educational research.

Monte Carlo methods invert the usual statistical paradigm. In statistical inference, it is common to attempt to infer the properties of a population from an available sample. The Monte Carlo method, instead, constructs an artificial population whose properties coincide with those of a system of interest, such as a Bayesian posterior distribution, and then obtains a large sample from that population in order to approximate its properties.

The simplest Monte Carlo methods proceed by drawing simple random samples from the synthetic population. However, for most distributions of interest, this task is itself intractable. MCMC methods instead construct a Markov chain for which the population of interest is an invariant or stationary distribution and appeal to so-called ergodic theorems.

Markov Chains: The Basis of MCMC

A *stochastic process* is a sequence of random variables indexed by time. This entry only considers the case in which the time index takes discrete values, which is by far the most common setting within MCMC methods. A Markov chain is, informally, a stochastic process that has the defining property that *given the current value of the process, the future is independent of the past*. That is, they are stochastic processes that have no memory; if you know their value at any given time, then the earlier history of that process has no influence on its future behavior.

There are a number of additional properties that are important when Markov chains are used in the MCMC setting. A probability distribution is an *invariant* or *stationary* distribution for a Markov chain if, once the marginal distribution of the chain at a given time coincides with that probability distribution, the marginal distribution at all subsequent times also coincides with that distribution. A Markov chain is *irreducible* if it can reach all important parts of the state space from any other part of the state space ("important" here means with positive probability under the invariant distribution). It is *aperiodic* if it is not possible to divide the space explored by the chain into several distinct subsets that the chain moves between in a deterministic, periodic order. It is *recurrent* if it will, on average, return to all important states infinitely often if allowed to run forever. All of these properties are important in the context of MCMC methods, as they prevent particular types of pathological behavior.

If a Markov chain has a distribution of interest and has a forgetting (or ergodic) property, then states separate by a long enough time period behave similarly to independent samples from that invariant distribution. Note that this is different from the defining property of Markov chains, which states that given a particular value, later values have no dependence upon the past, a conditional independence property; in contrast, the forgetting property tells us that without any conditioning, the distribution of the chain at a time $t + s$ is close to independent of that at time t for sufficiently large values of s, without requiring any conditioning. Under such conditions, it can be shown that averages taken over the collection of values obtained by simulating the Markov chain can be used to estimate expectations with respect to its stationary distribution consistently and that a central limit theorem holds for these estimators. Formalizing these results is technical, but one key point is that the variance in the central limit theorem depends critically upon the autocorrelation of samples: the more quickly the Markov chain forgets, the smaller the resulting variance.

Although the theory underlying MCMC methods is now somewhat developed, it is generally difficult to obtain precise quantitative statements about the accuracy of estimates obtained from finite samples in realistic problems. For this reason, it is important to empirically assess the performance of such algorithms, making use of so-called *convergence diagnostics*, noting that these diagnostics are sometimes able to detect problems but do not provide guarantees that an algorithm has converged. Such diagnostics attempt to establish that a chain has explored the full support of the distribution of interest and that it forgets its history fast enough for a chain of the length that has been simulated to be adequate for the inferential task at hand.

Common MCMC Algorithms

The main challenge with implementing MCMC algorithms in practice is constructing a Markov chain that has the distribution of interest as its invariant distribution and that forgets its past quickly enough for it to be useful. Some common approaches to the construction of a Markov chain with a particular invariant distribution are summarized in this section.

The Gibbs sampler is one of the most intuitive such schemes and can be employed whenever it is possible to simulate from the conditional distribution of any variable within the distribution of interest given a particular value for all of the other variables (often termed the *full conditional* distributions). One simple form of such an algorithm simply updates one component of the vector of random variables at a time by sampling from its full conditional distribution, given the current values of the remaining components.

In many real settings, the full conditional distributions of the distribution of interest are not available, and Gibbs sampling and related algorithms cannot be employed. The Metropolis-Hastings algorithm provides one generic and widely employed scheme that does not require access to these distributions. The basic idea is to use a proposal distribution (which, formally, is almost arbitrary but in practice strongly influences the performance of the algorithm) and a carefully constructed rejection mechanism (in which some proposals are rejected with the chain remaining at its current state and other proposals are accepted with the chain moving to the proposed value) in order to ensure that the chain has the desired stationary distribution.

More sophisticated algorithms have been developed that employ gradient information and other properties of the target distribution where it is available. Such methods include the Metropolis adjusted Langevin algorithm and Hamiltonian (or hybrid) Monte Carlo, both of which make use of ideas from physics to provide Markov chains that might better explore the support of the target distribution than the simple methods previously described. Although these methods can lead to better performance in difficult scenarios, and extend the range of problems that can be adequately addressed, they are somewhat technical and the details of their operation are beyond the scope of this entry.

In the context of generic optimization schemes, the *simulated annealing* algorithm and its variants are probably the most commonly used Monte Carlo schemes. In contrast to standard MCMC algorithms that aim to provide samples that allow the approximation of expectations with respect to a particular distribution, these algorithms aim to provide a simple point estimate of the location of the minimum of a particular function. Simulated annealing is motivated by a physical analogy with annealing in metals: a process by which the temperature of a molten metal is gradually lowered, allowing its constituent atoms to arrive in a low-energy configuration in which they become trapped at low temperatures. In the simulation context, this process is mimicked by employing standard MCMC kernels, each of which has an invariant distribution that is increasingly concentrated around the maximizer of the function of interest.

Practical Considerations

As it is desirable that the correlation between successive values in a Markov chain employed in Monte Carlo simulations is small, it is natural to consider *thinning* samples obtained in this way. That is, rather than storing every sample that is generated, only every *k*th sample is stored in order to reduce the dependence of a sample on its predecessor. It can be shown that this can only increase the variance of the resulting estimator and cannot be justified on grounds of statistical efficiency. However, in some settings, the storage of many samples is prohibitively costly and the loss of statistical efficiency can be more than compensated for by the reduced storage requirements in this setting.

As Markov chains employed in Monte Carlo settings are not initialized in equilibrium (i.e., the starting value is not drawn from the stationary distribution), they take some time to forget their initial conditions. Asymptotic results show that this will not harm the estimator in the large sample limit, but in order to reduce the bias resulting from out-of-equilibrium initialization in the finite sample case, it is common to discard some number of samples at the beginning of the simulated chain. Assessing how many samples to discard in this way is something of an art but can be guided by examining plots of the Markov chain to look for transient behavior and by employing certain convergence diagnostics. Although convergence diagnostics can in isolation never prove that a given chain has converged to stationary or is suitable for a particular purpose, they can allow the diagnosis of a number of failure modes and they should always be used to establish at least the absence of any obvious failure to converge.

Although the implementation of complex sampling algorithms may seem to present a technical barrier to the widespread application of these methods, software to facilitate their use by nonspecialists does exist. WinBUGS and Stan were both in widespread use as of early 2016.

Use of MCMC in Educational Research

The principal use of MCMC methods within the field of education is in the statistical analysis of complex models. Such models abound in educational research in which it is often desirable to model hierarchical relations; there are often large numbers of latent (unobserved) quantities, and there can be complex relationships between model variables. There is a natural synergy between Bayesian methods and Monte Carlo schemes and, in particular, Bayesian analysis of hierarchical models in which there may be missing data. Considerable structural complexity is made possible by the application of MCMC techniques to compute expectations with respect to the associated posterior distributions. Such methods have been widely applied to multilevel psychometric models such as those of item response theory. In these and other complex modeling settings, Bayesian inference facilitated by MCMC methods provides a flexible framework in which inference can be conduced and uncertainty characterized.

Adam Michael Johansen

See also Bayesian Statistics; Graphical Modeling; Item Response Theory; Maximum Likelihood Estimation; Monte Carlo Simulation Studies

Further Readings

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (Eds.). (2011). Handbook of Markov chain Monte Carlo. Boca Raton, FL: CRC Press.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (in press). Stan: A probabilistic programming language. Journal of Statistical Software.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). Markov chain Monte Carlo in practice. Boca Raton, FL: Chapman & Hall.

Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 270(4598), 671–680.

Levy, R., Mislevy, R. J., & Behrens, J. T. (2010). MCMC in educational research. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), Handbook of Markov chain Monte Carlo. Boca Raton, FL: CRC Press.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data and rated responses. Journal of Education and Behavioural Statistics, 24(4), 342–366.

Robert, C. P., & Casella, G. (2004). Monte Carlo statistical methods. New York, NY: Springer Verlag.

Seltzer, M., Novak, J., Choi, K., & Lim, N. (2002). Sensitivity analysis for hierarchical models employing level-1 assumptions. Journal of Education and

Behavioural Statistics, 27(2), 181–222.

Sinharay, S. (2004). Experience with Markov chain Monte Carlo convergence assessment in two psychometric examples. Journal of Education and Behavioural Statistics, 29(4), 461–488.

Spieghalter, D., Thomas, A., & Best, N. (2000). WinBUGS user manual. Version 1.3. London, UK: Imperial College.

Ser Hong Tan Ser Hong Tan Tan, Ser Hong

Gregory Arief D. Liem Gregory Arief D. Liem Liem, Gregory Arief D.

Mastery Learning

Mastery learning

1021

1025

Mastery Learning

Mastery learning, which is also known as learning for mastery, is a form of instructional practice pioneered by Benjamin S. Bloom in the 1960s and 1970s. At its core, mastery learning is based on the belief that individual differences exist in learning needs and styles. Students' predisposition to learn is also perceived to be malleable, and it is believed that all students can learn well and attain task mastery under favorable instructional conditions.

With the use of flexible time frames, teachers accommodate their instructional practices and adopt diverse methods to give students as much time as they need to fully comprehend the classroom material. This stands in contrast to the more rigid traditional teaching practices in which all students are taught in the same way and are given the same amount of time to learn the material. At the end of a fixed period, students' performances are assessed with summative assessment with little or no follow-up actions to address the problems students face in learning.

This entry first examines the importance of a constructive alignment among learning objectives and subsequent instructional techniques and assessments. Next, instructional practices for mastery learning are thoroughly examined. Benefits and concerns of mastery learning are then reviewed. Finally, contemporary developments in instructional practices are considered.

Constructive Alignment

Educational programs leveraging mastery learning have to adopt the principles of constructive alignment between the class learning objectives, teachers' instructional practices, feedback through formative assessment, corrective activities, and evaluation tasks about students' competence. In other words, learning objectives that reflect mastery in each learning unit must first be clearly defined, followed by instructional practices and grading systems that correspond to the same mastery standards. Constructive alignment is reflected in both horizontal and vertical curriculum alignment.

Horizontal curriculum alignment requires that the same learning objectives and materials covered in classrooms are being tested in assessments. In a similar vein, the materials tested in assessments are necessarily those that have been taught in classrooms and specified in the learning objectives.

On the other hand, vertical curriculum alignment speaks of the congruence in level of understanding between what is outlined in the learning objectives, taught in class, and assessed. Accordingly, levels of understanding are defined using the revised Bloom's taxonomy. The six levels outlined in the revised Bloom's taxonomy ordered by the complexity of cognitive skills are as follows: (1) remembering, which refers to retrieving information stored in memory, (2) understanding, which refers to making sense of knowledge such as through giving examples, (3) applying, which refers to using knowledge or executing procedures appropriately, (4) analyzing, which refers to drawing connections between different parts of knowledge, (5) evaluating, which refers to critiquing knowledge, and (6) creating, which refers to putting information together to form new knowledge.

Instructional Practices for Mastery Learning

In the pedagogy of mastery learning, teaching topics are broken down into smaller learning units that teachers focus on one at a time. Learning objectives and standards reflecting mastery are defined at the start of each learning unit. A typical instructional cycle begins with classroom delivery of the unit's content by the teacher, then includes formative assessment, corrective or facilitative activities, and a second formative assessment before moving on to the next learning unit.

The cornerstone of mastery learning lies in its view that learning is a process, not an outcome. Feedback in mastery learning primarily takes the form of formative assessment, which is characterized by nongraded tasks reflecting the learning objectives of each learning unit. Formative assessment helps to track students in their learning progress. Examples of formative assessment include written assignments, in-class presentations, and graphic organizers in which students can draw and put in place the knowledge they have acquired.

Regular formative assessment not only serves as a diagnostic tool but also helps to motivate students in their learning. Specifically, formative assessment is usually administered upon the completion of each learning unit. Its primary role is to identify what information students have mastered as well as where students’ weaknesses lie. Subsequently, students are encouraged to work on their weaknesses with additional support by the teachers who specifically target students’ problems. Formative assessment also helps to affirm students who have shown mastery in the learning unit that they are learning well, both in terms of content knowledge and in their method of study. For these students, they become more confident of themselves and less anxious about their academic progress.

It is not the intention of formative assessment to serve as an evaluation tool that marks the conclusion of a learning unit; the purposes of formative assessment are to specifically identify students’ level of understanding at present as well as their areas of weaknesses. Students’ progress are compared against the desired learning outcomes of the specific learning unit. Subsequently, teachers address students’ problems in learning through corrective activities to help narrow the gap between what students know at present versus what they have to learn.

Corrective activities are explicitly tailored to each student—they focus only on students’ problematic areas that surfaced from formative assessment. Accordingly, each student receives a personalized program in which teachers will address the student’s weaknesses in the learning unit. The remedial program gives students more time to work on their problems toward the desired learning outcomes.

Furthermore, corrective activities can either take the form of individually or group-based strategies. Teachers can make use of alternative delivery methods that better cater to students’ learning styles to focus on improving their weaknesses. Teachers can also involve students in other learning activities such as small-group collaborative learning and peer-to-peer helping sessions. The aim of corrective activities is to help students to better understand what is required to achieve the learning objectives of each learning unit to achieve task mastery achieve the learning objectives of each learning unit to achieve task mastery.

Such remedial activities help to ensure that students have first mastered the content of each learning unit, which may serve as building blocks for the next unit. Corrective activities are intended to work on more than the deficiencies in specific teaching materials; corrective activities also allow students to identify their own learning styles through alternative instructional methods, gain confidence and efficacy toward academic studies, and become independent learners.

Following corrective activities, another formative assessment is usually administered to check on students’ understanding. The second formative assessment is designed with the same learning objectives as the first, even though different questions are used. The level of difficulty in the second formative assessment should be of the same difficulty level as, if not more than, the first formative assessment. This way, any improvement in results could be attributed to the increase in students’ learning rather than an artifact of difficulty level of the second formative assessment.

Feedback from the second formative assessment could provide an indication of whether or not the corrective activities have been effective. Additionally, the second formative assessment gives students another opportunity to achieve academic competence. This encourages a *growth mind-set* in students, a term coined by Carol S. Dweck, which holds learning in high regard. A growth mind-set adopts the belief that knowledge can be developed through effort and to stand strong in the face of academic setbacks.

Although other students are working on corrective activities, students who have displayed competence at the learning unit engage in enrichment tasks to deepen their learning. Examples of enrichment tasks include special assignments chosen by students to further challenge themselves. This way, mastery learning affords the flexibility to cater instructional practices to the learning needs of each individual student. The ideal results arising from mastery learning are therefore to have all students master knowledge, where possible, and to reduce the discrepancy between students’ achievement performances.

Where evaluation is needed, criterion-referenced scoring is used in mastery learning. Accordingly, students are graded with reference to predefined standards to determine whether they have achieved the specified levels of mastery as outlined in the learning objectives. The purpose of criterion-referenced scoring is to evaluate students’ competence in terms of how well the referenced scoring is to evaluate students' competence in terms of how well the materials have been learned. Students are assigned the corresponding grades when their performances meet the predefined standards, and the grading is irrespective of other students' performances. Criterion-referenced scoring has the added advantage that it defines benchmarks of task mastery that students can potentially achieve when given the necessary support. This further promotes intrinsic motivation in students in their learning journeys.

Benefits and Concerns

Studies and research have shown that mastery learning can lead to extensive benefits in students' learning. Direct improvements to students' examination performance have been documented across subject areas at varying educational levels in different countries. Students also show longer retention for the knowledge learned through mastery learning. In addition to improvements to students' cognitive outcomes, mastery learning has the potential to create a supportive and empowering environment for all to learn. This may lead students to have higher self-confidence and self-concept toward their academic studies. At the same time, students may develop more positive attitudes and affect toward school, which can be observed from higher participation rates in classroom activities.

The use of mastery learning in the classrooms may also bring about two concerns. First, it may seem that the outcome of minimizing disparity between students' performances would lead all students to learn the same amount of knowledge and have the same achievements. Such an outcome is not ideal, as it means stifling the potential of students with higher abilities. A more accurate interpretation is that mastery learning allows for the opportunity and resources that all students need to actualize their potential. Every student is provided with the required support to achieve task mastery as specified in the learning objectives of each learning unit. As for students with higher ability, they could take on more challenging assignments while other students work on corrective activities.

Second, mastery learning may create the impression that corrective activities take up a lot of curriculum time. This raises a concern over the loss of materials being taught in class due to reduced curriculum time. Corrective activities may be more time consuming in the initial stages as students work on these activities under teachers' close supervision in class. However, over time, students would be able to grasp how corrective activities work and complete them as homework outside of class. More curriculum time would ultimately be freed up as corrective activities can be completed outside of class and students build a stronger foundation to work on subsequent learning units.

Contemporary Developments

Combined with other instructional methods, mastery learning continues to be implemented in present-day classrooms. Cooperative mastery learning is one example that combines cooperative learning with mastery learning. Specifically, cooperative learning has small groups of students coming together to discuss and assist each other in learning tasks. The merging of the two instructional methods helps to address limitations in each method. While some group members may not learn as well or achieve the desired level of competency in cooperative learning, cooperative mastery learning helps monitor each student’s progress toward learning outcomes through individual assessment. Additionally, cooperative mastery learning can help to reduce anxiety and encourage students to work on corrective activities as a group rather than on an individual basis. Cooperative mastery learning has the potential to yield greater positive results in students’ learning and motivation as compared to the use of mastery learning or cooperative learning as a stand-alone strategy.

Mastery learning is not synonymous with mastery goal structures in classroom and mastery achievement goals. Although mastery learning refers to a pedagogical practice, mastery goal structures in classroom refers to a set of situational cues in the classroom that encourage students to adopt mastery achievement goals. In mastery goal structures, learning is a process that values effort and diligence. Mistakes are part and parcel of the learning process and success is described in terms of improvements. In contrast, individuals who adopt mastery achievement goals are motivated to achieve competence defined in terms of the task objective standards and task mastery. Although they refer to distinct entities, mastery learning, mastery goal structures in classroom, and mastery achievement goals can potentially complement one another in promoting students’ learning and growth.

Looking ahead, mastery learning could be adopted in conjunction with other instructional practices to better address students’ learning needs that may change over time. The ever-evolving educational landscape calls for educators to retain the essence of adaptive educational programs and combine them with new innovative methods to teach students to learn in the dynamic world.

Ser Hong Tan and Gregory Arief D. Liem

See also Bloom’s Taxonomy; Classroom Assessment; Criterion-Referenced Interpretation; Formative Assessment; Formative Evaluation

Further Readings

Block, J. H. (1980). Promoting excellence through mastery learning. Theory into Practice, 19, 66–74. doi:10.1080/00405848009542874

Bloom, B. S. (1968). Learning for mastery. Evaluation Comment, 1(2).

Bloom, B. S. (1976). Human characteristics and school learning. New York, NY: McGraw-Hill.

Bloom, B. S. (1978). New views of the learner: Implications for instruction and curriculum. Educational Leadership, 35, 563–574.

Guskey, T. R. (2005). Formative classroom assessment and Benjamin S. Bloom: Theory, research, and implications. Paper presented at the American Educational Research Association, Montreal, Canada.

Kulik, C. C., Kulik, J., A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. Review of Educational Research, 60, 265–299. doi:10.3102/00346543060002265

Schunk, D. H. (2000). Learning theories: An educational perspective (3rd ed.). Saddle River, NJ: Merrill Prentice Hall.

Laura M. B. Kramer Laura M. B. Kramer Kramer, Laura M. B.

Matching Items

Matching items

1025

1027

Matching Items

Matching is a test item type where test takers can demonstrate their ability to connect ideas, themes, statements, numbers, expressions, or solutions with supporting evidence, definitions, equivalent expressions, and so forth. Elements of the item are traditionally presented in two columns or lists, and each element in one list is paired with at least one element from the other list. This entry further describes matching items, gives examples of several types of matching items, and discusses issues with scoring certain types of matching items.

Matching items have traditionally been a quick and efficient way to ask a series of related questions without being redundant. This item format is also very compact—two lists of 10 items take up far less paper than 10 individual multiple-choice questions—and is easy to lay out using even a basic word processing program. Because of the ease of both presentation and scoring, this is a common item type used in teacher-made classroom tests. From a more technical standpoint, an advantage of a matching item is that the test developer can obtain multiple responses about related content without compounding problems of local dependence.

For example, a test may ask two multiple-choice questions such as

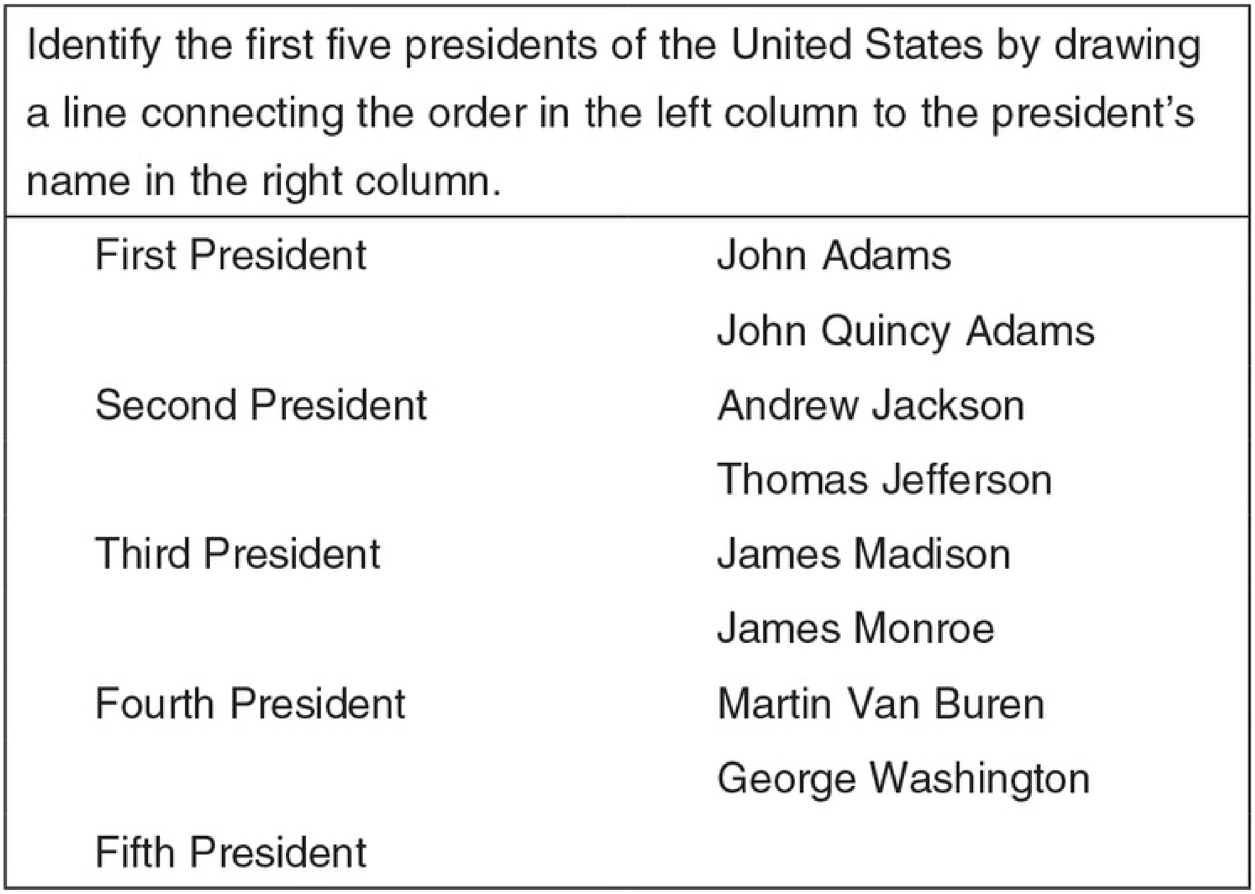
1. Who was the first president of the United States?
    1. Herbert Hoover
    2. James Madison
    3. Theodore Roosevelt
    4. George Washington

2. Who was the fourth president of the United States?
    1. John Adams
    2. Thomas Jefferson
    3. James Madison
    4. George Washington

These items’ answer choices make the items locally dependent. If an examinee correctly chooses D or incorrectly chooses B for Question 1, that selection will change the probability of selecting response options C and D in Question 2. Specifically, if a test taker chose D for Question 1, then Question 2 effectively has only three answer choices because George Washington could not be first AND fourth president, giving a one in three chance of guessing correctly if the examinee didn’t know the early presidents. The corollary is that if a test taker chose B for Question 1, Question 2 again effectively has only three answer choices, but now the probability of guessing correctly is zero.

However, the problem of local dependence can be somewhat reduced by making a matching lines item. Matching lines is one possible format for questions of this type, where, for example, the examinee literally draws a line to connect a word in column A with its definition in column B, as shown in Figure 1.

Figure 1 Simple Matching Lines Item



Identify the first five presidents of the United States by drawing a line connecting the order in the left column to the president's name in the right column.

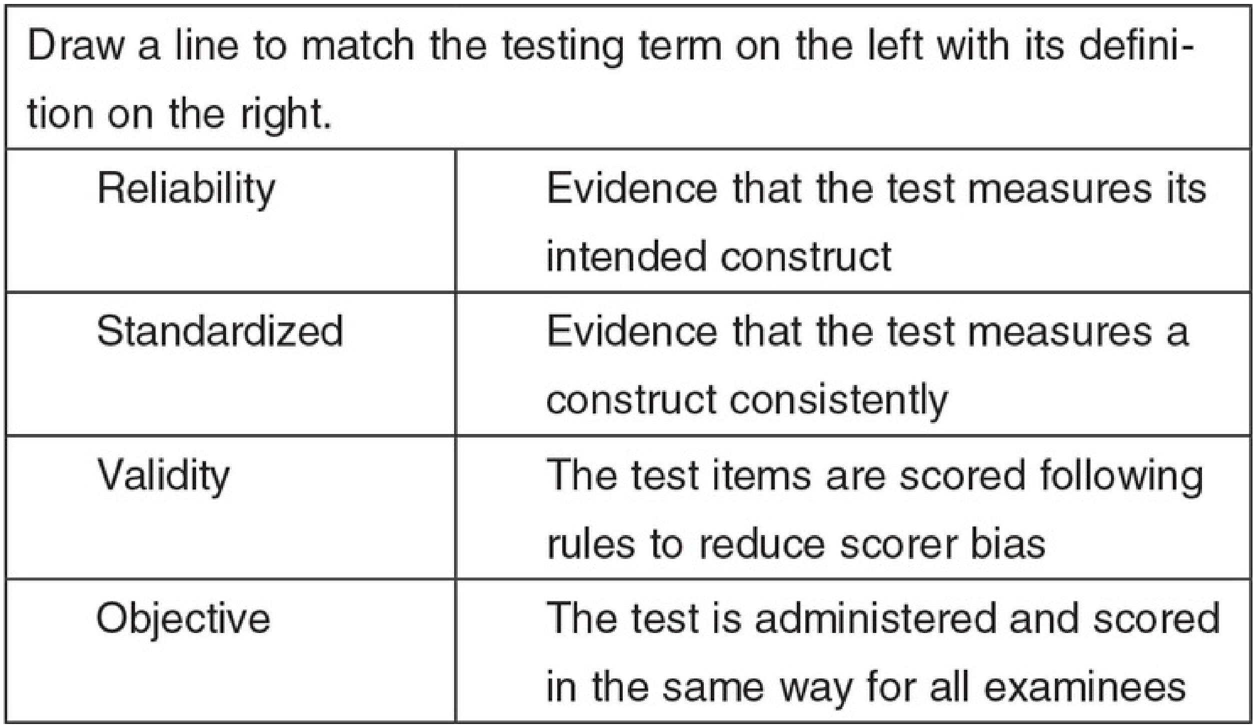
First President	John Adams
	John Quincy Adams
Second President	Andrew Jackson
	Thomas Jefferson
Third President	James Madison
	James Monroe
Fourth President	Martin Van Buren
	George Washington
Fifth President	

In this item type, the stem is an imperative statement rather than a question and should clearly direct the examinee how to answer the question. Ideally, there are more elements in the possible answers column than there are elements to be matched. In the example shown in Figure 1, the question asks for the first five presidents but gives eight choices. This is to prevent test takers from responding correctly through process of elimination rather than based on content knowledge.

However, this item format often requires hand scoring and is subject to errors in scoring depending on the perceptual acuity of both test taker and test scorer, the manual dexterity of the test taker, and the font size and page layout.

An example of a matching lines item where scoring might be challenging might look like that shown in Figure 2.

Figure 2 Matching item with difficult-to-interpret response



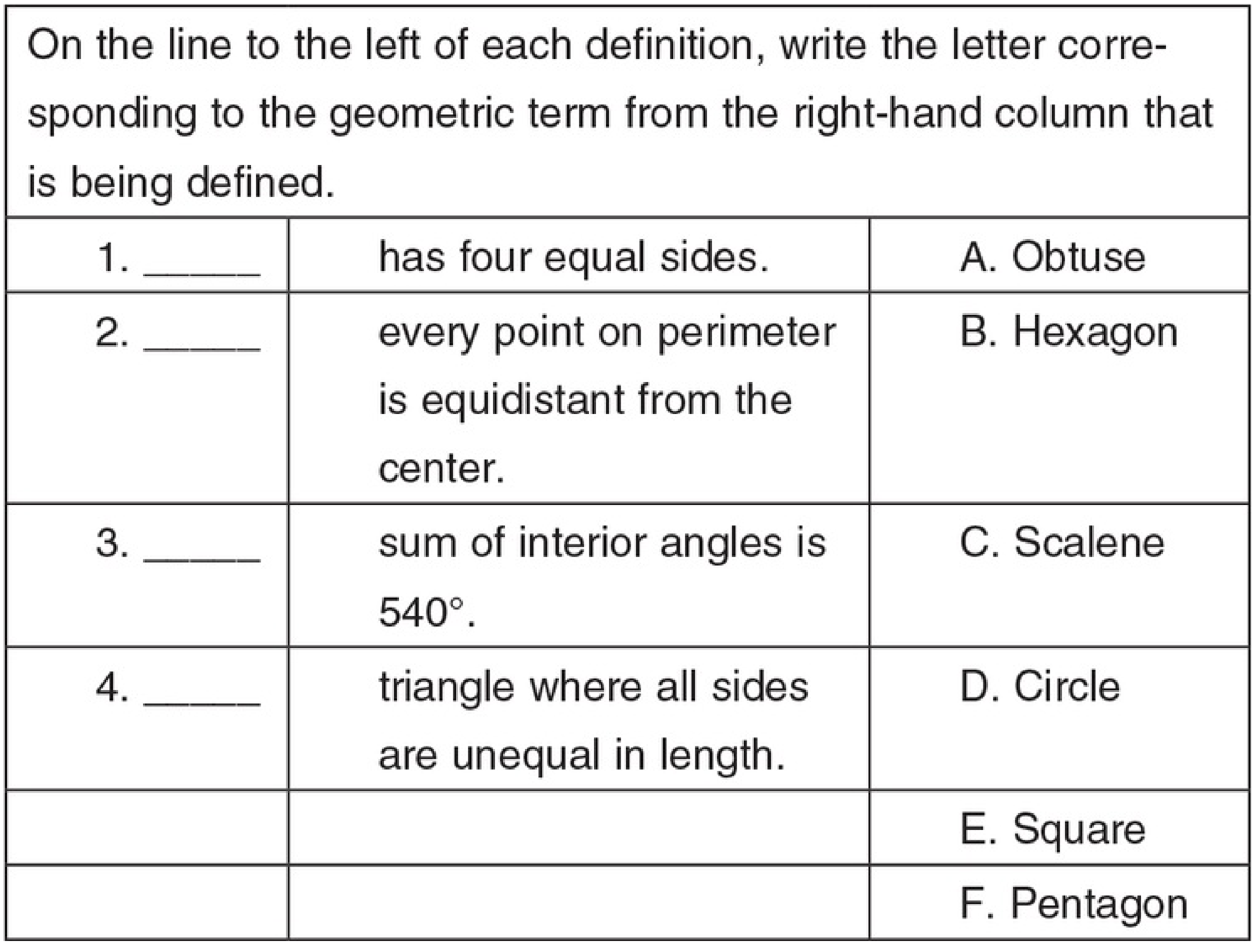
Draw a line to match the testing term on the left with its definition on the right.

Reliability	Evidence that the test measures its intended construct
Standardized	Evidence that the test measures a construct consistently
Validity	The test items are scored following rules to reduce scorer bias
Objective	The test is administered and scored in the same way for all examinees

This cagey test taker is uncertain about reliability and validity and hedges a bit by having the line from “reliability” end in the space between the two definitions and by having the line from “validity” shop short of the definitions, leaving it to the scorer to determine the trajectory of the line.

Another way to present a matching item to increase scoring efficiency is to have a list of “questions” in one column and a coded list of “answers” in the other column. A common convention is to have the “questions” numbered and the “answers” coded with letters. The test taker then writes the letter in a predetermined location next to its correspondent in the numbered column. Such an item might look like that shown in Figure 3.

Figure 3 Coded columns matching item



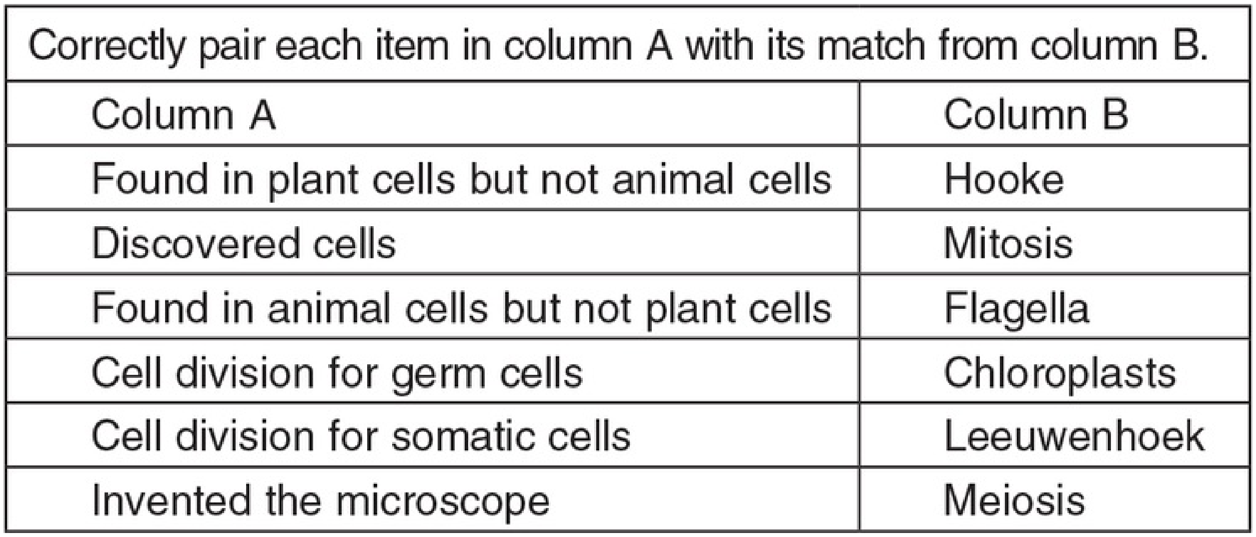
On the line to the left of each definition, write the letter corresponding to the geometric term from the right-hand column that is being defined.

1. _____	has four equal sides.	A. Obtuse
2. _____	every point on perimeter is equidistant from the center.	B. Hexagon
3. _____	sum of interior angles is 540°.	C. Scalene
4. _____	triangle where all sides are unequal in length.	D. Circle
		E. Square
		F. Pentagon

Each of the four statements in the left-hand column has a blank line in front of it, so the test taker can write in the letter corresponding to a geometric term from the right-hand column. There are many variations in this format. Scoring this format of a matching item is very fast and less prone to gaming (although again, cagey examinees might make a B and a D look as similar as possible and hope the scorer gives the benefit of the doubt).

Disadvantages to this type of item are that these items tend to measure content at a low level of cognitive complexity, such as the examples in this entry. Furthermore, although these items are easy to write, they are easy to write poorly. In addition to the concern about equal or unequal numbers of elements in each list, another common error in developing this type of item is using nonhomogenous content, as shown in Figure 4.

Figure 4 Matching item with nonhomogeneous content

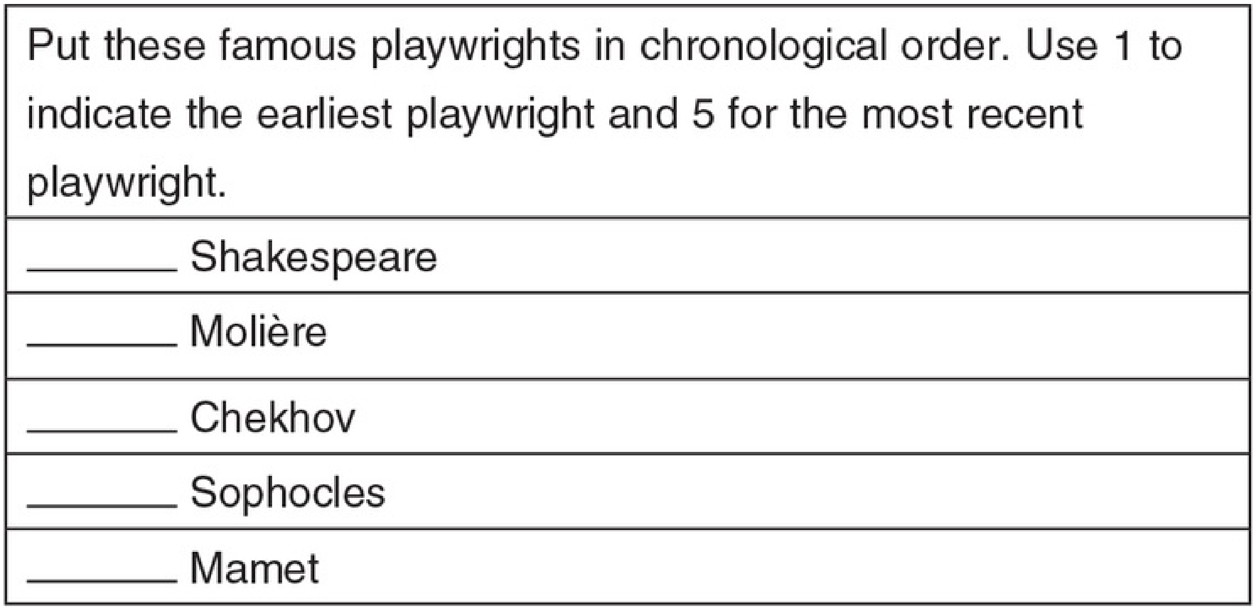


Correctly pair each item in column A with its match from column B.	
Column A	Column B
Found in plant cells but not animal cells	Hooke
Discovered cells	Mitosis
Found in animal cells but not plant cells	Flagella
Cell division for germ cells	Chloroplasts
Cell division for somatic cells	Leeuwenhoek
Invented the microscope	Meiosis

In Figure 4, there are really three pairs of elements in each column: two questions ask for names, two ask for cell structures, and two differentiate types of cell division. This item would be vastly easier than an item, for example, that asked test takers to pair six cell structures with their respective functions.

With the expansion of computer-delivered assessments, matching items can be readily machine scored. Matching lines items can be made less prone to gaming, and the issue of handwritten Bs looking like Ds goes away as well. Additionally, matching items lend themselves well to accommodation for students with disabilities who may use switch systems or other assistive technologies. Variations in matching items can be used in place of technology-enhanced items that use drag-and-drop response (which are inaccessible for students with motor impairments or who are blind or visually impaired), such as ordering, sorting, or categorizing. For example, a question may ask to put five playwrights in chronological order, which could be accomplished in either a drag-and-drop item interface or a matching interface similar to that shown in Figure 5.

Figure 5 Ordering item using matching



Put these famous playwrights in chronological order. Use 1 to indicate the earliest playwright and 5 for the most recent playwright.

_______ Shakespeare

_______ Molière

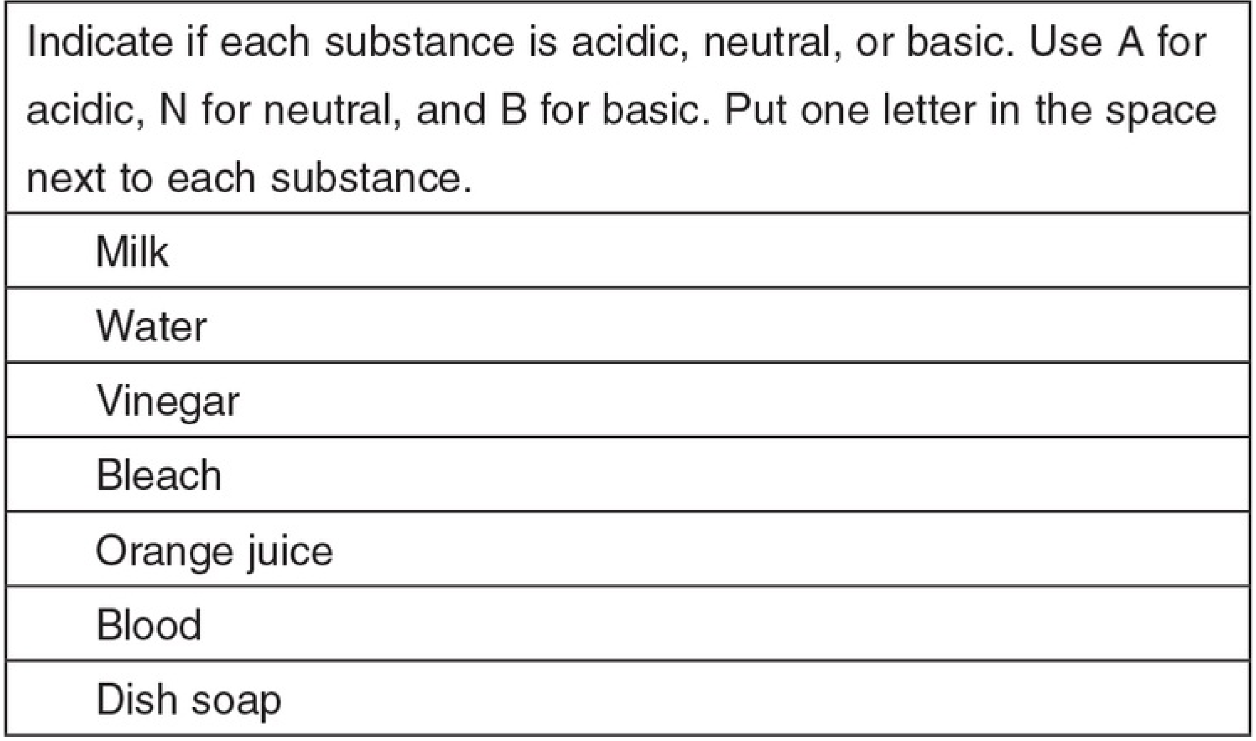
_______ Chekhov

_______ Sophocles

_______ Mamet

As a final example, an item might have students sort common substances into acids, bases, and neutral substances or place them on a scaled line representing pH. Again, this would not be accessible for students with certain types of disabilities or accessibility needs, so the item might be reworked as a matching type item as shown in Figure 6.

Figure 6 Sorting item using matching



Indicate if each substance is acidic, neutral, or basic. Use A for acidic, N for neutral, and B for basic. Put one letter in the space next to each substance.

Milk

Water

Vinegar

Bleach

Orange juice

Blood

Dish soap

Although matching items have been most commonly used in classroom assessments due to their parsimony and ease of hand scoring, recent developments in leveraging computer delivery for assessments makes this item type a viable choice for machine-scored, large-scale standardized assessment as well.

Laura M. B. Kramer

See also Local Independence; Multiple-Choice Items; Technology-Enhanced Items

Further Readings

Case, S. M., & Swanson, D. B. (2002). Constructing written test questions for the basic and clinical sciences (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Christoforos Mamas Christoforos Mamas Mamas, Christoforos

Matrices (in Social Network Analysis) Matrices (in social network analysis)

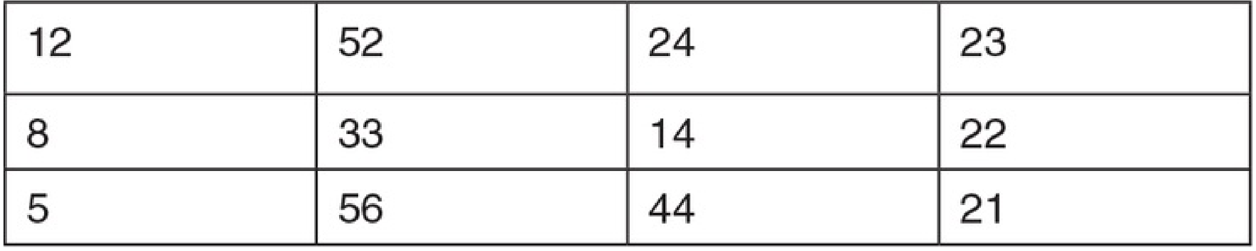
1027

1028

Matrices (in Social Network Analysis)

Generally speaking, a matrix is a rectangular arrangement of a set of elements or entries such as numbers or symbols that are arranged in rows and columns. The dimensions of the matrix in Figure 1 are three by four (3 × 4), as there are three rows and four columns. When discussing matrices, it is conventional to designate the number of rows as “m” and the number of columns as “n” and refer to the rows before the columns when describing the full size of a matrix. For example, Figure 1 displays a 3 (rows) × 4 (columns) matrix, which designates its full size. This entry describes the way matrices are utilized in social network analysis and defines common terminology such as ways and modes.

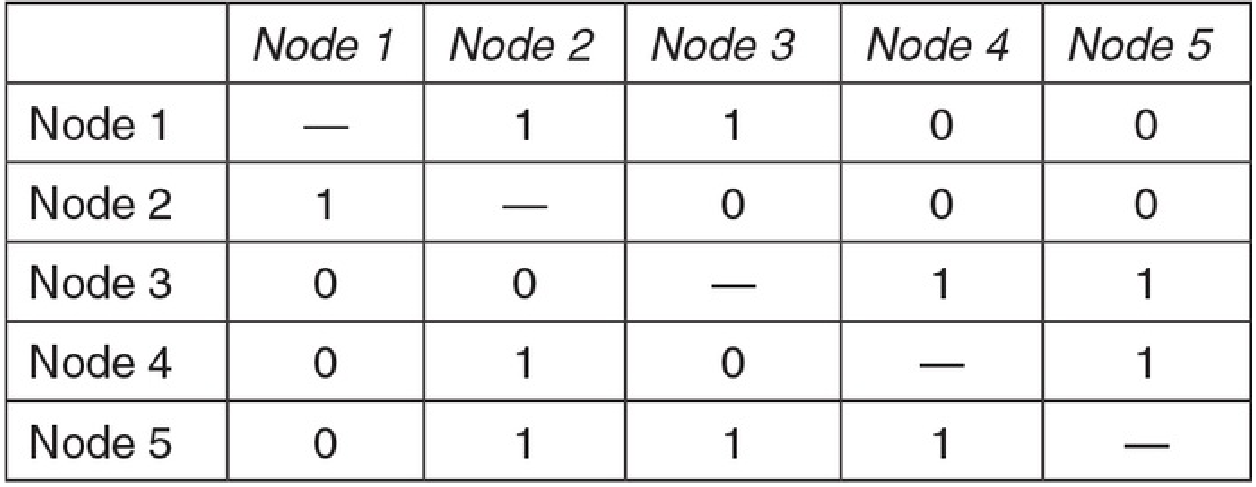
Figure 1 Example of a 3 x 4 matrix



12	52	24	23
8	33	14	22
5	56	44	21

In social network analysis, the most commonly used form of a matrix is the *adjacency* matrix. It is called an adjacency matrix because the entries indicate whether two nodes are adjacent or not. Most social network matrices are square with as many rows and columns as there are nodes in a data set. The elements or entries in the cells of the matrix record information about the ties between each pair of nodes. An adjacency matrix may be symmetric or asymmetric. For example, the matrix in Figure 2 represents a friendship network. The rows represent the source of directed ties, and the columns the targets. Node 1 nominates Nodes 2 and 3 as friends, but Node 3 does not reciprocate the friendship nomination. Therefore, this is an asymmetric matrix with directed friendship ties. If the ties represented in the matrix were undirected (e.g., ties representing the relation “is married to” or “talked to” where direction does not make sense), the matrix would necessarily be symmetric. The simplest matrix is binary, which means that if a tie is present, the numeral 1 is entered in a cell, and if there is no tie, 0 is entered. The first row and first column are not really parts of the matrix in Figure 2, but social scientists typically show their data as an array of labeled rows and columns for presentation purposes.

Figure 2 Asymmetric adjacency matrix



	Node 1	*Node 2*	*Node 3*	*Node 4*	*Node 5*
Node 1	—	1	1	0	0
Node 2	1	—	0	0	0
Node 3	0	0	—	1	1
Node 4	0	1	0	—	1
Node 5	0	1	1	1	—

Adjacency matrices of graphs are always square, as the one in Figure 2. They are also called one-mode matrices, as both the rows and columns refer to the same single set of nodes. However, in a two-mode matrix, the rows and columns refer to different sets of nodes. For example, imagine the nodes in the matrix rows of Figure 2 voting for different election candidates rather than selecting friends among one another. In this case, the columns would correspond to different candidates.

Matrices can be described as having ways and modes. The ways represent the dimensions of the matrix, such as when there are rows and columns, whereas the modes represent kinds of entities. A three-way matrix would then have rows, columns, and levels. Going back to the election example, suppose a researcher has data indicating which persons voted for particular candidates in different elections. This could be represented by a three-way, three-mode matrix, as in a data cube. Most studies, however, employ one-mode matrices that are the simplest to use. Overall, graphs of networks can be represented in matrix form, and mathematical calculations can then be performed to summarize the information in the graph that is useful in unpicking patterns of ties in social information in the graph that is useful in unpicking patterns of ties in social networks.

Christoforos Mamas

See also Matrix Algebra; Multitrait–Multimethod Matrix

Further Readings

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). Analyzing social networks. Thousand Oaks, CA: SAGE.

Daly, A. (2010). Social network theory and educational change. Cambridge, MA: Harvard Education Press.

Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods. Riverside: University of California. Retrieved from http://faculty.ucr.edu/~hanneman/

Mamas, C., Georgeson, J., & Kaimi, I. (2016). A mixed-methods approach to researching friendships and social interactions in mainstream schools in England and Cyprus. SAGE Education Case Studies.

John Joseph Dziak John Joseph Dziak Dziak, John Joseph

Matrix Algebra Matrix algebra

1028

1029

Matrix Algebra

Matrix algebra is vital for quantitative psychology, statistics, and computer science. It provides a compact way to express complicated mathematical operations. A matrix M is an array of numbers organized in rows and columns. The entry M_{ij} is the number in the *i*th row and *j*th column of M. Entries can be real or complex valued, but only real-valued matrices are considered in this entry. M may represent a data set, with participants or observations as rows and with variables as columns. This entry describes the basic operations of matrix algebra, how vectors are related to matrices, and common uses of matrix algebra.

Basic Operations

A matrix transpose (written $\mathbf{M}^{T}$ or **M**’) flips a matrix to exchange rows and columns, which can be written as $(M^{T})_{ij} = M_{ji}$. **M** is symmetric if $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$. Matrix addition or subtraction is defined for matrices of identical size and shape, and it adds or subtracts corresponding entries; thus $(\mathbf{A} + \mathbf{B})_{\mathrm{ij}} = \mathrm{A}_{\mathrm{ij}} + \mathrm{B}_{\mathrm{ij}}$.

A matrix can always be multiplied by a number: $(\mathrm{k}\mathbf{A})_{\mathrm{ij}} = \mathrm{kA}_{\mathrm{ij}}$. A product of matrices, **AB,** is defined when the number of columns of **A** equals the number of rows of **B** and consists of sums of cross-products of rows and columns: . **AB** need not equal **BA**, and they need not both be defined. An identity matrix **I** is a matrix such that for any **M** of the correct size, **MI** = **M** or **IM** = **M**. The identity matrix has a special diagonal structure: $\mathbf{I}_{ij} = 1$ for $i = j$ and 0 for $i \neq j$. While there is no matrix division in general, some square matrices **A** have a multiplicative inverse $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\,\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}$. Others do not have inverses; this happens if the matrix is not of full rank (i.e., if some row or column can be written as a linear combination of other rows or columns). This is called collinearity, rank deficiency, or singularity.

Vectors and Matrices

Vectors are matrices with only one row or column. A column vector $\mathbf{v}$ is often interpreted as the coordinates of a point in *r*-dimensional space and can be pictured as an arrow from the origin to the point. $\mathbf{Mv}$ represents some operation (e.g., rotating, stretching) on this $\mathbf{v}$, depending on the structure of $\mathbf{M}$.

The product $\mathbf{u}^T\mathbf{v}$ of two vectors is a single number, the sum of their cross-products. Thus, $\mathbf{v}^T\mathbf{v}$ is a sum of squares and is the squared length of $\mathbf{v}$'s arrow.

For a matrix $\mathbf{M}$, it is often possible to find vectors $\mathbf{v}$ such that $\mathbf{Mv} = \lambda\mathbf{v}$ for a constant λ. Then, $\mathbf{v}$ and λ are called an eigenvector and eigenvalue of $\mathbf{M}$. Usually, eigenvectors are scaled to be of standard length ($\mathbf{v}^T\mathbf{v} = 1$) when calculating eigenvalues. Eigenvalues and eigenvectors provide important information about $\mathbf{M}$ and are related to methods of decomposing (factoring) matrices into products of simpler matrices. Techniques such as principal component analysis are based on performing such a decomposition of the covariance matrix of a set of observed variables in order to study their interrelationships.

Common Uses

Matrices are often used to represent the coefficients of simultaneous linear equations, and inverses can be used in solving for their solution. For example, the normal equations defining the least squares solution of the regression coefficients β of $\mathbf{y}$ on $\mathbf{X}$ can be abbreviated as $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^{-1}\mathbf{y}$, so $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. If $\mathbf{X}^T\mathbf{X}$ is not full rank, β is undefined.

Software such as MATLAB (MathWorks) or the free package R (R Foundation for Statistical Computing) provides convenient functionality for matrix algebra, including finding inverses, eigenvalues, and decompositions.

John Joseph Dziak

See also Multivariate Analysis of Variance; R; SAS; SPSS

Further Readings

Harville, D. A. (1998). Matrix algebra from a statistician’s perspective. Technometrics, 40(2), 164.

Searle, S. R. (1982). Matrix algebra useful for statistics. Hoboken, NJ: Wiley.

Frederick Burrack Frederick Burrack Burrack, Frederick

Maturation

Maturation

1029

1030

Maturation

The *maturation effect* is defined as when any biological or psychological process within an individual that occurs with the passage of time has an impact on research findings. When a study focuses on people, maturation is likely to threaten the internal validity of findings. *Internal validity* is concerned with correctly concluding that an independent variable and not some extraneous variable is responsible for a change in the dependent variable.

Over time, people change and these maturity processes can affect findings. Most participants can, over time, improve their performance regardless of treatment. This can apply to many types of studies in the physical or social sciences, psychology, management, education, and many other fields of study.

Maturation Effects and Internal Validity

A number of maturation effects can occur during a short period of time, even within a few hours or days. How participants respond between pre-and post-data collection can change as a result of a good or bad mood at the time. Influences such as tiredness, boredom, hunger, and inattention can impact response. A participant may have had little rest prior to the data collection of a project, causing tiredness, or may be preoccupied by other responsibilities, causing inattention. These participant-based influences can be difficult to control and reduce the internal validity of findings.

Maturation effects that occur over the longer term include factors such as influences resulting from getting older, becoming better educated, changes in economic situations, and so forth. For example, participants who experience changes in their work expectations or in their financial status may respond differently irrespective of the intervention (independent variable) of the study. With particular populations, such as very young people or elderly people, small changes in age or situation can markedly impact physical, social, behavioral, and psychological response.

The issue to be questioned in a study is how confident one can be that the study can identify the observed changes in the dependent variable that are due to the treatment (i.e., intervention) and not due to maturation. Experimental design enables a researcher to be more confident that maturation is not responsible for change in the dependent variable. A simple experimental design can address maturation.

R O X O

R O O

The *R* represents participants randomly assigned to the treatment and control groups. The *O* represents the observation or measurement of the dependent variable. This design is used to achieve comparability with reasonable confidence that extraneous variables, such as maturation, are evenly distributed over both groups and controlled in this sense.

Frederick Burrack

See also Experimental Designs; Internal Validity; Random Assignment; Threats to Research Validity

Further Readings

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin.

Gall, M. D., Borg, W. R., & Gall, J. P. (2003). Quantitative research design. In Educational research: An introduction (7th ed.), pp. 287–431). White Plains, NY: Longman.

Tyler Hicks Tyler Hicks Hicks, Tyler

Maximum Likelihood Estimation Maximum likelihood estimation

1030

1032

Maximum Likelihood Estimation

Maximum likelihood (ML) denotes an important framework for estimation and inference. It is a theory about estimating models (i.e., recovering parameters from samples) rather than specifying models (i.e., constructing models). Its logic is premised on selecting the estimates that make the data the “most likely” (relative to the other possible estimates). Thanks to its versatility, ML is treated as the gold standard for estimating advanced models, such as multilevel and structural equation models. This entry provides an overview of ML estimation.

When explicating ML estimation, it is essential to keep distinct *models*, *algorithms*, and *theory*. A model defines the estimand (i.e., the parameter waiting to be estimated). An algorithm (or estimator) is the computational device used to obtain an estimate. A theory is the logical blueprint for building algorithms that yield warranted estimates for models. ML estimation is thus about estimation algorithms rather than models per se.

A likelihood can be defined as the conditional probability of the data given an estimate. The *likelihood lover’s principle* stipulates that modelers favor estimates assigning the highest likelihood to data. ML theory can take on a plurality of forms (e.g., full, restricted, robust) but likelihood lover’s principle unites them. Suppose two jars are full of marbles. There are eight red and two green marbles in the first jar and two red and eight green in the second jar. Suppose further a jar was selected and a red marble randomly drawn. One intuitively guesses the first jar was selected, and likelihood lover’s principle clarifies why this is a safe bet. The likelihood of a red marble, given the first jar, is 0.8; if it was the second jar, it is only 0.2. This logic exemplifies ML theory.

Likelihood functions are the building blocks of ML estimation. Input the data and such functions output their joint likelihood. Likelihood functions are not and such functions output their joint likelihood. Likelihood functions are not probability functions, but there is a one-to-one correspondence between these two types of functions. There are binomial, gamma, and normal likelihood functions, for example. To compute likelihoods for actual data, statisticians can use likelihood functions Education processes resemble natural lotteries; their outcomes, probability distributions. Only think about how IQ scores follow a normal curve for an example. Probability functions thus govern such distributions. Researchers working with educational data then need only select the matching likelihood function to compute likelihoods. Specifying the likelihood function is thus the first and foremost step in ML estimation.

An ML-based algorithm typically involves three more steps. All the steps can be briefly described as follows:

1. Construct a likelihood function. The model dictates the likelihood function (e.g., a normal likelihood function can be specified for data modeled as normally distributed).
2. Simplify the likelihood function and take its logarithms.
3. Take the partial derivative of the log-likelihood function with respect to each parameter and set the resulting equations equal to 0.
4. Solve the system of equations to find the parameters.

Solving systems of equations can be difficult when there are many parameters at play, even with the help of modern computers. For instance, implementing Step 4 may require the assistance of an algorithm (e.g., Newton-Raphson and EM algorithms). These algorithms use an iterative process of trial and error to converge upon a passable solution. As a caveat, there are no guarantees they will converge on the correct ML estimate or even converge.

Given that some minimal regularity conditions are satisfied, ML-based algorithms can be shown to yield estimates with desirable asymptotic properties. Suppose an infinite series of replication studies is conducted and a model repeatedly estimated with random data using an ML-based algorithm. The sampling density denotes the distribution of the ML estimates. As the sample size approaches the limits, the sampling densities of ML estimates will be (a) consistent (i.e., their peak settles on the truth), (b) normal (i.e., their shape approaches normalcy), and (c) efficient (i.e., their standard errors are as small as possible).

The well-understood sampling densities of ML estimates can sustain statistical inference. One can take the second derivative of the likelihood function to find inference. One can take the second derivative of the likelihood function to find the requisite standard error. This makes it possible to carry out hypothesis tests and confidence intervals for parameters. One can even derive fit statistics, such as likelihood ratio tests, to empirically build models using ML estimation. However, an important caveat is that likelihood ratio tests assume that compared models are nested (e.g., perhaps an extra parameter is added to the second model).

ML estimation can be fruitfully compared and contrasted with two alternate estimation theories, ordinary least squares (OLS) and Bayesian estimation. The goal of OLS estimation is to minimize misfit rather than maximize likelihood. Assuming population normalcy, OLS and ML estimation will yield equivalent estimates in linear regression. Yet, ML will outperform OLS in other modeling contexts (e.g., logistic regression).

Bayesian and ML are sister estimation theories. Bayesian estimation is foregrounded in the same likelihood functions as ML. Bayesian estimates thus share the same asymptotic properties as ML estimates. In other words, Bayesian and ML estimators will yield equivalent estimates given sufficiently large samples. However, a key difference between them is that ML estimators do not require users to specify a prior probability for the parameter.

When R. A. Fisher proposed ML estimation in the 1920s, he was ahead of his time. It was an impractical theory without modern computers to implement it. Estimation theories with friendlier algorithms, such as OLS, thus had a slight edge over ML in the past. This may be one reason why many introductory textbooks on regression and analysis of variance taught OLS rather than ML, even though ML is the more general method. In the 1960s, ML estimation finally became a viable option for most researchers thanks to the availability of computers. An understanding of ML estimation is thus a prerequisite for engaging in modern statistical modeling.

Tyler Hicks

See also Bayes’s Theorem; Bayesian Statistics; Estimation Bias; Inferential Statistics; Marginal Maximum Likelihood Estimation

Further Readings

Edwards, A. W. F. (1992). Likelihood (expanded ed.). Baltimore, MD: Johns

Hopkins University Press.

Eliason, S. R. (1993). Maximum likelihood estimation: Logic and practice. Newbury Park, CA: SAGE.

Myung, J. (2003). Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology, 47, 90–100.

MCMC

See Markov Chain Monte Carlo Methods

Christian Bokhove Christian Bokhove Bokhove, Christian

McNemar Change Test Mcnemar change test

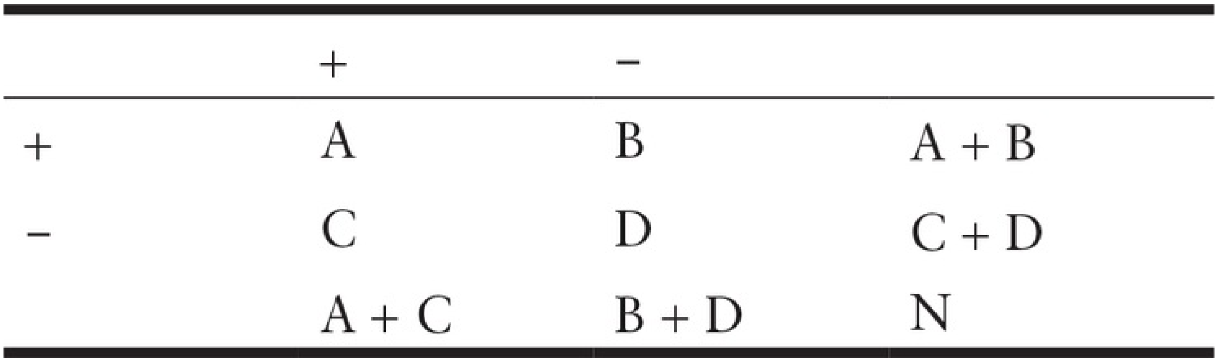
1032

1034

McNemar Change Test

The McNemar change test is a statistical test that can be used for paired nominal data. It can test differences on a dichotomous-dependent variable between two related groups. For dichotomous dependent variables, some like to think of it as similar to a paired *t* test. Typically, the test is used when researchers want to look at changes in participants’ scores by comparing the proportion of people who changed in one direction (e.g., an increase in test scores) to the proportion changing in the opposite direction (e.g., a decrease in test scores). It is a distribution-free (nonparametric) test. It can be used for pretest and posttest designs, matched pairs, and case-control studies. This entry further describes the McNemar change test and considers assumptions, characteristics, and applications of the test, concluding with examples.

The McNemar change test was first published in 1947 in the journal *Psychometrika* by Quinn Michael McNemar. The McNemar test is applied to 2 × 2 contingency tables, with a dichotomous variable and matched pairs of subjects. The test then determines whether row and column marginal frequencies are equal. This can be referred to as *marginal homogeneity*. Table 1 provides an example of such a 2 × 2 contingency table.



	+	−	
+	A	B	A + B
−	C	D	C + D
	A + C	B + D	N

The null hypothesis of marginal homogeneity states that the marginal probabilities for each outcome are the same (i.e., there is no difference)—the total rows are equal to the sum of columns. The mean of paired samples are equal and no (significant) change has occurred. The alternative hypothesis would state there is a significant difference—the total number of rows is not equal to the total number of columns, or that the paired sample means are not equal. Under the null hypothesis, if the frequencies in the cells B and C (discordants) are sufficiently large, has a chi-square distribution with one degree of freedom. Like other statistical tests, if the chi-square result is significant, the null hypothesis would be rejected. Most popular statistics programs like SPSS and R cater for the McNemar change test.

Assumptions for the McNemar Change Test

The McNemar change test makes several assumptions: 1. There is one categorical/nominal dependent variable with two categories, a dichotomous variable, and one categorical/nominal independent variable with two related groups. Examples are passing or failing a test (pass or fail), two groups (treatment A and treatment B), or stress-level groups (high and low).

2. The two groups of the dependent variable must be mutually exclusive and not overlap. It should not be possible that a study participant can be a member of both groups.

3. The participants should be a random sample of the target population.

Typically, the last assumption is the one that is violated most often.

Discussion of the McNemar Change Test

When the number of discordants (cells B and C in Table 1) is small (generally B + C <25), chi-square is not approximated well by the chi-square distribution any more. An exact binomial sign test would be more appropriate, where B is compared to a *binomial distribution* with size n = B + C and p = .5. In 1948, Allen L. Edwards provided a continuity corrected version of the McNemar test. Another option is the mid-p McNemar test (mid-p binomial test), which is calculated by subtracting half the probability of the observed b from the exact one-sided p value, then double it to obtain the two-sided mid-p value.

One characteristic of the McNemar test is that the elements of the main diagonal (A and D in Table 1) do not contribute to the decision about which treatment condition is more favorable. This means that the sum B + C can be small and statistical power of the test previously described can still be low even when the total N, A + B + C + D is large (see the example provided later in this entry).

There are several extensions and alternatives to the McNemar test that might mitigate some of these drawbacks.

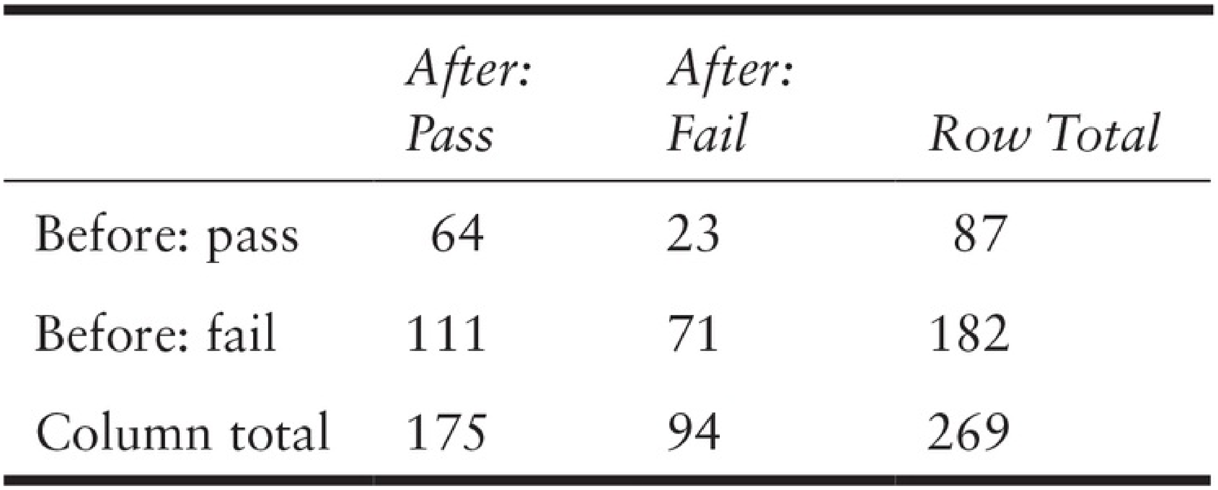
- The Cochran’s Q test is an extension of the McNemar’s test for more than two treatment groups.
- The Liddell’s exact test is another exact alternative to McNemar’s test.
- The Stuart-Maxwell test is a generalization of the McNemar test, which can be used for testing marginal homogeneity in a square table with more than two rows/columns. The Bhapkar’s test is an alternative to the Stuart–Maxwell test.

Applications and Examples

Examples of questions that can be answered with a McNemar change test are as follows:

- Is there a change in the proportion of voters prior to and following a presidential debate?
- Does the proportion of success versus failure significantly change after treatment?

Another example might be when an education researcher attempts to determine whether a new mathematics program has an effect on students’ achievement. Counts of students are given in Table 2, with the result of a math test (test: *pass* or *fail*) before the program given in the rows, and the result of a math test after the program in the columns. The test requires the same subjects to be included in the before-and-after measurements (matched pairs).



	After: Pass	*After: Fail*	*Row Total*
Before: pass	64	23	87
Before: fail	111	71	182
Column total	175	94	269

In this example, the null hypothesis of marginal homogeneity would mean there was no effect of the program. From the data, the McNemar test statistic would be , a test statistic of 57.79. This value yields a *p* value of *p* < .001, and therefore, the null hypothesis would be rejected: the program had an effect: .

Christian Bokhove

See also Chi-Square Test; Distributions; Hypothesis Testing; Inferential Statistics; *p* Value; Random Selection

Further Readings

Edwards, A. L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. Psychometrika, 13(3), 185–187.

Lancaster, H. O. (1961). Significance tests in discrete distributions. Journal of the American Statistical Association, 56(294), 223–234.

Liddell, D. (1976). Practical tests of 2 × 2 contingency tables. Journal of the Royal Statistical Society, 25(4), 295–304.

McNemar, Q. M. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157.

Rice, J. (2007). Mathematical statistics and data analysis (3rd ed.). Belmont, CA: Duxbury Press.

Jamie C. McGovern Jamie C. McGovern McGovern, Jamie C.

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Measurement Invariance Measurement invariance

1034

1037

Measurement Invariance

Measurement invariance is when a test or scale is found to measure the same construct in the same way across different groups of people. Measurement invariance is important for a number of reasons. First, this concept is closely related to test bias. When measurement invariance is not present, then the measure (e.g., an intelligence test, an academic achievement test, or a personality measure) in question may be biased because it is not functioning the same way across different groups, and people’s scores are being affected by their demographic grouping rather than just their true ability or possession of a particular quality. In an applied setting, biased measures can lead to poor assessments and inappropriate decisions regarding diagnoses and services. In a research setting, biased measures can mislead researchers and bring them to unwarranted or incorrect conclusions.

Measurement invariance is also important because the scores of people from different demographic groups should not be compared if measurement invariance has not been supported for a particular instrument. Researchers are often interested in differences between groups on different constructs, such as anxiety levels or academic achievement scores. Some examples of groups that are often of interest to researchers are gender, age, grade-level, socioeconomic status, ethnic, and cultural groups. Comparing the scores of groups such as these without first establishing measurement invariance of the instrument in question would be ill-advised. This is because score comparisons that occur in the presence of a lack of measurement invariance have little meaning due to the fact that the measure is not capturing the same construct in the same way across the groups of interest. Essentially, without measurement invariance, one would be comparing apples to oranges. The remainder of this entry will review methods comparing apples to oranges. The remainder of this entry will review methods for evaluating measurement invariance, namely, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and differential item functioning (DIF).

Methods for Evaluating Measurement Invariance

There are three main methods for investigating invariance. The first method is through the use of EFA. The second method is through the use of CFA. Bruno D. Zumbo, Stephen G. Sireci, and Ronald K. Hambleton suggest that CFA and EFA methods can be used in a complementary fashion when stricter tests of measurement invariance using CFA methods do not work. Under these circumstances, EFA methods may be helpful to use. A third main method to examine measurement invariance is DIF. DIF is a method to determine whether the responses on an item differ across groups when controlling for the attribute or latent variable being measured by the instrument.

EFA Methods

EFA methods allow researchers to determine how similar the structure of a measure is across groups. To explore factorial invariance with EFA methods, a researcher would first determine the factor structure for each group separately. This process involves an examination of the instrument for an underlying structure that is indirectly derived from many respondents’ item level data. EFA helps elucidate which items appear to measure the same underlying factors or constructs. For different groups, the same items may be most salient (or load) on the same or different factors. Once the factor structure is determined through an EFA for each group of interest, such as males and females, the factor coefficients of the items on the matched pairs of corresponding factors are compared. For example, if a measure consisted of two factors for males and for females and this was determined through conducting EFAs, then the factor coefficients of the items on the first factor for males would be compared to the factor coefficients of the items on the corresponding factor for females, and the factor coefficients of the items on the second factor for males would be compared to the factor coefficients of the items on the corresponding factor for females to see whether they are similar across males and females.

Indices such as the Pearson *r* correlation coefficient, the coefficient of congruence, and the salient variability similarity index may be calculated by researchers to determine how similar the matched pairs of corresponding factors are for the different groups. Pearson *r* correlation coefficient involves the computation of a correlation coefficient between the factor coefficients of the items on the matched pairs of corresponding factors for the different groups of interest. Values may range from −1.00 to +1.00, with values close to +1.00 suggesting similarity of the matched pair of corresponding factors across groups. Although this technique was popular in the past, Cecil R. Reynolds notes that it is not recommended for current use because the assumption underlying the Pearson *r* statistic that variables are bivariately normally distributed may be violated when one compares the factor coefficients of a matched pair of corresponding factors and the method used to correct this problem, such as conversion to Fisher *z*s prior to the calculation of a Pearson *r* correlation coefficient, may not work. The coefficient of congruence, a parametric statistic, directly compares the factor coefficients of a matched pair of corresponding factors across groups. Values of 0.90 or higher, derived from a statistical formula, are generally interpreted to mean that the matched pair of corresponding factors in question are similar across groups. Although the coefficient of congruence is a common method used to determine whether the matched pair of corresponding factors are similar across groups, it does have its limitations. Reynolds and Patricia A. Lowe suggest that its limitations include the use of the coefficient of congruence when the variances of the two groups are not equal, which may decrease the magnitude of the coefficient of congruence value, and the use of the coefficient of congruence when an orthogonal rotation procedure is used in performing the EFAs. The salient variable similarity index, a nonparametric statistic, indicates how similar the matched pair of corresponding factors are across the groups, with +1.00 indicating very high similarity and any negative value indicating dissimilarity and likely bias. Reynolds notes that the salient variable similarity index is often used in tandem with the coefficient of congruence, as the salient variable similarity index is not influenced by unequal variance-covariance matrices nor factor size.

CFA Methods

In contrast to EFA methods, CFA methods allow researchers to determine how different the structure and functioning of a measure is across groups. Using CFA methods to examine whether a measure is invariant across groups involves a multiple-step process in which more and more parts of the structural equation model are constrained, and at each step, fit indices and changes in those fit indices are examined to determine whether the fit of the model to the data has become worse. The first step is to run CFAs on each group separately and to find a model for each group that fits the data well. Model fit is determined by examining the fit index values. Common fit indices used to evaluate model fit include the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) in addition to the chi-square (χ^2) statistic. The CFI and TLI are considered to be incremental fit indices. The SRMR is viewed as an absolute fit index, and the RMSEA is also categorized sometimes as an absolute fit index. The range of values for the CFI, TLI, RMSEA, and SRMR range from 0.00 to 1.00, with CFI and TLI values close to 1.00 and RMSEA and SRMR values close to 0.00, suggesting a good model fit to the data. A nonsignificant χ^2 value would also indicate a good model fit. Different researchers have suggested different values for these different fit indices to serve as guidelines to indicate an adequate or a good model fit to the data. The second step is to test whether a multigroup model fits all of the groups of interest; if it does, then this is called configural invariance, and it means that the latent structure is the same across groups (i.e., the number of factors and indicators [e.g., items] on each factor are similar across the groups). The third step is to determine whether the corresponding factor loadings are similar across the groups, which is called weak factorial invariance. The fit of the weak factorial invariance model is compared to that of the configural model, and if the fit is not significantly worse, then weak factorial invariance is suggested. Researchers may use a nonsignificant change in the χ^2, a change in the CFI value or the RMSEA value, or a combination of changes in these fit indices between the more-and less-restricted nested models to determine whether invariance seems tenable across groups. Other researchers may use the RMSEA value of the alternate model and see whether it falls in the 90% confidence interval of the null model to determine whether invariance appears tenable across groups of interest. The fourth step is to determine whether the corresponding indicator (e.g., item) means or thresholds, dependent on which parameter estimator is used, are relatively equivalent across groups, which is called strong factorial invariance. To determine whether strong factorial invariance seems tenable, the fit of the strong factorial invariance model is compared to that of the weak factorial invariance model, using the same process described in the third step.

Some researchers advocate taking a fifth step to investigate whether strict factorial invariance is supported. In this fifth step, one would test whether the corresponding indicator residuals or error terms are similar across groups. This would be accomplished through comparing the fit of the strict invariance model to that of the strong invariance model, using the same process described in the third step. If strict factorial invariance is supported, it suggests that group differences in scores are due completely to actual differences between the groups in their amounts of the underlying construct being measured. However, researchers such as Todd D. Little maintain that demonstrating strict factorial invariance is unnecessary. As William Meredith notes, it is generally accepted that differences in groups’ scores on a measure can be meaningfully compared when strong or partial strong factorial invariance is supported for that measure. Partial strong invariance is when some items may not operate the same across groups.

Once strong or partial strong factorial invariance is supported, the observed scores of the groups of interests, such as males and females, on the measure can be compared using statistical tests, such as independent *t* tests, to determine whether differences exist between the two groups. However, the observed mean scores used in this analysis would have measurement error. Another option would be to compare the latent factor means between groups. The advantage of this second approach, which would involve performing a latent means analysis, is that measurement error would be removed from the latent variable and a true difference on the latent variable of interest could be examined. In latent mean analysis, the latent factor mean of one group (e.g., males) is constrained and the latent factor mean of the other group (e.g., females) is not. If a positive latent mean estimate results, then the group whose latent factor mean was not constrained (e.g., females) would have a higher level of the latent variable than the group whose latent factor mean was restricted (e.g., males). In contrast, if a negative latent mean estimate results, then the group whose latent factor mean was restricted (e.g., males) would have a higher level of the latent variable than the group whose latent factor mean was not constrained (e.g., females).

DIF

DIF is the third method commonly used for examining measurement invariance. DIF assesses item bias. DIF analyses, using such techniques as the Mantel-Haenszel technique, logistic regression, or structural equation modeling, compare the difference between two groups’ responses on individual items on a measure. Significant results of these comparisons indicate that the groups are responding differently on one or more items. When this occurs, it suggests that responding differently on one or more items. When this occurs, it suggests that the content of the items that are functioning differently needs to be examined more closely to determine if they are biased and should be removed from the measure.

In summary, when a test or scale is determined to measure the same construct in the same way across different groups of people, it can be said to possess measurement invariance. Exploratory and confirmatory factor analytical methods and DIF can be employed to investigate the measurement invariance of an instrument. Determining whether a measure is invariant across different groups is particularly important if a researcher is investigating test bias or comparing the scores of different groups.

Jamie C. McGovern and Patricia A. Lowe

See also Confirmatory Factor Analysis; Differential Item Functioning; Exploratory Factor Analysis; Latent Class Analysis; Test Bias

Further Readings

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research, 32, 53–76. doi:10.1207/s15327906mbr3201_3

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. Psychometrika, 58, 525–542. doi:10.1007/bf02294825

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), Handbook of cross-cultural neuropsychology (pp. 249–285). New York, NY: Kluwer Academic.

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), The handbook of school psychology (4th ed.), pp. 332–374). Hoboken, NJ: Wiley.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. Journal of Applied Psychology, 91, 1292–1306. doi:10.1037/0021-9010.91.6.1292

Zumbo, B. D., Sireci, S. G., & Hambleton, R. K. (2003, April). Re-visiting exploratory methods for construct comparability: Is there something to be gained from the ways of old. In Construct Comparability Research: Methodological Issues and Results, National Council on Measurement in Education symposium, Chicago, IL.

Nicole M. Nickens Nicole M. Nickens Nickens, Nicole M.

Measures of Central Tendency

Measures of central tendency

1037

1040

Measures of Central Tendency

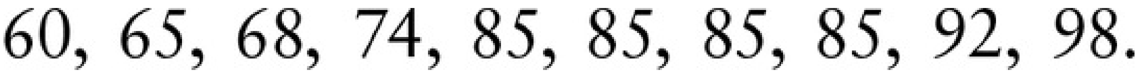
A measure of central tendency is a value that is typical of a data set. Any measure of central tendency is considered to be representative of a whole data set or distribution; so by itself, it provides a description or summary of the data set as a whole. A conceptual understanding of measures of central tendency is basic to understand many aspects of educational measurement and thus essential for inclusion in this resource book. This entry defines and describes the most commonly used measures of central tendency and presents some advantages and disadvantages of using each one. The mean, median, and mode are the most commonly used measures of central tendency. In addition, the relationship between specific measures of central tendency and the effect of the shape of the distribution on measures of central tendency is described.

Mean

The mean is the average of a set of values, calculated by adding all of the values together and dividing by the total number of values. The formula can be written as

$$\bar{x} = \sum x \, / \, n,$$

where ∑*x* is the sum of all of the values in the data set, *n* is the number of values in the data set, and is the mean or average of those values. For example, a teacher may want to determine the average test score on the first exam of the semester. For 10 students in the class, the data set (test scores out of 100 points) is



60, 65, 68, 74, 85, 85, 85, 85, 92, 98.

To find the mean, add all of the test scores together for a total of 743, and then divide by 10, which is the total number of scores. The average test score for Exam 1 is 74.3.

Advantages of Using the Mean

Because calculation of the mean uses every value included in the data set, it is a good representative value of the data. Note that in the example provided, the mean of 74.3 does not actually appear in the raw data; this is typical. Another advantage of using the mean is that it is resistant to differences or variations of data sets drawn from the same population. If the teacher in the example checks the average test score from that same group of 10 students for each exam given during the semester, it is likely that those averages will remain quite similar from exam to exam.

Disadvantages of Using the Mean

The primary disadvantage of using mean to represent an entire data set is that it is sensitive to extreme values or outliers. Extreme high or low scores can create a skewed distribution, meaning that outliers can pull the average one direction or the other and provide misleading information about the data set as a whole. This effect is especially strong when the data set is small. Therefore, the mean is not an appropriate measure of central tendency for distributions with outliers, particularly small distributions.

Median

The median is the middle value in a data set; it lies at the middle position when all the values in a set are arranged in an ascending or descending order. Because of its position in the data set, it literally divides the frequency distribution into two equal parts, resulting in half of the values in a data set lying at or above the median and half of the values lying at or below the median. Therefore, the median represents the 50th percentile in a distribution. To calculate the median, simply identify the value at the middle of the ordered distribution. If the number of values in a data set are odd, then (n + 1) / 2 value is the median. If the number of values in a data set are even, it is found by calculating the average of *n* / 2 and (*n 2 + 1) value. Note that the values in the data set are not part of the calculation, only the place the value holds in the ordered distribution. To find the median of the test scores provided earlier in this entry, you would use the calculation for the even number of values because there are 10 test scores. Therefore, the average of the value found at 10* 2 (the fifth value is 85) and the value found at 10 / 2 + 1 (the sixth value is 85) in the ordered set is the median.

60, 65, 68, 74, 85, 85, 85, 85, 92, 98.

The average of the two values is 85 and thus the median value.

Advantages in Using the Median

The median is easy to identify or compute and conceptually easy to understand. Unlike the mean, which can be misleading as a measure of central tendency if the distribution is skewed, the median is not affected by extreme scores (outliers). Although the mean can only be computed for values on a ratio or interval scale, the median is useful for data that may be on an ordinal scale as well as numerical (ratio or interval) scales.

Disadvantages in Using the Median

An advantage in using the mean as a representative value for the data set is that the whole data set enters into its computation; conversely, the median does not consider or compute the value of each item in the data set. It relies upon the placement of a value in an ordered set rather than upon all of the information the data can provide.

Mode

Mode is defined as the value that occurs most frequently in the data. To use the test score data set one more time, with

60, 65, 68, 74, 85, 85, 85, 85, 92, 98,

the mode can easily be identified as 85, the value that occurs more times than any other value in the set.

Although mean and median can be calculated for any data set, a data set in which each value occurs only once (or each value occurs the same number of times) does not have a mode. On the other hand, it is possible for a data set to have more than one mode; this occurs when two or more values of equal highest frequency are included in the data set. When this occurs, a bimodal distribution can be seen; the data distribution creates two peaks instead of the familiar bell curve-shaped unimodal distribution. Two peaks is a visually easy way to recognize that two different "maximum occurrence" values, or modes, occur in one distribution.

Although mean and median can be used as a single value to provide accurate summary information about an entire data set, the mode is rarely used as a summary statistic except to describe a bimodal distribution.

Advantages in Using the Mode

In addition to the ease of calculation for the mode, it is the only measure of central tendency that can be used for data measured in a nominal scale (a scale in which numbers are used as names of categories and do not represent quantities).

Disadvantages in Using the Mode

Mode is a very simplistic measurement of central tendency; because it is not algebraically defined, it is not used in statistical analysis. Also, variation in the frequency of observation is more likely when the sample size is small.

Relationship Among Measures of Central Tendency

The mean, median, and mode have logical relationships to each other and can offer different information on the data set as a whole. In a normal distribution, or bell curve, the mean, median, and mode are all equal and all are at the very center of the distribution or the 50th percentile.

In a data set with values that are clustered at extremes of very high values or very low values, the distribution will be skewed, as previously described. The mean in a skewed distribution will be pulled in the direction of the cluster of values, the opposite direction of the tail of the distribution. With a cluster of high extreme values, the mean will be pulled toward the higher end of the scale, and with a cluster of low extreme values, the mean will be pulled toward the lower end of the scale.

When extreme values are clustered at the high end of the scale, the distribution is skewed to the left. In such a distribution, the mode is the highest value of the three measures of central tendency, the mean is the lowest value, and the median lies between the mode and mean. When extreme values are clustered at the low end of the scale, the distribution is skewed to the right. In such a distribution, the mean is the highest value of the three measures of central tendency, the mode is the lowest value, and the median lies between the mode and mean.

According to Karl Pearson, the father of the discipline of mathematical statistics, in a moderately skewed distribution, if the values of any two measures of central tendency are known, the value of the third can be calculated. Pearson expressed the relationship among the three measures in this way:

$$\text{Mode} = 3\,\text{median} - 2\,\text{mean, and}$$

$$\text{Median} = \text{mode} + 2/3\,(\text{mean} - \text{mode}).$$

In moderately skewed distributions, the distance between the mean and median is about one-third the distance between the mean and mode.

Measure of Central Tendency as the Best Representation of a Data Set

Mean is the most frequently used measure of central tendency; it is widely understood and generally considered the best measure of central tendency for most data sets. The mean is most useful when there are not extreme scores in the data set. There are situations and types of distributions when one of the other measures of central tendency is preferred as a more accurate representation of the data set. Median is preferred to mean when there are extreme scores in the distribution, the data set is small, there are no big gaps in the middle of the set, data are measured in an ordinal (qualitative instead of quantitative) scale, or when some values in the data set are undetermined. Mode is the preferred measure when data are measured in a nominal scale and when the data set has many repeated data points.

Nicole M. Nickens

See also Measures of Central Tendency; Normal Distribution; Skewness

Further Readings

Adamson, K. A., & Prion, S. (2013). Making sense of methods and measurement: Measures of central tendency. Clinical Simulation in Nursing, 9(12), 617–618.

Manikandan, S. (2011). Measures of central tendency: Median and mode. Journal of Pharmacology & Pharmacotherapeutics, 2(3), 214–215.

Salkind, N. J. (Ed.). (2007). Encyclopedia of measurement and statistics. Thousand Oaks, CA: SAGE.

Nicole M. Nickens Nicole M. Nickens Nickens, Nicole M.

Measures of Variability

Measures of variability

1040

1042

Measures of Variability

Observations within a data set are not of equal value; they vary along a given scale. The extent to which they vary between and among themselves can be indicated by measures of variation or variability. Measures of variability show the amount of dispersion in the data set or, in other words, how much the observations or values are spread out along the scale. Dispersion within a data set can be measured or described in several ways, including the range, interquartile range, and standard deviation. This entry provides a definition, description, and calculation of each measure of variability along with advantages and disadvantages in using each. It also includes a discussion of standard deviation in a normal distribution, or the empirical rule, and Chebyshev's theorem.

Measures of central tendency (such as mean, median, and mode) each provide a single value that represents or is descriptive of the whole data set; measures of variability each provide a single value that represents how much spread exists in the data set. Measures of central tendency combined with measures of variability can offer an accurate summary description of a whole data set. A data set that contains values or observations that are much higher or much lower than the mean (extreme scores or outliers) has high dispersion. One measure of variability may be more appropriate than another depending upon the data set.

Range

The range of a set of data is the difference between the largest and smallest value in the data set; it is calculated very simply by using this formula:

$$\text{Range} = \text{maximum value} - \text{minimum value},$$

and it is very useful not only for showing dispersion in a data set but also for comparing variability between and among similar data sets. In the previous example, the professor of the college algebra course may want to compare dispersion of Exam 1 scores for three different classes or to compare dispersion of scores for Exams 1, 2, and 3 for the same class.

Although the range is the easiest measure of variability to find, using range to describe dispersion is not always the most appropriate choice. Calculation of range relies on only two values, so it is very sensitive to extreme values. If the highest and/or lowest value is/are an outlier(s), the range will provide an inaccurate picture of how much spread actually exists in the data set. Using the previous college algebra example, the professor wants to find the amount of dispersion in the set of scores from the first exam; there are 25 students in the class. If the highest score in the class is 90 of 100 and the lowest score is 20 of 100, the range is 90 – 20 = 70. This indicates a great deal of dispersion on a 100-point scale. However, if the majority of the scores cluster around the class average of 75 on the exam, and the scores of 20 and 90 are both outliers, then the range of 70 is misleading.

Interquartile Range

For data sets that contain outliers, the interquartile range can be calculated as a measure of variability; the interquartile range is not sensitive to outliers because it considers variability only within the middle 50% of the data set. The interquartile range is related to the median, which is a measure of central tendency that divides a distribution in half with 50% of the scores below and 50% of the scores above it. The distribution or data set can be further divided into fourths or quartiles. If there are 16 values in the data set, each quartile will contain 4 values; the data are not divided into 25% portions of total value but rather 25% of the number of observations. If the values in the data set are in order of ascending value, the first quartile contains the lowest 25% of the values; the second quartile contains the next 25% of the values; the third quartile contains the second highest 25% of the values; and the fourth quartile contains the highest 25% of the values. The interquartile range can be calculated by subtracting the first quartile from the third quartile:

$$\mathrm{IQR} = Q_3 - Q_1.$$

Find the median of the first quartile and subtract it from the median of the third quartile. This calculation is a measure of how the values of the data set are spread or clustered around the mean, and extreme scores outside of the middle 50% of the values do not impact the calculation. Thus, the interquartile range is more appropriate than the range as a measure of variability for data sets with outliers. However, like the range, the interquartile range is still dependent upon only two of the values in the data set and thus its representation of the whole data set may still be misleading.

Standard Deviation

The standard deviation, unlike the range or interquartile range, uses every value in the data set and therefore is a more powerful measure of dispersion. It is very widely used. The standard deviation is a measure that provides a summary or average of the amount that every value in the data set varies from the mean. A small standard deviation indicates that most of the values in the data set are clustered around the mean, whereas a large standard deviation indicates that there is more spread away from the mean.

The formula for calculating the standard deviation is the square root of the variance. The variance is the average of the squared differences from the mean. To calculate the variance, first find the mean (average) of the data set, then find the difference between each value in the data set and the mean of the data set (value − mean). Each of the differences should be squared and added together so that the average of that total can be calculated. The variance alone will not provide very useful information. For example, if teachers hope to find the standard deviation of grade point averages on a 4.25 scale in their advanced placement classes, they might calculate the mean (average) grade point average and then find the difference of each advanced placement student’s grade point average from the average; those differences would be squared and then added together. This is the variance, and it would certainly be a number far larger than a 4.25. The number produced would not be helpful by itself.

Once the variance is identified, the square root of the variance is the standard deviation. There are two different formulas to find the standard deviation, which one is used depends upon whether the values are a sample of a larger population or whether they represent an entire population. To find the standard deviation of a sample, the average will be found by calculating the sum of the squares of all deviations divided by the number of values (*n*−1).

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

In this formula, *s* = sample standard deviation, Σ = sum of, = sample mean, and *n* = number of values in sample.

To find the standard deviation of a population, the average will be found by calculating the sum of squares of all deviations divided by (*n*−1).

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{n}}.$$

In this formula, σ = population standard deviation, Σ = sum of, μ = population mean, and *n* = number of scores in the population.

Conceptually, it seems to make sense that to find the average of the dispersion in a data set (standard deviation), the calculation would require only to find the total of the differences between the mean and each value and figure the average of that total to determine the variance. However, because some of the differences between the value and the mean are positive and some are negative, if the differences are added together they will cancel each other out and the sum will be zero. That is the reason the averages must be squared before adding them.

Empirical Rule

A normal distribution is said to occur in many data sets when the values deviate from the mean value due to chance in such a way that most of the values are clustered around the mean, and only a few values are outliers (extreme highs or lows). In a normal distribution, the standard deviation can be used to determine the proportion of the total values that lie within a particular segment or range of the distribution. The empirical rule states that in a normal distribution, 68% of the values will fall within one standard deviation of the mean (34% of the values fall between −1 standard deviation of the mean and the mean, and another 34% fall between +1 standard deviation of the mean and the mean); 95% of the total values fall within two standard deviations of the mean; and 99.7% of the total values fall within three standard deviations from the mean. If the normal curve is broken up into sections according to standard deviations, it can be seen that starting with the mean and moving in either direction, the proportion of values contained in each section under the normal curve is always the same according to the empirical rule. Starting with the mean and moving to one standard deviation away from the mean in one direction holds 34% of the values in the data set; two standard deviations from the mean is an additional nearly 14%; three standard deviations from the mean is an additional nearly 2%.

Many data sets can fall into a normal distribution. If the standard deviation is small, the data are tightly clustered around the mean and the curve will be steep. If the standard deviation is large, the curve of the distribution will be flattened because the data are more dispersed or spread away from the mean.

Chebyshev’s Theorem

The empirical rule states how values in a normal distribution are proportioned under the bell curve as divided by standard deviations; however, the empirical rule applies only to normal distributions. Chebyshev’s theorem can be applied to any data set. According to Chebyshev’s theorem, at least 75% of the values lie within two standard deviations from the mean, and at least 89% of the values lie within three standard deviations from the mean. This theorem states the minimum proportion of values that will be contained within each segment of the distribution as defined by standard deviations from the mean.

Nicole M. Nickens

See also Interquartile Range; Normal Distribution; Standard Deviation; Variance

Further Readings

Gravetter, F. J., & Wallnau, L. B. (2017). Statistics for the behavioral sciences (10th ed.). Belmont, CA: Wadsworth–Thomson Learning.

Moore, D. S., Notz, W. I., & Fligner, M. A. (2015). The basic practice of statistics (7th ed.). New York, NY: W. H. Freeman.

Pagano, R. R. (2012). Understanding statistics in the behavioral sciences (10th ed.). Boston, MA: Cengage Learning.

Salkind, N. J. (Ed.). (2007). Encyclopedia of measurement and statistics. Thousand Oaks, CA: SAGE.

Wilcox, R. R. (2009). Basic statistics: Understanding conventional methods and modern insights. Oxford, UK: Oxford University Press.

John T. E. Richardson John T. E. Richardson Richardson, John T. E.

Median Test

Median test

1042

1045

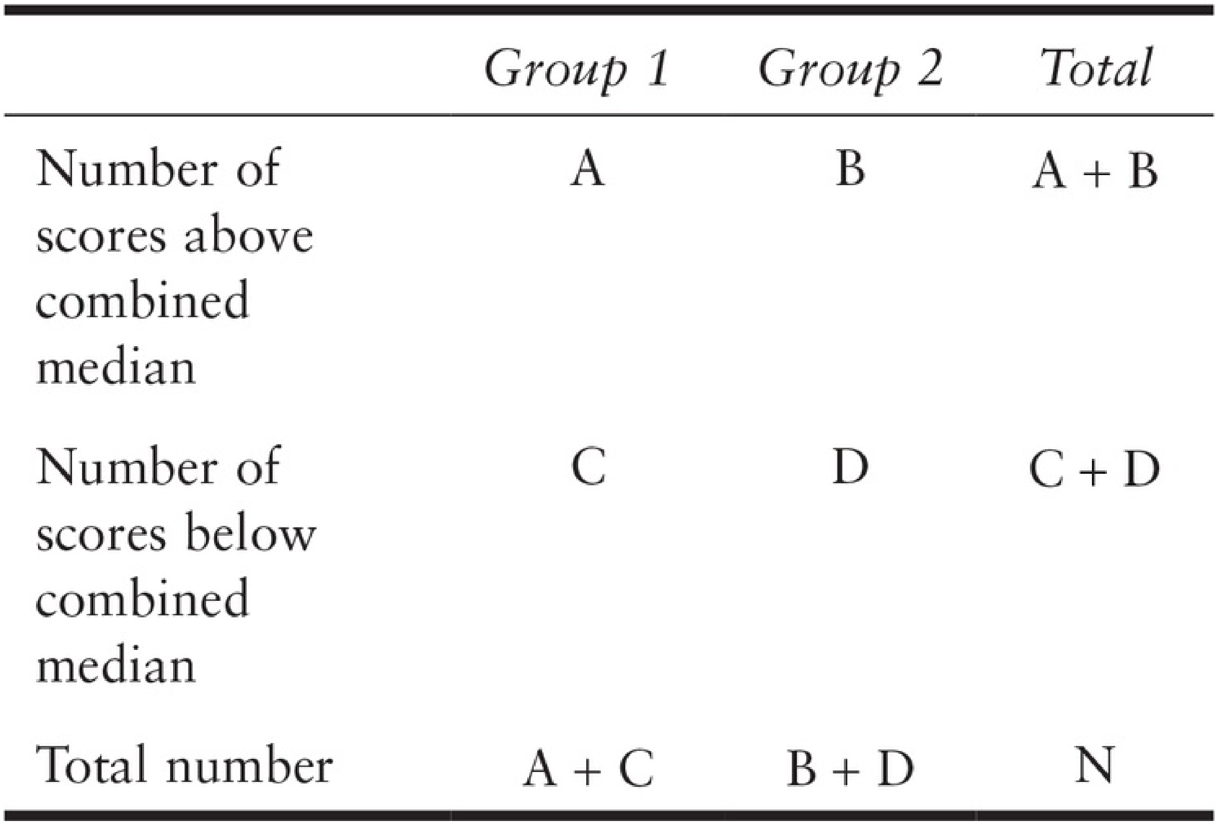
Median Test

A very simple design in quantitative research involves the random allocation of a sample of *N* individuals to two different groups. The groups are exposed to different treatments, and the research question is whether there is any difference between the two groups on some criterion variable. Classically, this question is addressed using Student’s *t* test for independent groups or the equivalent one-way between-subjects analysis of variance. However, this procedure assumes that the criterion variable in question (a) is measured on an interval or ratio scale, (b) is normally distributed, and (c) has the same variance in both of the groups. The median test was devised for use in situations in which one or more of these assumptions is not met. This entry describes the derivation of the median test, examines different ways of analyzing the contingency tables that result, and concludes by discussing the test’s power and power efficiency.

Origins of the Median Test

The median of a set of scores is a measure of their central tendency defined as the value below which 50% of the scores fall. If the number of scores is odd, the median is the centermost score when they are ranked from the lowest to the highest. If the number of scores is even, the median is the average of the two centermost scores when they are ranked from the lowest to the highest.

The median test is employed to test the null hypothesis that the scores obtained by two independent groups are drawn from populations with the same median. The first step is to find the overall median of the combined data (i.e., without regard to group membership). The scores within each group are then categorized in terms of whether they fall above or below the overall median. If the null hypothesis is true, then roughly half of the scores in each group should fall above the overall median and half should fall below the overall median. If the null hypothesis is false, the proportions in question would be expected to be different. Table 1 shows the outcome in a schematic form. This is simply an example of a 2 × 2 contingency table.



	Group 1	*Group 2*	*Total*
Number of scores above combined median	A	B	A + B
Number of scores below combined median	C	D	C + D
Total number	A + C	B + D	N

Source: Adapted from Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences* (p. 111). New York, NY: McGraw-Hill. Copyright by the McGraw-Hill Book Company, Inc.

The median test was first described by Dutch biologist Jacob Westenberg in 1948. He referred to the earlier writings of Ronald A. Fisher on the analysis of 2 × 2 contingency tables and advocated the use of the Fisher exact probability test (described later in this entry) to analyze data of the sort shown in Table 1. Westenberg’s initial account assumed that the two groups being compared contained the same number of cases (so that the total number of cases was by definition an even number), but in 1950, he published a more general account allowing for unequal sample sizes. The same year, a U.S. statistician, Alexander M. Mood, independently derived the median test. He proposed that the Fisher exact probability test should be used to analyze the resulting contingency table if either group contained 10 cases or fewer, but that for larger samples, Karl

Pearson’s chi-square test could be used instead. One of Mood’s colleagues, George W. Brown, demonstrated how the median test could be extended to encompass designs involving more than two groups. In this case, Pearson’s chi-square test would have to be used.

The median test assumes that the original observations are measured on at least an ordinal scale, that they are independent of one another, and that those within each group are drawn from the same population. However, it does not make any assumptions about the parameters of the populations from which the data are drawn, and hence, it is an example of a nonparametric test. A minor issue is how one treats scores that coincide with the overall median. This will arise by definition if the number of scores is odd, but it will also arise if the number of scores is even and the two centermost values are the same. Westenberg grouped these cases with those that fell above the overall median, whereas Mood grouped them with those that fell below the overall median. This probably makes little practical difference.

The Analysis of 2 × 2 Contingency Tables

The probability of obtaining any particular pattern of frequencies in a 2 × 2 contingency table depends on whether the two sets of marginal totals are regarded as fixed in advance or as random variables subject to sampling error. The median test constitutes a situation where both sets of marginal totals are fixed in advance. Referring to Table 1, the column totals (*A* + *C*) and (*B* + *D*) are fixed by the research design, whereas the row totals (*A* + *B*) and (*C* + *D*) are fixed by virtue of the definition of the median. On this model, the exact probability of any particular pattern of cell frequencies under the null hypothesis of no association between the two variables follows the hypergeometric distribution. This defines the probability of obtaining precisely *A* “successes” and *C* “failures” in a random sample of size (*A* + *C*) that is drawn from a population of size *N* in which there are (*A* + *B*) “successes” and (*C* + *D*) “failures.” This can be shown to be equal to

$$\frac{(A+B)!(C+D)!(A+C)!(B+D)!}{[N!A!B!C!D!]}.$$

Fisher advocated the use of this formula to calculate the exact probabilities of the observed set of frequencies and possible more extreme sets of frequencies that might have been observed under the null hypothesis of no relationship between the two dichotomous variables. This procedure is nowadays generally known as the Fisher exact probability test. Even so, Fisher recognized that the exact treatment of 2 × 2 contingency tables using the limited computational resources available at the time was extremely laborious, and instead, he recommended the use of Pearson's chi-square test as an approximation. He had previously shown that, when Pearson's chi-square statistic was obtained by estimating the expected cell frequencies in a 2 × 2 contingency table from the marginal totals, it could be calculated directly from the following formula and would have one degree of freedom:

$$\frac{N(AD-BC)^2}{[(A+B)(C+D)(A+C)(B+D)]}.$$

This formula is often cited in statistics textbooks, but there is actually an error in its derivation. The correct approximation to the hypergeometric distribution is

$$\frac{(N-1)(AD-BC)^2}{[(A+B)(C+D)(A+C)(B+D)]}.$$

How accurate an approximation are these variants of the chi-square statistic? John T. E. Richardson used analytic methods to calculate their Type I error probabilities (i.e., the likelihood that statistically significant results would be obtained when the null hypothesis was true). Using the conventional level of statistical significance of .05, Richardson found that, for very large samples (N = 1,000), both formulas led to Type I error probabilities between .04 and .06; however, these varied more widely even for moderately large samples (N = 100). As a consequence, many authorities consider that Pearson's chi-square test is not sufficiently accurate to be useful with 2 × 2 contingency tables when both sets of marginal totals are fixed in advance. It follows that it should not be used in connection with the median test.

However, there is also a problem with the Fisher exact probability test. Any discrete statistic will be systematically conservative when used in classical hypothesis testing. In general, any such statistic will typically have two possible adjacent values such that the first has an exceedance probability that is greater than the threshold probability level α under the null hypothesis, whereas the second has an exceedance probability that is less than α under the null hypothesis. When carrying out tests of statistic inference against a prespecified significance level, the null hypothesis will be rejected only when the latter value is attained or one more extreme. However, it follows that the effective proportion of Type I errors in such a case will normally be less than α.

To illustrate this point, refer to Table 1. Suppose that $n = 10$ and that the marginal totals are 5, 5, 5, and 5. Therefore, the value of A can vary between 0 and 5. Under the null hypothesis of no association between the two variables, the results will be statistically significant if A is 0 or 5 ($p = .008$ by a two-sided test according to the Fisher exact probability test) but not otherwise. For instance, if A is 1 or 4, $p = .206$. Even though the nominal Type I error rate is .05, the actual Type I error rate is .008. In other words, the Fisher exact probability test is very conservative, especially when used with small samples (as Fisher himself intended).

One solution would be to reject the null hypothesis on a certain proportion of those occasions when the probability of the obtained results or of more extreme results, $p(oe)$, is greater than α but when the probability of results more extreme than those actually obtained, $p(e)$, is less than α. To be more precise, one should reject the null hypothesis on $[\alpha - p(e)]/[p(oe) - p(e)]$ of those occasions, chosen entirely at random. Keith D. Tocher proved that, if combined with the Fisher exact probability test, this decision rule offered the uniformly most powerful, unbiased test of statistical inference for 2 × 2 contingency tables in which both sets of marginal totals were fixed. Nevertheless, the introduction of random elements into a statistical procedure aroused philosophical misgivings on the part of other statisticians, and the technical complexity of Tocher’s procedure means that it is seldom used in research practice.

Power and Power Efficiency

The *power* of a statistical test is the probability of rejecting the null hypothesis when it is false. Its complement is the probability of *not* rejecting the null hypothesis when it is false, in other words, the probability of making a Type II error. In general, nonparametric tests tend to be less powerful than the corresponding parametric test because they use less of the information that is contained in the data. For instance, the median test only uses information concerning whether each observation is above or below the overall median, whereas Student’s *t* test employs the actual values of the observations.

The power of two different statistical tests in the same research design can be compared using the notion of *power efficiency*. This notion relies upon the fact that the power of a test in a particular situation depends (other things being equal) on the sample size. Suppose that Test 1 is the most powerful statistical test when used in a particular research design with data that meet its underlying assumptions. Test 2 is a less powerful test in the same design, in that it would need to be used with a sample of N_2 cases to match the power that is achieved by Test 1 with N_1 cases (where $N_2 \geq N_1$). The power efficiency of Test 2 is N_1 / N_2, often expressed as a percentage.

When used with the Fisher exact probability test, Mood showed that the power efficiency of the median test in comparison with Student's *t* test for independent samples approached a value of 2 / π or 63.7% as the overall sample size increased. Using Monte Carlo simulation, Boris Freidlin and Joseph L. Gastwirth found that the power of the median test was consistently poor across a range of conditions. They concluded that the median test should be retired from general use in favor of other procedures such as the Mann-Whitney test.

John T. E. Richardson

See also Chi-Square Test; Fisher Exact Test; Mann-Whitney Test; Measures of Central Tendency; *t* Tests; Type I Error; Type II Error

Further Readings

Brown, G. W., & Mood, A. M. (1951). On median tests for linear hypotheses. In J. Neyman (Ed.), Proceedings of the second Berkeley symposium on mathematical statistics and probability (pp. 159–166). Berkeley: University of California Press. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200500226

Fisher, R. A. (1934). Statistical methods for research workers (5th ed.). Edinburgh, Scotland: Oliver & Boyd.

Freidlin, B., & Gastwirth, J. L. (2000). Should the median test be retired from general use? American Statistician, 54, 161–164. doi:10.1080/00031305.2000.10474539

Mood, A. M. (1950). Introduction to the theory of statistics. New York, NY: McGraw Hill.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. Annals of Mathematical Statistics, 25, 514–522. doi:10.1214/aoms/1177728719

Richardson, J. T. E. (1990). Variants of chi-square for 2 × 2 contingency tables. British Journal of Mathematical and Statistical Psychology, 43, 309–326. doi:10.1111/j.2044-8317.1990.tb00943x

Tocher, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika, 37, 130–144. doi:10.1093/biomet/37.1-2.130

Westenberg, J. (1948). Significance test for median and interquartile range in samples from continuous populations of any form. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 51, 252–261. Retrieved from http://www.dwc.knaw.nl/DL/publications/PU00018486.pdf

Westenberg, J. (1950). A tabulation of the median test for unequal samples. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 53, 77–82. Retrieved from http://www.dwc.knaw.nl/DL/publications/PU00018756.pdf

Milica Miocˇevic´ Milica Miocˇevic´ Miocˇevic´, Milica

David P. MacKinnon David P. MacKinnon MacKinnon, David P.

Mediation Analysis

Mediation analysis

1045

1049

Mediation Analysis

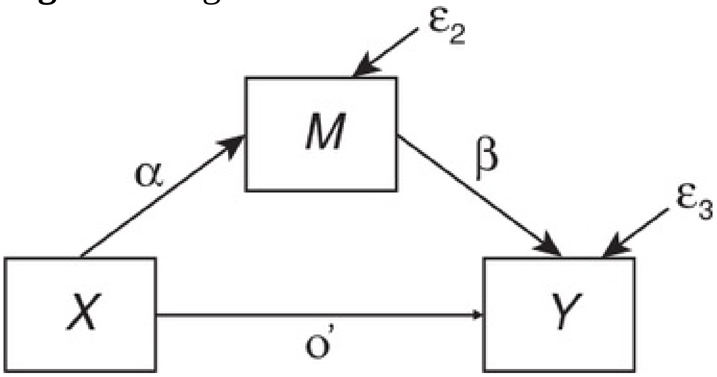
Mediation analysis is used in social sciences, biology, epidemiology, and other fields in order to evaluate the mechanism through which an independent variable (X) affects a dependent variable (Y). The variable transmitting the influence of the independent variable onto the dependent variable is called the mediator (M), and the indirect effect through the mediator is called the mediated effect. In prevention research, mediation analysis is used to improve future interventions by discovering the mediator(s) through which the intervention (X) affects the outcome of interest (Y); for example, an intervention designed to reduce smoking (X) affects self-efficacy (M), which in turn affects the number of cigarettes smoked (Y). The mediated effect in this example represents the indirect effect of the intervention on the number of cigarettes smoked through participants' self-efficacy. In fields such as social psychology, mediation analysis is used to uncover the intermediate variable that transmits the effect of an independent variable onto a dependent variable; for example, exposure to a certain message (X) affects attitudes (M), which then affect behavior (Y). The mediated effect is the indirect effect of the message on the behavior through the attitude.

It is also possible to have a model with two parallel mediators or several sequential mediators. The effect of age (X) on typing proficiency (Y) is mediated through skill (M_1) and manual dexterity (M_2) in a parallel two mediator model. In this case, there could be several effects of interest: the mediated effect through skill alone, the mediated effect through manual dexterity alone, and the total mediated effect through skill and manual dexterity. Some theories posit a sequence of mediators between an independent variable and a dependent variable, and in order to evaluate these effects, one would fit a sequential mediator model. For example, being assigned to receive a treatment versus placebo (*X*) affects sleep quality (M_1), which then affects alertness (M_2), which affects the number of automobile accidents (*Y*). In this model, the mediated effect has three paths, from *X* to M_1, from M_1 to M_2, and from M_2 to *Y*. There are several ways of computing point and interval estimates of the mediated effect. The following sections describe the simplest mediation model, mention several more complex mediation models, and discuss recent methodological advances in the field of mediation analysis.

Single Mediator Model

The single mediator model is the simplest example of mediation analysis. In this model, the independent variable (*X*) affects the mediator (*M*), which in turn affects the outcome (*Y*; see Figure 1).

Figure 1 Single mediator model



The single mediator model is described using three equations:

$$M = \alpha X + \varepsilon_2, \text{ and}$$

$$Y = \tau X + \varepsilon_1, Y = \tau' X + \beta M + \varepsilon_3,$$

where *X* is the independent variable, *M* is the mediator, *Y* is the dependent variable, τ is the coefficient for predicting *Y* from *X*, α is the coefficient for predicting *M* from *X*, and τ′ and β are the coefficients for predicting *Y* from *X* and *M*, respectively. The residual terms ε_1, ε_2, and ε_3 are assumed to follow normal distributions with means of zero and variances of , , and , respectively. The last two equations are sufficient for estimating the mediated effect, αβ.

Point and Interval Estimates of the Mediated Effect

There are several ways to test for mediation: testing the significance of a series of regression coefficients as in Baron and Kenny’s classic causal steps approach, testing the significance of the paths that comprise the mediated effect, computing the mediated effect as the product of coefficients and dividing it by the appropriate standard error of the mediated effect, and computing the mediated effect as the difference in coefficients and dividing it by the appropriate standard error of the mediated effect. When *M* and *Y* are continuous and the three equations are estimated using OLS regression, the point estimate of the mediated effect can be computed either as the difference of coefficients τ−τ′ or as the product of coefficients αβ. In more complex models (i.e., with multiple sequential mediators and/or with latent variables), the mediated effect is computed as the product of coefficients, and thus, the remainder of this essay will only refer to the mediated effect as the product of coefficients.

The mediated effect in the single mediator model is interpreted as the number of units of change in *Y* for a one-unit increase in *X* through *M*. This is easiest to conceptualize when *X* is binary, with the two levels coded 0 and 1. In the example of the prevention program (*X*) that aims to reduce smoking (*Y*) through self-efficacy (*M*), the mediated effect represents the indirect effect of being in the treatment (*X* = 1) versus control (*X* = 0) condition on cigarettes smoked (*Y*) through self-efficacy (*M*).

In addition to computing the point estimate of the mediated effect, it is also advisable to quantify the uncertainty surrounding the estimate by computing an interval estimate. The most common ways of constructing interval estimates for the mediated effect are either distribution-based methods or resampling methods. Distribution-based methods require assumptions about the distribution of the mediated effect, and the general form of the confidence interval for the mediated effect using these methods is

$$\hat{\alpha}\hat{\beta} + z_{\alpha/2}(s_{\hat{\alpha}\hat{\beta}}) \leq \alpha\beta \leq \hat{\alpha}\hat{\beta} + z_{1-\alpha/2}(s_{\hat{\alpha}\hat{\beta}}),$$

where is the sample estimate of the mediated effect, $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the critical values from the assumed distribution (with [1 − α] being the desired level of confidence), and is the sample standard error of the mediated effect.

Before 2002, it was common to assume that the mediated effect is normally distributed and use critical values from the normal distribution to construct intervals for the mediated effect; however, the product of two normal distributions (e.g., two regression coefficients) is not normal. For this reason, using critical values from the normal distribution in the computation of confidence interval for the mediated effect produces intervals with suboptimal statistical properties.

Unlike the normal distribution, the distribution of the product is usually not symmetric and thus can better accommodate the asymmetry of the mediated effect. The computation of the confidence intervals for the mediated effect based on the distribution of the product has been simplified with programs called PRODCLIN and RMediation. Simulation studies that compared statistical properties of confidence limits for the mediated effect obtained using normal theory and the distribution of the product have found that confidence limits using critical values from the distribution of the product have superior statistical properties relative to confidence limits using critical values from the normal distribution.

Another way of constructing interval estimates of the mediated effect is using resampling methods. The bootstrap is one type of resampling method, and findings from simulation studies suggest that the confidence intervals obtained using the percentile bootstrap have comparable statistical properties to confidence intervals obtained using the distribution of the product and that both methods are superior to normal theory for interval estimation in mediation analysis. The bootstrap methods, however, are more easily generalized to more complex mediation models.

The interval estimates of the mediated effect obtained using normal theory, distribution of the product, and bootstrap methods are interpreted in terms of confidence. A 95% confidence interval for the mediated effect means that upon repeated sampling, 95% of the intervals for the mediated effect will contain the true value of the mediated effect. Note that this is not the same as stating that the mediated effect lies between the two interval limits with 95% probability.

Some theories require the extension of the single mediator model to include moderators, multiple mediators, latent variables, multilevel modeling, and/or repeated measures. Mediation analysis can be done within more complex models following the same general principles as in the single mediator model. Methods for estimating the mediated effect have been described for models with moderation and mediation, parallel mediation models, mediation models with a sequence of at least two mediators, structural equation models, multilevel models, and longitudinal mediation models.

Methodological Advances in Mediation Analysis

Theories in social sciences are often phrased in terms of one variable influencing other variables, and data are collected in order to quantify the empirical support for a hypothesis and/or model given the observed sample. Recent methodological advances have produced guidelines for causal inference and probabilistic interpretations of findings in mediation analysis.

It is often assumed that random assignment to levels of the independent variable (*X*) will lead to confounders being equally distributed between groups and thus to the validity of the conclusion that any difference in outcome (*Y*) between groups is due to the independent variable. In order to make causal inference claims for mediation, one needs to make several assumptions. The assumption of sequential ignorability states that variables omitted from the analysis do not bias the relationship between *X* and *M* and that variables omitted from the analysis do not bias the relationship between *M* and *Y*. In other words, sequential ignorability means assuming no unmeasured confounders. The ignorability assumption for the *X* to *M* relation can be satisfied by random assignment of *X*. However, ignorability of the *M* to *Y* relation is not satisfied by randomization because the level of the mediator cannot be randomly assigned, and thus, the β path is not causal. With actual data, it is not possible to know whether the sequential ignorability assumption has been satisfied; however, sensitivity analysis can be used to evaluate the robustness of the observed findings to the violations of the sequential ignorability assumption.

The distribution-based and resampling methods are based on a classical (frequentist) view of probability in which support for the existence of an effect is found by refuting the null hypothesis of no effect. In this framework, a 95% confidence interval for the mediated effect means that upon repeated sampling, 95% of intervals will contain the true value of the mediated effect. An alternative way of approaching hypothesis testing and parameter estimation is using Bayesian statistics. In Bayesian statistics, parameters have distributions, and data

Bayesian statistics. In Bayesian statistics, parameters have distributions, and data are treated as fixed. Thus, inference is made about parameters conditional on data, unlike in classical statistics in which one computes the likelihood of the observed data given the hypothesis. Bayesian inference has three ingredients: the prior distribution, the likelihood function, and the posterior distribution used to make inferences. The prior distribution is a statement of current knowledge before the data are observed. The posterior distribution is obtained by updating the prior distribution with the observed data using Bayes's law.

In the context of mediation analysis, prior distributions can be assigned to parameters in equations (referred to as the method of coefficients) or to the covariance matrix of the independent variable(s), mediator(s), and dependent variables (referred to as the method of covariances). Like all parameters in the model, the mediated effect αβ has a posterior distribution. In order to make probabilistic inferences, the posterior for the mediated effect can be summarized using point and/or interval summaries. For example, a researcher can compute the probability that the mediated effect is greater than some meaningful value (e.g., zero), and Bayesian credibility intervals for the mediated effect give the answer about the probability that the true mediated effect lies between the two interval limits. The ability to use existing information in the mediation analysis and probabilistic statements about the value of the mediated effect are the two main advantages of the Bayesian over the frequentist approach.

Mediation analysis is commonly used in many areas of research because it provides information about the process by which an independent variable affects a dependent variable. The methodology for the single mediator model has been extended to include multiple independent variables, mediators, outcomes, moderators, and latent variables as well as longitudinal and multilevel data. Methodological research in statistical mediation analysis is ongoing, and most recent developments focus on the computation of causal estimates and probabilistic interpretations of the mediated effect.

Milica Miocˇevic´ and David P. MacKinnon

See also Analysis of Covariance; Correlation; Path Analysis

Further Readings

Cox, M. G., Kisbu-Sakarya, Y., Miocˇevic´, M., & MacKinnon, D. P. (2014). Sensitivity plots for confounder bias in the single mediator model. Evaluation

Review, 37(5), 405–431.

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. Prevention Science, 10(2), 87–99.

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. Multivariate Behavioral Research, 47(1), 61–87.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. Multivariate Behavioral Research, 36(2), 249–277.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. Evaluation Review, 17(2), 144–158.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behavior Research Methods, Instruments, & Computers, 36(4), 717–731.

Valeri, L., & Vanderweele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods, 18(2), 137–150.

Vanderweele, T., & Vansteelandt, S. (2013). Mediation analysis with multiple mediators. Epidemiologic Methods, 2(1), 95–115.

Nicole Mittenfelner Carl Nicole Mittenfelner Carl Carl, Nicole Mittenfelner

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

Member Check

Member check

1049

1050

Member Check

The phrase *member check* refers to a set of processes in which researchers “check in” with participants in a qualitative study so that participants can consider and respond to their comments in the data and/or to researchers’ interpretations of the data. Member checks are often (and ideally) used in combination with other methods to establish a study’s validity, particularly a study’s credibility. Member checks are important in educational research studies because determining if and how participants “see themselves” and relate to the data and the researcher’s interpretations (or if and how they do not) can help to ensure the credibility and reliability of the research process, including data collection, analysis, and reporting.

Member checks can also be referred to as participant or respondent validation strategies; some researchers prefer these terms because they signify a more process-oriented and relational approach rather than a one-time, transactional interaction. Regardless of the term, the goal of member checks is to help researchers determine whether and how they understand participants’ experiences, perspectives, and realities.

Member checks can be technical processes, which may involve having participants verify the accuracy of statements in transcripts. Member checks can also be analytical in which participants respond to or engage with a researcher’s interpretations at different stages of analysis. This entry focuses on the role of member checks in relationship to study credibility, provides an overview of the processes of member checking, and discusses some critiques of and challenges to conducting member checks.

Establishing Credibility

Conducting member checks is a method that qualitative researchers use to enhance the validity or trustworthiness of a study; specifically, member checks are often used to establish credibility, which means that the research findings are believable to research participants. Validity in qualitative research should not be thought of as simply being achievable through a checklist of items. Because validity refers to the quality and rigor of a study, researchers should engage in validity techniques systematically throughout an entire study, and member checks are an important way to engage participants at various stages of data collection and analysis. It is in this regard that member checks can inform the credibility of a study because conducting member checks means that participants have had opportunities to react to the data and/or researchers' interpretations. Member checks should be conducted with fidelity to participants and their experiences in such a way that participants' feedback and challenges are seriously considered and help inform the interpretive frames of a study.

Process Overview

Member checks can occur in a variety of ways. For example, they may take place informally during data collection (i.e., during an interview or focus group to check for understanding) or more formally during a follow-up interview, meeting, or conversation. Member checks can involve checking in with participants to determine accuracy, engaging in sustained dialogue with participants, and/or involving participants in collaborative data analysis processes.

The processes that researchers select vary depending on the qualitative approach, study goals, research questions, and issues of participant availability and access. The process should involve sufficient time in the overall research design and timeline so that researchers can actively engage with and respond to participants' critiques, interpretations, and additions. Ideally, member checks are employed at multiple points throughout a study so that they are proactively and intentionally structured into a study's research design and not tacked on at the end of a study. When done with fidelity, participant validation processes often result in additional data that can contribute to a study's ability to capture and represent complexity. These processes (e.g., follow-up interviews) should be systematically recorded and analyzed in the same ways as all other forms of data.

To be clear, member checks can be technical—moments when participants are asked to respond to the accuracy of data—or they can be analytical—times when participants may respond to analytical themes, codes, and/or findings. Member checks ideally occur continuously throughout data collection and analysis. That is why some contemporary methods scholars prefer the term *participant validation* because it better captures the processes by which qualitative researchers continuously engage with participants to help make sure they understand, to the fullest extent possible, participants' experiences and perspectives.

Critiques and Challenges

When not considered holistically, member checks have the potential to denote that there is a single truth to be determined and then confirmed, which runs contrary to the interpretivist paradigm of qualitative research, which contends that there is no single truth but rather multiple, situated realities. When participant validation processes are approached as holistic ways of engaging with and understanding participants' experiences, and not solely as a form of verification, they become a vital part of establishing validity in qualitative research.

It is important to think through potential challenges that could arise during member checks and to make sure that researchers make the most of participants' time. This involves researchers reflecting on what they hope to achieve as a result of checking in with participants. For example, if sharing transcripts, what are participants supposed to do with the transcript (i.e., see if they still agree with what they stated? Add to points they have made? Clarify language and concepts?) and how should researchers respond to them? If discussing the researcher's interpretations, how should researchers react if the participant disagrees with their interpretations? Researchers should consider multiple ways to approach member checks, receive feedback from participants, and integrate that feedback in meaningful ways that relate to the study's data and analysis.

Member checks can help qualitative researchers conduct rigorous and valid studies by assessing and challenging the researcher's interpretations, but it is important to remember that solely conducting member checks does not mean that a study is considered valid. Engaging participants in analysis and interpretation is an important process that should, when appropriate to the research design, be used to strengthen the rigor and validity of a qualitative research study.

Nicole Mittenfelner Carl and Sharon M. Ravitch

See also Naturalistic Inquiry; Qualitative Data Analysis; Reliability; Triangulation; Trustworthiness; Validity

Further Readings

Barbour, R. S. (2001). Checklists for improving rigour in qualitative research: A case of the tail wagging the dog? British Medical Journal, 322, 1115–1117.

Cho, J., & Trent, A. (2006). Validity in qualitative research revisited. Qualitative Research, 6(3), 319–340.

Ellingson, L. L. (2009). Engaging crystallization in qualitative research: An introduction. Thousand Oaks, CA: SAGE.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. Educational Resources Information Center Annual Review Paper, 29, 75–91.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: SAGE.

Ravitch, S. M., & Carl, N. M. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: SAGE.

Jacqueline Remondet Wall Jacqueline Remondet Wall Wall, Jacqueline Remondet

Merit

Merit

1050

1051

Merit

The word *merit* can be defined as the quality of being good, important, or useful; it can also refer to a quality or conduct deserving of praise or reward. Historically, those in authority determined merit; however, with the advent of the scientific revolution, a more objective method of determining merit began.

In evaluation science, criteria are established to determine the value of that which is being assessed. Postmodern philosophy has contributed to evaluation science through the critique of rationality and logical positivism. Postmodern ideas correlate with changes in the definition of merit that are described here.

In evaluation settings, merit is an element of the purpose of an evaluation, where it is an approach to identifying the value of the evaluand or that which is being evaluated. When using evidence-based methods, merit is therefore objectified through defining relevant criteria to determine the value of that which is received, as determined by those receiving benefit from the evaluand. For example, if evaluating the effectiveness of an educational program, completing the program is thought to have value to those who have graduated Evaluation science also explores considerations of value for others connected to the evaluand, in addition to those who are expected to receive benefit. In the case of an educational program, these considerations of value could involve those who have commissioned the evaluation (e.g., funders and administrators), those who are responsible for that which is being evaluated (e.g., administrators), others who are part of that which is being evaluated (e.g., employees or volunteers), and those who are indirectly impacted by that which is being evaluated (e.g., groups and society more broadly). With these multiple perspectives potentially groups and society more broadly). With these multiple perspectives potentially having similarities and differences, merit is determined by the value not only for those receiving the benefit of that which is evaluated but also for others in the setting.

In determining merit, it is important to also consider two related terms, *worth* and *significance*. Along with merit, worth and significance provide the foundation for evaluation. Merit has been determined to be inherent, intrinsic, and context free; a value element of the evaluand. Worth also illustrates value, but it is identified as extrinsic in that it is contextually determined and often incorporates a cost-related value in a comparative fashion. Significance is associated with importance or size and also reflects how those impacted the view of the evaluand. As such, worth and significance may be influenced by contextual or situational factors outside of merit.

The critical feature of merit is the recognition that it must be based on evidence and logic as applied in evaluation work and used to determine whether and how well an evaluand meets its objectives or outcome criteria or provides value.

Jacqueline Remondet Wall

See also Applied Research; Evaluation; Formative Evaluation; Process Evaluation; Summative Evaluation

Further Readings

Posavac, E. (2015). Program evaluation: Methods and case studies. Abingdon, UK: Routledge.

Scriven, M. (1991). Evaluation thesaurus. Newbury Park, CA: SAGE

Paul B. Ingram Paul B. Ingram Ingram, Paul B.

Michael S. Ternes Michael S. Ternes Ternes, Michael S.

Meta-Analysis

Meta-analysis

1051

1054

Meta-Analysis

Meta-analysis is an observational method for summarizing research. It provides a context for quantitatively understanding multiple studies on a topic through synthesis of individual effect sizes and variances observed across studies. Utilizing these effect sizes and variances from numerous studies, meta-analysis calculates an overall estimate of effect for a given phenomenon. For instance, meta-analysis can be used to evaluate the efficacy of behavior therapy on phobias, even if studies on this topic vary greatly in their estimation of influence on outcome. In addition, a meta-analysis provides information on the significance of the effect and precision of the estimate. This entry first looks at types of meta-analysis and how researchers prepare for a meta-analysis and calculate meta-analytic estimates. It then discusses the reporting of meta-analytic results and methodological issues with meta-analysis.

Types of Meta-Analyses

Meta-analytic estimates are calculated in one of two ways, either through fixed or through random effect models. Fixed-effect models assume that all variabilities observed between study effect sizes are due to the differences in sampling error. Underlying this sampling error is a single, common effect (e.g., the true effect). Fixed-effect models assume a null hypothesis of no treatment effect in every study.

Fixed-effect models are appropriate if it can be assumed that all study characteristics are the same (e.g., identical conditions surrounding the characteristics of the participants, the researchers, and the experimental dosage). This assumption is rarely appropriate; multiple studies are seldom conducted in an equivalent manner. Indeed, even if a researcher were to find what the researcher believes to be appropriate conditions for a fixed-effect meta-analysis, heterogeneity of effect sizes should be tested to ensure the assumptions inherent to a fixed-effect model are tenable.

Conversely, random-effects models assume that there is no single estimate of effect. Instead, the actual influence of a given factor will vary between studies as a function of study characteristics. Effect sizes observed in studies from which a random-effects meta-analysis is calculated are assumed to be drawn from a normally distributed, random sampling of effect estimates with a mean of zero.

The assumption that a random-effects model varies across the study context is important because it provides a role for moderating influence, a factor unavailable in fixed-effect models. For instance, a study of parenting interventions for disruptive child behavior using a fixed-effect model would assume that the intervention would have the same effect across all studies; a random-effects model estimates the effect of the intervention as a function of other characteristics, such as sampling variability or a moderator such as parent-child attachment.

Preparation for Meta-Analysis

Essential to any meta-analysis is thorough preparation, beginning by specifying a precise research question. Once a clear intended thesis has been decided, determinations of how strictly to impose standards of quality can be made through the defining of criteria for study inclusion. There are many arguments for implementing strict guidelines for inclusion; however, there remains a need to present a thorough and comprehensive review, and some arguments for strictness are counterbalanced by the nature of meta-analysis. Strict guidelines on study quality can impact meta-analytic results and may be more comprehensively explored using moderator analyses.

After solidifying what types of studies will be included in the analysis, care should be taken to exhaustively search for all relevant studies. This search often involves creating a key word, or set of key words, that will be used in database searches (e.g., PsycINFO, ERIC, Google Scholar). If unpublished studies are to be included, then all efforts must be taken to discover as many of these sources as possible, such as contacting authors of published works and scouring conference proceedings. A forward and backward search of included studies is useful to ensure the comprehensive inclusion of all relevant studies.

After gathering sources, the process of coding information of interest begins. There are six basic domains of information that have been suggested as useful to code: report identification (i.e., year of publication, authors); setting (i.e., context of the study); participants (e.g., age, race/ethnicity, patient status); methodology (i.e., how studies were conducted, variables observed); treatment and any procedures involved; and all statistical information from the study necessary to compute the desired effect size and all relevant metrics. Some studies will present more than one effect size (e.g., a study that reports treatment means for both a male and female sample independently). There are two ways that this situation can be approached. One is to synthesize the two treatment reports to an averaged metric. The other is to incorporate both figures separately into the meta-analysis.

Calculation of Meta-Analytic Estimates

Once preparation is complete, a common metric must be calculated to accurately summarize the data. There are three standards upon which effect sizes are based: correlations, mean differences, and binary data.

Correlation coefficients are effect sizes, but given that multiple studies will likely utilize different measures to generate these figures, merely combining the coefficients would be inappropriate. Thus, correlation coefficients of interest are recorded and then transformed into Fisher's *z* scale where analyses are performed using this transformed value. This standardized metric limits influence from various measurement and study-specific concerns. Once results (e.g., mean effect size and its confidence intervals) are calculated, Fisher's *z* scores are converted back to the *r* metric for interpretation.

Mean differences are a common and useful method for determining an effect size. Although mean differences can act as an effect size, the multiple studies included in a meta-analysis utilize differing methods of assessment, making a standardized metric necessary. Mean differences for each study are converted to Cohen's *d* in order to calculate the mean effect size. However, Cohen's *d* can overestimate the standardized mean difference—particularly in smaller sample sizes of studies. To correct for this, Hedges’s g can be utilized as a second standardized metric. Unlike with the Fisher z calculations for correlation coefficients, the resulting summary effect calculated in the Hedges’s g metric can be reported as the final form of the mean effect size estimate.

Some studies will seek to understand the risk ratio (also called odds ratio) for a specified event or want to know the difference in risk. In this case, binary data are utilized to construct these ratios. Much like mean differences and correlation coefficients, odds and risk ratios are computed utilizing a standardized metric using natural logs. Thus, odds ratios become log odds ratios, and risk ratios become log risk ratios. Summary effects are calculated in the log metrics before being converted back to the raw form for interpretation. An exception to this practice is in risk difference, which can be calculated with the unstandardized data.

Regardless of which metric is utilized, all effect sizes compute a summary effect size for the study. The effect size does not rely on null hypothesis statistical testing to demonstrate its impact. Instead, effect sizes have a small, medium, or large impact based on their observed values (e.g., 0.3, 0.5, or 0.7, respectively). Summary effect can then be explored further, determining the degree to which variance is adequately explained by the underlying model of analysis or to what degree the figure is representative of data that are heterogeneous.

One commonly used measure of heterogeneity is Cochran’s Q statistic. Q is a measure of weight squared deviations and explains the amount of true variance observed versus random error. Upon rejecting the null hypothesis, the associated Q value indicates unexplained variance beyond random noise. Another metric of heterogeneity is I^2. This statistic, much like the Q statistic, estimates observed variance. I^2 provides the percentage of unexplained variance due to something other than chance. Tau-squared (T^2) is another parameter of heterogeneity and is often used to estimate the variance in true effects and is particularly important within random-effects meta-analyses.

Reporting of Meta-Analytic Results

Conducting a meta-analysis requires effectively reporting the research synthesis process to ensure that design decisions and synthesized results may be interpreted effectively. Reporting should enable meta-analytic replications using the same search criteria and methodological decision-making criteria. Metathe same search criteria and methodological decision-making criteria. Meta-analytic methods should contain information about inclusion and exclusion criteria, sources of information and full search strategy used to obtain sources of information utilized in the meta-analysis (e.g., electronic databases searched, search terms used), operationalization definitions of analyzed variables, processes for how data were extracted from the source material, methods to assess bias, and planned analyses such as moderation evaluation or meta-regression.

A strong a priori operationalization of outcome and moderator definitions helps assuage criticism that meta-analysis compares unlike phenomenon. Results should include description of study characteristics and identification on included studies, results of analyses for bias, and a reporting of summary statistics. Statistic summaries should allow readers to understand important demographic characteristics related to the study samples but also information about how studies contributed effect sizes to each analysis. This inclusion of information is particularly important when not all studies offer the same types of information, as is common in moderator analyses. These reporting components are not exhaustive and may need to be supplemented, depending on the nature of the meta-analysis. These reporting patterns should be considered when interpreting a meta-analysis.

Methodological Issues

Although there are many benefits to meta-analysis, particularly in its lack of reliance on null hypothesis statistical testing, it is not without its potential concerns. One such criticism is long-standing and is often termed *commensurability*. Because meta-analysis is the reduction of independent studies and data, there exists a concern that this reduction involves the comparison of “apples to oranges.” This concern is birthed from the logic posited by researchers that different studies, while related enough to meet inclusion criteria, will utilize different measures and constructs for the investigations of interest. Combining varied constructs into one metric is perceived as illogical due to the potential incongruence of the various characteristics involved and therefore the inability to create a unified metric.

Commensurability is far from the only threat to meta-analysis as a statistical methodology. The quality of studies included is also a concern. When multiple studies are pooled, there are increased opportunities for the influence of study errors through biased or unreliable data. In an attempt to counter this influence errors through biased or unreliable data. In an attempt to counter this influence, some meta-analyses will incorporate metrics of study quality within moderator analyses. However, this approach is not a flawless way to control for variability due to design quality within included studies.

Raising the standards for studies included in an analysis could be another avenue for mitigating the viability of the study quality concern. This approach potentially results in selection bias. Selection bias is the idea that the way in which studies are chosen for inclusion in a meta-analytic review is not always objective. If a systematic bias exists in how studies are permitted for inclusion, there is the opportunity for skewed results. Because meta-analysis is supposed to be comprehensive in its findings, such an effect on results potentially invalidates any findings.

A problem related to selection bias is publication bias. Often published works are the most accessible resource for researchers when collecting data for a meta-analysis. There are many databases of research publications online and in print, offering millions of published works. Each of those works includes a list of citations to further a researcher’s foraging for data.

In a system of research that relies on null hypothesis statistical testing, the process of publication can favor papers and projects that demonstrate statistically significant, and often novel, results. Even if a submitted work is significant, it faces highly selective examination by journal reviewers. The end result is a large number of empirical research studies that are not readily available, if they are preserved at all. Every year countless posters and treatise symposiums presented at any number of conferences may contain information pertinent to a meta-analytic review. This information may never be published and may suffer from the publication bias referred to as the file drawer effect. Unless exhaustive measures are taken by researchers to acquire unpublished, potentially nonsignificant, findings for inclusion, results from any analysis are likely to be biased or incomplete.

Paul B. Ingram and Michael S. Ternes

See also Effect Size; File Drawer Problem; Sample Size; Selection Bias

Further Readings

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., &

Oxman, A. D. (1993). Should unpublished data be included in meta-analyses? Current convictions and controversies. Journal of the American Medical Association, 269(21), 2749–2753.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449–472.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. Psychological Methods, 3, 486–504.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine, 6(7), e1000097. doi:10.1371/journal.pmed.1000097

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 125–138). New York, NY: Russell Sage.

Mira B. Kaufman Mira B. Kaufman Kaufman, Mira B.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Metacognition

Metacognition

1055

1057

Metacognition

The term *metacognition* refers generally to knowledge of and control over one's own cognitive processes. Within the context of educational research, metacognition is associated more specifically with the processes by which students can both understand and adjust their thinking and learning strategies to expand the limits of their existing knowledge. Metacognitive educational practices emphasize two goals—self-awareness and self-regulation—and both can be applied pragmatically in a range of academic contexts. This entry describes the integration of metacognition into theories of learning and education, the common metacognitive strategies that are taught to students to increase their learning abilities, and the assessment of metacognition in educational contexts.

The Integration of Metacognition Into Educational Research

Initially, metacognition was studied within the realm of developmental science. John Flavell is credited with first using the term while investigating the cognitive processes of young children. Researchers determined that metacognition develops gradually over the course of life, raising the question of whether metacognitive development, like other similarly developing cognitive skills, including communication and problem solving, could be supplemented by additional teaching.

Scientists next investigated the presence of metacognitive processes in expert populations to determine whether awareness of one’s cognition contributes to attaining higher levels of knowledge and skill. Metacognition was discovered to play a crucial role in the development of *adaptive expertise*, a term Giyoo Hatano and Kayoko Inagaki used to describe the ability to both excel at tasks previously completed and solve new, unrelated problems. Additional research confirmed that the metacognitive practices exhibited by experts could be taught to novice populations and could contribute to significant gains in learning and problem solving. In the 1990s, committees of both the American Psychological Association and the National Research Council publicly supported the use of metacognitive strategies in educational programs and contributed to the implementation of school reforms that included the addition of metacognitive instruction to the school curriculum. The integration of metacognition into theories of learning and education has led to increasingly widespread use of metacognitive strategies as an educational tool.

Metacognitive Strategies

In 2012, Kimberly Tanner developed the commonly used guide *Promoting Student Metacognition*, which outlines a set of metacognitive strategies that help students navigate the process of completing an academic task in biology class. The guide highlights the importance of utilizing two components of metacognition: metacognitive knowledge and metacognitive regulation. Despite the specific subject matter referenced in the guide, Tanner’s strategies have been adapted by educational instructors to fit a variety of academic areas, as they call on the pragmatic application of both metacognitive knowledge and regulation to increase academic performance.

Metacognitive Knowledge and Metacognitive Regulation

Metacognitive knowledge refers to the awareness of how one thinks and learns. There are three types of metacognitive knowledge. *Person variables* can recognize one’s own cognitive processes, including the strengths and weaknesses or styles of learning that one exhibits in different contexts. *Task variables* can determine about a particular task with respect to one’s own cognition, including how long it will take to complete or how difficult it is in comparison to a different task. *Strategy variables* are the tactics that one is aware of using regularly and flexibly to complete tasks and bolster the learning process. Metacognitive regulation is the capacity to use all three variables of metacognitive knowledge to actively control and enhance one's cognitive abilities. Both components of metacognition inform the strategies students can employ to gain awareness of and control over their learning abilities.

Learning Strategies

The first metacognitive strategy that Tanner's *Promoting Student Metacognition* endorses is the careful preassessment of students' existing comprehension of the task they are to complete, including individual person variables that may influence their performance. This preassessment can then be used to inform the planning phase, which should address the students' goals (e.g., what they are supposed to learn from the task) and utilize task variables to guide their actions (e.g., what first steps should be taken, how long the task should take to complete).

While completing the task, students are expected to monitor their progress. This strategy allows students to constantly evaluate their thinking processes, noting how well they are doing or how much they understand. During this phase, students are also able to regularly modify strategy variables (e.g., recognizing when to move on from a certain aspect of the task, determining how to solve a more difficult part) to proceed successfully.

Following the task's conclusion, students are instructed to complete a thorough postassessment, reflecting on whether and how their comprehension of the subject changed by the end of the task. This component provides students with the opportunity to identify aspects of the task that were most confusing as well as the methods by which they can improve their completion of the task in the future. Finally, students are encouraged to record in writing the results of the task, whether the thought processes they used for this task can be applied in other areas and how they can further adjust their learning strategies to enhance their cognitive abilities.

Assessment and Implications of Metacognition in Education

Several methods have been used to measure metacognition in an educational setting. Most common are self-report questionnaires and structured interviews that ask students directly about their awareness and regulation of their cognitive processes. Both domain-specific (relating to metacognition in an individual academic area) and domain-general (relating to metacognition across a range of academic areas) inventories can be implemented to assess students' metacognitive knowledge and regulation. Self-report and interview methods are contrasted with technologically based methods, such as the online collection of process data during an academic task. For example, a computer might record patterns of eye movement during a reading task that indicate when a student looked at an earlier part of the text or paused on a specific sentence. These data can be used to evaluate the student's comprehension or self-monitoring during the task.

The question of which research method should be used to assess metacognition has met with controversy, as each has advantages and limitations. Ideally, multiple methods of assessment should be used to measure students' metacognitive processes and these should then be compared. Although the methods of metacognitive assessment prove divisive, the use of metacognition as an educational tool does not. The integration of metacognition into the classroom curriculum in the late 1900s has been widely accepted as a means of improving students' information processing, problem solving, memorization, and test-taking abilities, thereby increasing learning capacities in student populations overall.

Mira B. Kaufman and Marc H. Bornstein

See also Cognitive Development, Theory of; Educational Psychology; Learning Theories

Further Readings

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. The nature of intelligence, 12, 231–235.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34, 906–911. doi:10.1037/0003-066X.34.10.906

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. W. Stevenson, H. Azuma, & K. Hakuta (Eds.), Child development and education in Japan (pp. 262–272). New York, NY: W. H. Freeman.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. Theory Into Practice, 41, 219–225. doi:10.1207/s15430421tip4104_3

Tanner, K. D. (2012). Promoting student metacognition. CBE–Life Sciences Education, 11, 113–120. doi:10.1187/cbe.12-03-0033

Brenda Hannon Brenda Hannon Hannon, Brenda

Methods Section

Methods section

1057

1058

Methods Section

The methods section of a research paper contains information about participants, the research design, the materials/apparatus (i.e., equipment, stimuli), the procedures, and the data analysis plan (when necessary). In smaller studies, the research design and procedures sections are combined. This entry describes each portion of the methods section and the process of writing this section.

Many researchers consider the methods section to be the most important section of a research paper because it provides the reader with information that is not only necessary for replication but also necessary for ultimately judging the validity of the study. Information in the methods section allows readers to determine, for instance, whether the study used the correct procedures for its particular research questions and whether the study is free of confounding variables. For this reason, it is important that a methods section provides accurate and specific descriptions as to how the experiment was done and the rationale for why each specific experimental procedure was selected and that this information is provided in a clear and logical fashion. When there is a large amount of information to convey, information should be arranged using subheadings for clarity purposes. Finally, sections should be written with the most important information first and the least important information last.

Participants

The participants subsection is the first subsection of a methods section. This subsection should provide the reader with enough information to judge whether the research can be generalized to a larger population. At a minimum, it (a) informs the reader about who was in the experiment (i.e., the number of participants, how they were selected, the number of participants assigned to each experimental condition), (b) describes the population tested (i.e., age, gender, ethnicity, and other variables relevant to the experiment), and (c) reports any restrictions that were imposed on the participant pool (e.g., excluded learning disabled and deaf). An example of a participants section is “Sixty children from St. Helen’s School (27 boys, 33 girls) ranging in age from 10.5 to 12 years old, received $10 for participating in this study. All children completed all of the conditions.”

Research Design

The research design subsection, which is the second subsection of a methods section, should describe all baseline and experimental conditions. It should provide a description (i.e., operationalization) for all of the independent variables, including the levels of manipulation for each variable, whether an independent variable is a between-subjects variable or a within-subjects variable, and a discussion of the counterbalancing of the between-subjects variables. In addition, the research design section describes the dependent measure(s), including the unit of measure. Finally, the research design subsection typically lists the order of the tasks as they are administered.

Materials/Apparatus

This section provides a description of equipment, physical settings, and stimuli that are important to the experiment. When describing each of these components, it is important to first evaluate the importance of that component to the overall study. For instance, with respect to equipment, the greater the need for precise measurement, the greater the need to be more specific about the description of the equipment. If, for instance, a researcher is collecting reaction time data, then the researcher needs to describe the hardware (both processor and monitor) and software (e.g., programing language, data recording) used.

With respect to the environment of the experiment, the less common the environment, the greater the need for a more thorough description. For instance, if the data were collected in a lab room, then simply stating that it was collected in a lab room will likely suffice. However, if a researcher is collecting data in a local park, then it might be necessary to include details such as a description of the park's structures, the weather, and what people and animals were in the park to enable future replication of the experiment.

For stimuli, a number of experimental considerations need to be described in detail so that the study can be replicated. The first consideration is whether the stimuli are preexisting questionnaires/psychometric measures or experiment-generated stimuli. If the stimuli are preexisting questionnaires/psychometric measures, the materials/apparatus section needs to include (a) their reliability estimates, (b) how many subtests each measure has, (c) how many items per subtest, and (d) some example items of each subtest. If the stimuli are newly generated for this experiment (i.e., not preexisting questionnaires/psychometric measures), the materials/apparatus section needs to include (a) every parameter that is relevant to each new task (e.g., number of words, word length), (b) the number of items, (c) number of trials, (d) how the stimuli are counterbalanced, (e) if the stimuli were normed, and (f) reliability, if applicable.

Procedure

The purpose of the procedure section is to provide the reader with a summary of each step executed in the experiment. This summary of steps must be logical, concise, and precise. The procedure subsection includes the order of the tasks and, where logically appropriate, the instructions given to the participants.

Data Analysis

The last step in the methods section is an optional data analysis subsection. The decision to include this section is often based on the complication of the data analyses. If the data analyses are more complicated, researchers tend to include this subsection. If the data analyses are less complicated, researchers tend to not include this subsection.

The purpose of this section is to describe how the data will be presented in the results section (e.g., mean vs. median and accuracy vs. reaction time), to identify the statistical tests that have to be computed to infer results, and the level of significance that is set in order to identify statistically significant differences.

Other Special Considerations

Different research areas (e.g., education, medicine, physics) have different guidelines and specifications about labeling and formatting subsections of a methods section of a research paper. Different research areas also have varying specifications about terminology. Educational research typically follows the guidelines of the American Psychological Association.

Brenda Hannon

See also Abstracts; American Psychological Association; APA Format; Journal Articles; Literature Review; Results Section

Further Readings

Kallet, R. H. (2004). How to write the methods section of a research paper. Respiratory Care, 49, 1229–1232.

Amy Clark Amy Clark Clark, Amy

Minimum Competency Testing Minimum competency testing

1058

1061

Minimum Competency Testing

Minimum competency testing is a type of criterion-referenced assessment that requires examinees to demonstrate a minimum threshold of knowledge, skill, or ability in order to be deemed competent in the construct being measured. The assessments are usually used as part of a decision-making process, often as an exit exam or to award examinees with a credential based on their score. The entry that follows describes the application and history of minimum competency testing, including a highlight of relevant court cases regarding students' opportunities to learn content measured by minimum competency assessments.

As the name suggests, minimum competency exams tend to measure basic skills because they represent the bare minimum an individual would need to know in a content area. Minimum competency tests are not intended to differentiate between individuals along the full ability continuum but rather distinguish between individuals who have demonstrated adequate knowledge in the content area from those who have not. Furthermore, specific scores on the exam may not even be reported in lieu of a general indication of either pass or fail.

Minimum competency exams have many different applications. State and district education agencies have implemented tests of basic skills for annual grade promotion and for demonstrating graduation requirements in K–12 education. Credentialing exams, including those for certification and licensure, also require the examinee to demonstrate a minimum threshold of knowledge in the area in order to be deemed qualified for the credential being awarded. In each of these instances, high stakes are associated with the results of the assessment. Diplomas may be withheld, students may be retained in a grade for an additional year, or employment prospects may be impacted if the examinee does not demonstrate adequate knowledge of the necessary skills.

Because of the high stakes associated with results of minimum competency tests, the assessments must be carefully developed to meet the test specifications and follow guidelines set forth by *The Standards for Educational and Psychological Testing*. Test developers must provide evidence to support the intended interpretation and use of assessment results. Furthermore, consideration must be given to the reduction of measurement error and the impact of construct irrelevant variance on the results of the test. Due to the consequences associated with the results, the measure may eventually need to stand up against the rigor of a potential lawsuit, should individuals be negatively impacted by consequences associated with the assessment results.

Depending on the purpose of the test, the agency responsible for defining the knowledge, skills, and abilities required to demonstrate minimum competency may vary. In K–12 education applications, it may be the state education agency; local education agencies at the district level have also put forth basic skill requirements necessary for grade promotion. In credentialing applications, the assessment vendor may define the test specifications through job task analysis, and degree and preparation programs may cover the same content standards to prepare examinees prior to their taking the test. Similarly, the responsible agency must also set standards to determine the requirements for passing, and additionally define whether a specific overall score is required, whereby knowledge in one area can compensate for lack of knowledge in another or if the examinee must achieve a specified score on any subscore areas as well.

Proponents of minimum competency testing argue that setting minimum expectations for the knowledge, skills, and abilities that must be demonstrated for grade promotion or as a graduation requirement holds all students to a consistent standard of achievement in K–12 education. In credentialing applications, minimum competency requirements may protect the public against mistakes by requiring individuals to demonstrate a minimum level of competence prior to being certified or licensed for a given profession. By setting and communicating a specific standard of achievement, minimum competency testing can encourage students to take ownership of their education to ensure they meet the minimum requirement. Students know what is expected of them and must work toward the goal of achievement in those areas. Additionally, the implementation of minimum competency tests may also provide some impetus for schools to make programmatic changes that will lead to greater success for students who have traditionally struggled and potentially contribute to reducing the achievement gap.

Those opposing minimum competency testing argue that the negative aspects of such tests far outweigh the benefits. Imposing minimum competency thresholds may lead to increases in high school dropout rates and may adversely impact students with disabilities. Additionally, complaints have been lodged regarding whether adequate opportunity to learn the required content has been demonstrated prior to implementing graduation or grade promotion requirements. Other critics have argued the tests don't go far enough and that the minimum requirements are still too minimal to demonstrate that students are ready for postsecondary education or the workforce. Furthermore, the technical adequacy of such tests has also been questioned, specifically whether the tests themselves are backed by a strong validity argument supporting the interpretation of scores for their intended use, and additionally fueled by arguments that such high-stakes decisions should not be based off the results of a single measure but rather a collection of evidence attesting to the student's knowledge, skills, and abilities.

Both arguments for and against minimum competency tests have their merit, which has led to variation in implementation policy of minimum competency tests in K–12 education over time. The section that follows describes a brief history of the minimum competency testing movement in U.S. public education.

History

The modern conceptualization of minimum competency testing became increasingly popular in the 1970s as part of many state K–12 education programs. Led by efforts in Florida, California, and others, states quickly implemented minimum competency expectations in the mid-1970s due to an outcry that high school diplomas were being awarded based on education requirements that lacked rigor. The introduction of minimum requirements for graduation generally impacted a small proportion of students, primarily at the lower end of the distribution. Students in advanced coursework generally passed the test of basic skills on their first attempt.

By 1978, 33 states had implemented minimum competency standards as part of the requirements for earning a high school diploma. The remaining states that had not yet enacted minimum competency requirements were either in the process of approving legislation or had enacted state board of education studies in possible pursuit of adopting standards for minimum competency. The use of minimum competency tests also began to flourish at the district level in response to the sweeping legislation at the state level.

However, there was a shift in the widespread adoption of minimum competency requirements at the state level following the ruling in the *Debra P. v. Turlington* (1981) case in Florida. The minimum competency movement began to give way to state accountability methods and standards-based assessment. By 1984, 19 states had minimum competency requirements in place, and in 1996, the number of states requiring minimum competency exams for graduation had further reduced to 17. With the rise of the Common Core State Standards beginning in 2010, some states further chose to retire their high school minimum competency exams that did not cover the content of the standards. In 2015, California suspended its state minimum competency exam requirements for high school and allowed students who had failed the exam between 2006 and 2015 while meeting all other graduation requirements to receive their diplomas due to the assessment not aligning with the Common Core standards.

As state accountability policies change under the Every Student Succeeds Act adopted in 2015, the use of minimum competency testing for graduation requirements may continue to change and evolve. By looking to past implementation and court decisions regarding minimum competency testing, the impact on student outcomes when implementing these tests can be improved.

Court Cases Related to Minimum Competency Testing

The *Debra P. v. Turlington* case is perhaps the most well-known and widely referenced court case regarding minimum competency testing. It represented the first lawsuit to address the fairness surrounding the use of minimum competency testing as a graduation requirement and whether any students were disadvantaged as a result.

The lawsuit originated in response to Florida’s implementation of a 1976 minimum competency graduation requirement whereby students had to pass a functional literacy exam to receive a high school diploma, beginning with the graduating class of 1979. Students who did not pass the exam were to receive a certificate of completion in lieu of a traditional high school diploma. Students were given three opportunities to take the exam. However, by the third administration, a gap was evident in the passing rates between Caucasian and African American students. Based on the results of the exam, diplomas would be

African American students. Based on the results of the exam, diplomas would be withheld from approximately 2% of Caucasian students as compared to 20% of African American students who would otherwise graduate in 1979.

Ten individuals initiated a lawsuit in October 1978, alleging the assessment was biased against African American students due to the difference in testing outcomes across Caucasian and African American examinees. Their effort eventually resulted in a class action lawsuit on behalf of all students potentially impacted by the implementation of the minimum competency graduation requirement. Following the appeals process, the final ruling was that the state could not withhold diplomas from students on the basis of test results without first demonstrating the measure fairly assessed content that all students were being taught in their classrooms.

In addition to the *Debra P. v. Turlington* case, two 1983 lawsuits were filed on behalf of students in special education who did not meet minimum competency requirements and were denied high school diplomas: *Brookhart v. Illinois State Board of Education* and *Board of Education of Northport v. Ambach*. Rulings in these cases indicated students in special education must be given at least the same amount or more time as students in the general population to meet requirements for minimum competency when making graduation decisions.

Opportunity to Learn

As was demonstrated in the *Debra P. v. Turlington* case, and also referred to in educational and psychological testing standards, students must be given ample opportunity to learn the content being assessed when the organization responsible for curriculum and instruction is also imposing the testing requirement as is often the case in K–12 education applications. Opportunity to learn introduces complexity to the interpretation of assessment results, whereby students who have not been instructed on the content the assessment measures would not be expected to perform well and as such should not be penalized by the results of the assessment. Opportunity to learn also has fairness implications, whereby all students must be given the opportunity to learn the content regardless of their race, ethnicity, first language, or access to educational resources as is the case in rural districts or for students in special education.

In instances where minimum competency testing requirements are enacted, particularly for K–12 grade promotion or as a graduation requirement, the organization implementing the requirement must ensure students and teachers organization implementing the requirement must ensure students and teachers are given enough notice of the construct being assessed, so that students have the opportunity to learn all content measured by the exam.

Amy Clark

See also Accountability; Common Core State Standards; Criterion-Referenced Interpretation; Standard Setting; Standards-Based Assessment

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bishop, J. H., Mane, F., Bishop, M., & Moriarty, J. (2001). The role of end-of-course exams and minimum competency exams in standards-based reforms. Brookings Papers on Education Policy, 4, 267–345.

Linn, R. L., Madaus, G. F., & Pedulla, J. J. (1982). Minimum competency testing: Cautions on the state of the art. American Journal of Education, 91, 1–35.

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. The Phi Delta Kappan, 59(9), 585–588.

Popham, W. J. (1981). The case for minimum competency testing. The Phi Delta Kappan, 63(2), 89–91.

Pullen, D. (1983). *Debra P. v. Turlington:* Judicial standards for assessing the validity of minimum competency tests. In G. F. Madeus (Ed.), The courts, validity, and minimum competency testing (pp. 3–20). New York, NY: Springer.

Resnick, D. P. (1980). Minimum competency testing historically considered. Review of Research in Education, 8, 3–29.

Ryan W. Schroeder Ryan W. Schroeder Schroeder, Ryan W.

Phillip K. Martin Phillip K. Martin Martin, Phillip K.

Minnesota Multiphasic Personality Inventory Minnesota multiphasic personality inventory

1061

1064

Minnesota Multiphasic Personality Inventory

The Minnesota Multiphasic Personality Inventory (MMPI) was published in 1943 by the University of Minnesota Press. Since then, subsequent versions of the test have been created and published including the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; 1989), the MMPI-Adolescent (1992), and the MMPI-2-Restructured Form (2008). All these tests are standardized psychometric self-report questionnaires designed to assess a number of personality patterns and psychological disorders. This entry focuses on the MMPI-2, given that it is the most frequently used personality assessment tool in clinical psychological evaluations, it is commonly utilized for a variety of nonclinical purposes, and it has a substantial research base documenting its reliability and validity. The entry discusses the development of the MMPI and MMPI-2 and the psychometric considerations, administration, use, and interpretation of the MMPI-2.

Development of the MMPI and MMPI-2

In the early 1930s, psychologist Starke Hathaway and neurologist J. C. McKinley sought to develop a single multifaceted/multiphasic personality inventory for use in clinical settings. To do so, they compiled more than 1,000 items, some of which were from existing psychological inventories and others created from information gathered from psychiatric textbooks and the authors’ own clinical experiences. After removing duplicate items and items considered insignificant for the inventory, 504 items remained. Using these items, Hathaway and McKinley empirically derived a series of scales, which differentiated patients with specific clinical disorders from normal controls. Additional items were later added, primarily to assess sexual orientation, ultimately resulting in a total of 566 items. Eventually, 10 clinical scales were developed to differentiate patients with clinical conditions from individuals without those conditions. Validity scales were also developed and included within the test to assess consistency of responding, as well as over-and underendorsement of reported symptoms.

In 1989, the MMPI was revised and restandardized, resulting in the creation of the MMPI-2. When revising the test, the researchers sought to maintain continuity between the original version and the updated version. To do so, they made as few changes as possible to the clinical scales. Only a small number of items were deleted, and the majority of changes that were made were in rewording items to improve readability and/or eliminate outdated and sexist language. Although a number of new items were added to the test, the purpose of adding these new items was to create additional scales to assess content areas that were not covered by the original MMPI. After revision, the MMPI-2 contained a total of 567 items. The test was standardized on a normative sample of 2,600 adults (aged 18 years and older) from across the United States, whose age, ethnicity, and marital status were consistent with national census data.

Psychometric Considerations of the MMPI-2

The MMPI-2 manual reports internal consistency coefficients ranging from .34 to .87 for the clinical scales and from .72 to .86 for additional content scales. Given the generally low internal consistency of the MMPI-2 clinical scales, it is not surprising that factor analyses indicate that many of the scales are multidimensional. Regarding test-retest coefficients, the MMPI-2 manual indicates that the values range from .56 to .93 for the clinical scales and from .77 to .91 for additional content scales for a subsample of individuals from the normative sample. The test manual also reports that the length of interval between testing sessions had no effect on test-retest reliability. Follow-up studies have found similar results: test-retest correlations for college students were comparable to those from the MMPI-2 normative sample, and 5-year test-retest correlations for more than 1,000 men in an aging study were only slightly lower.

Aspects of test validity can be drawn from many of the MMPI-2 publications, which total over 3,500. Beyond this, though, it has been argued that because there is continuity between the MMPI and MMPI-2, studies from the original

MMPI can also be applied to the MMPI-2. When combining this literature, the number of publications increases to over 20,000, and multiple articles directly aimed at establishing test validity become evident. These articles include diverse experimental designs including classification paradigms, incremental validity comparisons, and empirical correlations for items, scales, and groups of scales (i.e., code types).

A meta-analysis of more than 400 control and psychiatric samples indicated that the MMPI was effective in discriminating between control and psychiatric groups, psychotic and neurotic groups, and anxiety and depression disorder groups. Additionally, extra-test correlates of scores on individual scales and code types have been found to be predictive of behaviors for nonclinical adults, college students, medical patients, psychiatric patients, prison inmates, and others. Overall, the diversity and number of studies within the MMPI/MMPI-2 research base helps provide evidence for construct and criterion validity of the test.

Administration and Use of the MMPI-2

The MMPI-2 can be used with adults aged 18 years and older. The test contains short individual items addressing emotions and behaviors, and these items are to be marked “true” or “false” depending on how the person has recently been feeling and behaving. Administration is straightforward and can be conducted via a self-report booklet or computer.

The test usually takes a person 1–1½ hours to complete depending on factors such as reading level (the test is designed to require a sixth-grade reading level, although some researchers recommend an eighth-to ninth-grade reading level) and degree of intellectual and cognitive impairment (individuals with an IQ of 70 or below will probably have difficulty completing the MMPI-2). In cases of physical, cognitive, or reading ability limitations, an audio reading of test items can be administered to patients in place of using the booklet administration. Although the English version of the test is the most common, the test has been translated into over 50 languages including Spanish, French, Italian, German, Chinese, Korean, and Hmong.

Once administered, the MMPI-2 can be scored via hand (using provided scoring keys), computer, or a mail-in scoring service. Raw scores are converted to linear

T scores for two of the clinical scales and all of the validity scales; however, most of the scales utilize uniform *T* scores. Across scales, the *T* scores have a mean of 50 and a standard deviation of 10. The test is copyright protected, and it must be purchased from an authorized supplier in order to administer and score. The cost varies based on method of administration, scoring, and interpretation.

Interpreting the MMPI-2

Once the test has been scored, the process of interpreting the profile can begin. Multiple sources of test interpretation assistance are available including an interpretation manual produced by the test developer, interpretation books published by experts on the test, and computerized interpretative printouts. Generally, interpretation begins by evaluating the validity scales to determine whether the testing is valid or not. The multiple validity scales can be conceptualized as examining three constructs: item omission (Cannot Say Scale), item response consistency (Variable Response Inconsistency Scale and True Response Inconsistency Scale), and honesty of responding (Infrequency Scale, Back Infrequency Scale, Infrequent Psychopathology Scale, Lie Scale, Correction Scale, Superlative Scale, Fake Bad Scale, and Response Bias Scale). The latter set of validity scales help the examiner to determine whether the individuals are presenting themselves in either an overly negative or positive manner.

After testing has been determined to be valid, interpretation of the main clinical scales and additional content scales can occur. The following are clinical scales originally named after the diagnoses they were meant to identify: Hypochondriasis (now called Scale 1), Depression (now called Scale 2), Hysteria (now called Scale 3), Psychopathic Deviate (now called Scale 4), Masculinity-Femininity (now called Scale 5), Paranoia (now called Scale 6), Psychasthenia (now called Scale 7), Schizophrenia (now called Scale 8), Hypomania (now called Scale 9), and Social Introversion (now called Scale 0). These can be interpreted individually and/or in a combined, code-type, fashion. The code types are often considered the core of the interpretive process, and much empirical research has demonstrated that certain code-type productions correlate with certain diagnoses, behaviors, personality traits, and feelings. However, these code types are based on probabilistic likelihoods and may or may not apply to certain individuals who obtain them. Thus, the additional content scales on the MMPI-2, along with the individual’s history and presentation, are also evaluated to confirm the likelihood that the code type is presentation, are also evaluated to confirm the likelihood that the code type is accurate.

Uses of the MMPI-2 in Various Settings

The MMPI-2 has widespread applications in psychological and psychiatric settings. The test can help to assess for a range of dimensions regarding response style, attitude, and approach. Specifically, validity scales can help psychologists to determine whether patients are presenting their symptoms with some degree of inconsistency, inaccuracy, and/or deception. Clinical and additional content scales can help psychologists clarify diagnostic issues including the presence or absence of psychopathology, types of probable clinical syndromes, and severity and chronicity of psychopathology.

The clinical and additional content scales can also aid psychologists in assessing suitability for psychotherapy, as measurable factors such as ego strength, socialization, ability to contain impulses, and ability to experience and accurately interpret emotions predict compliance with and successful outcomes related to psychotherapy. The test can also be readministered during the course of or after therapy to determine what degree of success the therapeutic interventions had in reducing symptomatology, increasing coping mechanisms, and improving self-esteem.

In medical settings, psychologists generally use the MMPI-2 to screen for psychopathology and substance abuse problems, determine whether somatic symptoms are due to organic or functional etiologies, understand the psychological effects of chronic or severe medical conditions and treatments, and understand how patients are likely to psychologically respond to medical interventions. Research has even indicated that the MMPI-2 might be useful at predicting who is likely to develop serious medical conditions such as coronary heart disease. In a survey of physicians who had access to psychologists administering the MMPI in medical settings, the vast majority of physicians indicated that the MMPI results were useful with their patients.

In personnel settings, the MMPI and MMPI-2 have been used to screen for psychopathology and to predict a number of variables including openness to evaluation, social facility, addiction potential, stress tolerance, and overall adjustment. In fact, research has indicated that the MMPI/MMPI-2 can be used successfully in selecting nurses, physician’s assistants, psychiatric residents, clinical psychology graduate students, clergy, successful business people, airline clinical psychology graduate students, clergy, successful business people, airline pilots, firefighters, police officers, probation officers, Air Force cadets, military aviators, nuclear power plant personnel, and others. Routine use of the MMPI-2 for personnel selection is generally not recommended, however, as many jobs require certain abilities or training experiences, whereas personality and psychological factors are considered to be relatively less important. In addition, the Americans with Disabilities Act has established guidelines about screening individuals, and thus, when the MMPI-2 is used in vocational settings, administration of the test is usually only undertaken after conditional offers of employment have been provided to the applicant.

Another area where the MMPI/MMPI-2 has been used extensively is in correctional settings. Researchers have found that inmates produce valid profiles approximately 80% of the time, and these valid profiles can sometimes be used to classify prisoners and to predict their behavior while incarcerated and after release. Normative data on criminal offenders have been compiled, and scores on various scales can be used to determine who might need mental health assessment and treatment due to a variety of factors including substance abuse, thought disorder, mood disorder, hostility, manipulation/exploitation, and anger. Additionally, other scales can be used to predict degree of hostility, conflict with authority, response to supervision, and likely benefit from academic and vocational programming.

Ryan W. Schroeder and Phillip K. Martin

See also Personality Assessment; Personnel Evaluation; Standardized Tests; *Standards for Educational and Psychological Testing*; *T* Scores; Testing, History of

Further Readings

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). MMPI-2: Manual for administration, scoring, and interpretation (Rev. ed.). Minneapolis: University of Minnesota Press.

Friedman, A. F., Lewak, R., Nichols, D. S., & Webb, J. T. (2000). Psychological assessment with the MMPI-2 (2nd ed.). Mahwah, NJ: Psychology Press.

Graham, J. R. (2012). MMPI-2 assessing personality and psychopathology (5th ed.). Oxford, UK: Oxford University Press.

Greene, R. L. (2011). The MMPI-2/MMPI-2-RF: An interpretive manual (3rd ed.). Boston, MA: Allyn & Bacon.

Nichols, D. S., & Kaufman, A. S. (2011). Essentials of MMPI-2 assessment. Hoboken, NJ: Wiley.

Wilfridah Mucherah Wilfridah Mucherah Mucherah, Wilfridah

Minority Issues in Testing Minority issues in testing

1065

1066

Minority Issues in Testing

Researchers have studied multiple areas involving the assessment of minority students’ learning and how schools’ assessment practices affect the education of these students. This entry looks at how minority students are affected by teachers’ choice of assessments to evaluate students’ progress and achievement, by the increasing importance of statewide standardized test results, and by the assessments used for participation in programs intended for gifted and high-achieving students.

Some aspects of student assessment are outside of teachers’ control. For example, the system they use for grading is typically determined by the school or school district. However, teachers still make many decisions about student assessment, such as how assignments are structured or the amount, length, and types of exams. It follows, then, that teachers’ individual differences in perspective can shape their assessment behaviors. Even the kinds of questions a teacher writes to assess students on an exam may unintentionally reflect that teacher’s unconscious biases or a lack of effort to integrate a variety of cultural perspectives in the classroom.

Several kinds of assessments are used to evaluate student learning. They include traditional tests, performance-based tests, projects, oral presentations, portfolios, and journals. Assessments fall into two general categories based on their purpose: formative and summative. Formative assessment occurs before or during instruction. Formative assessment is used to guide the teacher in planning and to help students identify areas that need work. Therefore, formative assessment helps form instruction. On the other hand, summative assessment occurs at the end of instruction. Its purpose is to inform the teacher and the students of the students’ level of accomplishment. The main differences between formative and summative assessment involve how the results are used and when formative and summative assessment involve how the results are used and when the measurement takes place (while students are learning or after learning has occurred). Teachers rely heavily on summative assessment, and statewide standardized testing also is categorized as summative assessment.

Educational policy changes, particularly the No Child Left Behind Act of 2001 (NCLB), have led to an increasing importance placed on statewide standardized testing that has significantly affected students, especially minority students. Although student performance on the National Assessment of Educational Progress rose in the decade after enactment of NCLB, there was not a corresponding narrowing of the achievement gap between White students and their African American and Latino counterparts. The Every Student Succeeds Act, adopted in 2015 to replace NCLB, gives states more flexibility in setting goals for student proficiency but maintains the old law’s annual testing requirement. William Penuel and colleagues have argued that states should choose school quality and student success indicators for accountability systems that are more inclusive and promote equity.

Assessment tools that emphasize high-stakes testing raise questions regarding culturally relevant instructional practices. Research examining African American students’ performance on high school exit exams has shown that when the classroom curriculum incorporates students’ input as well as other culturally relevant aspects, it improves students’ critical reading skills and performance on tests. In addition, there is research showing that when minority students are being evaluated, they tend to personally internalize failure, and that being placed in remedial classes can make students feel like lesser students and feel that they are perceived as lesser students. These feelings, in turn, negatively impact students’ performance on tests. There is also research indicating that the performance of African American and Latino students on high-stakes tests can be undermined by stereotype threat, which refers to the risk of confirming a negative stereotype of one’s group.

Creating a relaxing classroom climate tends to be beneficial to the learning of minority students. It is therefore important to understand the psychological effects of assessment on minority students and to use this understanding to build relationships with them based on mutual respect. Similarly, educators need to consider cultural differences and use this understanding to enhance assessment tools.

Minority students are underrepresented in Advanced Placement (AP) courses.

Admission to AP courses is based on students' GPA, teacher recommendations, and sometimes by parental request. Students enrolled in AP courses tend to do well on the PSAT test, so it has been suggested as a tool that could help identify minority students for participation in AP courses. Equally important is the representation of minorities and economically disadvantaged students in gifted programs. The assessment tools used to identify gifted students may not favor low-income and minority students, specifically African American and Latino students.

Wilfridah Mucherah

See also African Americans and Testing; Giftedness; Latinos and Testing; National Assessment of Educational Progress; Test Bias

Further Readings

Heillig, J., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. Educational Assessment and Policy Analysis, 30(2), 75–110.

Houchen, D. (2013). "Stakes is high": Culturally relevant practitioner inquiry with African American students struggling to pass secondary reading exit exams. Urban Education, 48(1), 92–115.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? Contemporary Educational Psychology, 26(3), 291–310.

Penuel, W., Meyer, E., & Valladares, M. R. (2016). Making the most of the Every Student Succeeds Act (ESSA)—Helping states focus on school equity, quality and climate. Boulder, CO: National Education Policy Center.

Richardson, C. C., Gonzalez, A., Leal, L., Castillo, M. Z., & Carman, C. A. (2016). PSAT component scores as a predictor of success on AP exam performance for diverse students. Education and Urban Society, 48(4), 384–402. doi:10.1177/0013124514533796

VanTassel-Baska, J., Johnson, D., & Avery, L. D. (2002). Using performance tasks in the identification of economically disadvantaged and minority gifted learners: Findings from project STAR. Gifted Child Quarterly, 46(2), 110–123. doi:10.1177/001698620204600204

Julianne Michelle Edwards Julianne Michelle Edwards Edwards, Julianne Michelle

W. Holmes Finch W. Holmes Finch Finch, W. Holmes

Missing Data Analysis

Missing data analysis

1066

1069

Missing Data Analysis

Missing data refers to a common problem that researchers face in all fields. In the context of testing and survey research, this phenomenon can occur due to participants running out of time, not knowing the correct answer, simply not wanting to answer the question, and even issues within the measurement. In the context of longitudinal research, missing data may result from study attrition. Missing data may also result from simple problems with data entry and management. In short, research paradigms, the phenomenon of missing data, is a potential problem.

The researcher must decide how to deal with missing data prior to using statistical analyses. Many (though not all) such analyses require each variable to be complete prior to its inclusion. If it is not dealt with appropriately, the presence of missing data can lead to biased statistical tests and parameter estimates. Luckily, there are myriad approaches for dealing with missing data that can help to reduce the problems associated with this occurrence. The selection of an appropriate technique for dealing with missing data is, however, largely dependent on the type of missing data that is present. This entry briefly reviews the types of missing data, then introduces some of the main methods for dealing with this potential problem, and concludes with some limitations of dealing with missing data.

Types of Missing Data

Missing data is typically described as coming from one of three sources: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR occurs when there is no systematic cause to a data value being missing. For example, an MCAR item response was left blank by the respondent completely by accident. With MAR, the missing data is not truly random in nature, but the variable associated with the missing data has been measured by the researcher. For example, if males are more likely to leave an item on a survey unanswered, and the researcher has collected data on the gender of the respondents, then the missing values would be considered MAR. Finally, MNAR occurs when the missing data is directly linked to the missing value itself. MNAR data would occur if an examinee taking a test were to leave an item missing because the examinee did not know the correct answer. Each of the methods that are discussed in the following section may be appropriate for certain types of missing data but not for others. This entry explores the various methods for dealing with missing data. Specifically, this entry elaborates on traditional missing data methods, maximum likelihood estimation, multiple imputation (MI), and methods for MNAR.

Missing Data Methods

Traditional methods for dealing with missing data are still commonly used throughout research, though in many cases, they have not proven to be very effective. These methods include deletion methods, single imputation methods, averaging items, and last observation. Deletion methods are commonly used and, while never optimal, are less harmful for MCAR data than for the other types described previously. When deletion methods are used, the cases with missing data are simply eliminated from the data set and not used in the analyses. However, the type of deletion method will dictate when missing data is removed. Perhaps the most popular and convenient method is listwise deletion (LD). When LD is used, all cases that have missing data are removed from the data set. This results in a data set with only those cases that are 100% complete. Data analyses are then conducted on this subset of complete data.

Another deletion method that is commonly used is pairwise deletion. Unlike LD, pairwise deletion does not completely eliminate cases with missing data to create a single complete data set. Instead, it only removes cases with missing data if, and only if, the analysis being used requires the variable with the missing data. For example, a researcher may use pairwise deletion in conjunction with estimation of a correlation matrix. A data point with a missing value for one estimation of a correlation matrix. A data point with a missing value for one variable will only be removed for the correlations using that variable but will be included in calculation of the other correlation coefficients. In contrast, with LD the individual would be removed from all correlation calculations, even though the individual’s responses were only missing data for one of the variables. Although deletion methods are popular, they create the disadvantage of removing data, which can in turn lead to inaccurate parameter estimation when the missing data is not MCAR, and low statistical power for all types of missing data.

Another traditional approach for dealing with missing data is single imputation. Unlike deletion methods, single imputation methods do not remove cases with missing data. Instead, the missing data is replaced through the generation of a replacement value for each missing data point. The data analysis of interest (e.g., regression) would then be conducted on this revised data set that includes data for all observations. However, like deletion methods, when most single imputation methods are used, biased parameter estimates are still produced for MAR data and even MCAR data. Single imputation methods will also underestimate sampling errors, as the replaced values are treated as real data and not distinguished as missing. But even though there are disadvantages in using single imputation methods, there are a variety of approaches to choose from that vary based on how the one replacement value is generated, with some methods performing better than others. These methods include arithmetic mean imputation, regression imputation, stochastic regression imputation, hot-deck imputation, and similar response pattern imputation. Of these methods, the missing data must be MCAR, except for stochastic regression imputation. Rather, stochastic regression imputation can be used for MCAR and MAR without producing biased estimates. For this reason, stochastic regression imputation is one of the best single imputation methods.

Stochastic regression imputation is unique among single imputation methods, though it does share some traits with regression imputation. Both approaches rely on a regression model to impute the missing data. However, what makes stochastic regression imputation unique is that it adds a random value to the prediction from the regression model. By adding this random number to the values imputed from the regression model, stochastic regression imputation alters the imputed values from

$$yi = \beta 0 + \beta 1,$$

where yi is the replacement value, β0 is the intercept, and β1 is the slope of the where yi is the replacement value, β0 is the intercept, and β1 is the slope of the missing value, to

$$yi = \beta 0 + \beta 1 + zi,$$

where zi is the random value and is generated from the normal distribution with a mean of 0, and variance equal to the variance of residuals from the regression model. This additional, random information acknowledges the fact that the imputation is merely an estimate of what the actual value would have been and that the imputation itself is almost sure to be not exactly correct.

The final two traditional methods differ from the methods previously discussed, as these methods are used for specific types of research. First, averaging the available values for the variable in question is common, particularly when using an instrument that computes a scale score with multiple items that measure a single construct. When missing data occurs for this type of scenario, this method will average the items the participants did respond to in order to create the scale score. For example, if a participant responded to only 18 of the 20 items, this participant's scale score would be the average of the 18 items that were responded to and then multiplying the average by the total number of items (e.g., 20).

Last, observation carried forward is specific for longitudinal designed research, and the last observation is used to fill the missing time points. For example, if a participant drops out of the study in the 8th week of a 10-week study, the participant's data from the 7th week will be used for the remaining 3 weeks.

Although traditional methods have been used, and are still utilized today, these methods can lead to biased estimates. Therefore, traditional methods should be avoided when full information maximum likelihood estimation (FIML) or MI can be used. FIML and MI are both excellent approaches to use for either MCAR or MAR data. In such cases, neither biased estimates are produced nor statistical power is maximized because all available information is used in the observed data. FIML is a popular method for estimating parameters for latent variable models and in regression. Essentially, FIML estimates the parameter values for the model, filtering out observations with missing data when that data value would be used for parameter estimation. On the other hand, when the observation is not missing a data point being used in the parameter estimation, it is included. For example, if observation A is missing a value for variable *x*, but not for *y* or *z*, then parameter estimation involving *x* will not include observation

A, but estimation involving *y* or *z* will include observation A.

MI fundamentally differs from FIML as it is a data imputation approach. However, instead of a single imputed value being generated for a missing data point, multiple values are generated. MI replaces the missing data by using available information from all variables. Imputed values are generated for each missing value and a random value is added to each, much as with stochastic regression imputation. This is done *m* times, and the analysis of interest is then conducted using each of the resulting data sets. By creating multiple data sets, MI acknowledges the uncertainty inherent in the imputed values. MI can incorporate information from all available variables into the imputation process, providing more accurate imputations.

One popular MI method is joint modeling multiple imputation. Joint modeling multiple imputation works by first making an assumption about the probability model underlying the data (e.g., multivariate normal, multinomial). Next, parameter estimates are calculated from the Bayesian posterior distribution created using the Markov Chain Monte Carlo method of data augmentation based on the probability model, observed data, and a prior distribution. The resulting posterior distribution includes imputed values. This process is repeated *m* times to create complete data sets, each of which is used in the analysis of interest (e.g., regression) with the results then combined. There are also several relatively new imputation methods, including multivariate imputation by chained equations, random forest imputation, and extensions of multivariate imputation by chained equations that incorporate recursive partitioning.

Each of the more advanced methods previously described can be used with MCAR and MAR data. Very recently, methods have been described for use with MNAR data, though these are less popular in practice than MCAR-and MAR-based methods. This relative lack of popularity is largely due to the assumptions underlying these MNAR methods and lack of methods to check these assumptions. Violations of these assumptions can result in severely biased parameter estimates. Thus, extreme caution must be used when attempting to use MNAR methods. Given these limitations, the MNAR methods will not be elaborated on here; however, for those interested readers, two MNAR models to consider are the selection model and the pattern mixture model. More research should continue to be devoted to MNAR data. This may ultimately lead to more reliable missing data methods for MNAR data in the future.

Limitations

Each of the methods described in this entry have been used and continue to be used throughout all fields of research. However, not all of these are appropriate for use in all (or sometimes any) situations. Thus, the researcher must make some considerations prior to selecting a method for dealing with missing data. The type of missing data that is present should be considered. As previously mentioned, some missing data methods are more appropriate for certain types of missing data (i.e., MAR, MCAR, and MNAR) than others. Because of this, it is important for researchers to consider the mechanism underlying the missing data present in their data and select from only those methods that can adjust their data based on the missing data type. Another issue to consider is the availability of the missing data methods in the software that is being used. Depending on the type of missing data technique, statistical software such as Amos, EQS, LISREL, MPLUS, NORM, SAS, SPSS, or R may be required.

Julianne Michelle Edwards and W. Holmes Finch

See also Maximum Likelihood Estimation; Structural Equation Modeling

Further Readings

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40–49. doi:10.1002/mpr.329

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. Educational and Psychological Measurement, 64, 419–436.

Enders, C. K. (2010). Applied missing data analysis (methodology in the social sciences). New York, NY: Guilford Press.

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P., & Catalano, P. (2010). Multiple imputation in a large-scale complex survey: A practical guide. Statistical Methods in Medical Research, 19(6), 653–670. doi:10.1177/0962280208101273

Leite, W., & Beretvas, S. N. (2010). The performance of multiple imputation for Likert-type items with missing data. Journal of Modern Applied Statistical Methods, 9(1), 64–74.

Wolkowitz, A. A., & Skorupski, W. P. (2013). A method for imputing response options for missing data on multiple-choice assessments. Educational and Psychological Measurement, 73(6), 1036–1053. doi:10.1177/0013164413497016

Joseph A. Maxwell Joseph A. Maxwell Maxwell, Joseph A.

Mixed Methods Research Mixed methods research

1069

1073

Mixed Methods Research

The term *mixed methods research* is generally used to refer to research that combines quantitative and qualitative research approaches and methods in the same study. Some researchers include studies that combine different quantitative methods, or different qualitative methods, but the term *multimethod research* is more commonly used for these. Many prominent mixed methods researchers add that such studies should involve an actual integration of the results of the two methods, rather than simply being separate strands of a study with no real interaction. This entry explains the differences between qualitative and quantitative research and describes the history of mixed methods research, key issues in its development, important concepts and strategies in its use, and current controversies in the field.

Qualitative and Quantitative Research

There is no agreement on a precise distinction between quantitative and qualitative research; both include quite diverse approaches and methods, and there are multiple differences between the two, none of which are entirely definitive. The simplest (and common) distinction, that quantitative research involves numbers and qualitative involves only words, is clearly inadequate; both fields use numbers (although quantitative research relies much more heavily on them), and the numbers/words distinction fails to capture other commonly invoked differences between these approaches, including the use of artificial versus natural settings, a primary reliance on deductive versus inductive strategies, and a positivist versus constructivist epistemology. None of these distinctions adequately capture the diversity of strategies within each approach, and none definitively distinguish the two approaches.

However, the distinction is extremely meaningful to researchers in both communities and was central to the development of mixed methods research. A different way of distinguishing the two approaches, in terms of their strategy for explanation, was proposed by the evaluation researcher Lawrence Mohr and may be helpful in clarifying the differences. Mohr identified two types of explanation, which he termed *variance theory* and *process theory*. Others had earlier presented similar distinctions but had not developed them as systematically. Variance theory is based on the concept of a variable, a property of something that can vary, and can be measured or categorized. This concept is fundamental to quantitative research; essentially, all such research involves the creation and correlation of different variables, or comparison of the values of particular variables, across persons or other units of analysis. The use of variables allows precision in counting or measuring social phenomena, determining differences between individuals or groups on particular variables, and identifying relationships between variables.

Qualitative research, in contrast, makes very little use of variance theory; although some qualitative researchers use the term *variable*, they do not employ it in the same way as quantitative researchers. Instead, they focus on describing the phenomena studied (behavior, meaning, experience, and social organization) in a specific context and understanding the processes (physical or mental) that connect these phenomena, thus being labeled as process theory by Mohr.

These differences are fundamental to understanding the explicit development of mixed methods research because conflicts between advocates for the two approaches were intrinsic to this development and because the complementarity between quantitative and qualitative research, in terms of their strengths and limitations, provides the main rationale for combining the two approaches. Quantitative research is better at answering “what” and “how much” questions, such as, “Did this educational program make a difference in academic achievement for these students, and how much of a difference?” Qualitative research is better at answering “how” and “why” questions, such as, “How was the program experienced and understood by participants, and how did this shape their responses; how was it influenced by the particular context in which it occurred; and how and why did it achieve these results?” The answers to both types of questions are important for policy and practice, and a mixed methods study is much more capable of answering both.

History of Mixed Methods Research

In 1959, the explicit emergence of mixed methods research has been traced to work by Donald Campbell and Donald Fiske on what they called the *multitrait–multimethod matrix*, followed by increasing discussion of the possibility of combining quantitative and qualitative methods in the 1970s and 1980s. However, the actual use and integration of quantitative and qualitative methods and data have a much longer history, particularly in the natural sciences, although the latter are almost never addressed in the mixed methods literature. In the physical sciences, the joint use of both methods can be found at least as far back as Galileo’s telescopic observations in the early 1600s, which combined visual description (e.g., of the topography of the moon, and the phases of Venus) with mathematical calculation and measurement. Similarly, field research in geology involves both descriptive observation and quantitative measurement. In biology, the work of ethologists such as Karl von Frisch, Konrad Lorentz, Niko Tinbergen, and Jane Goodall has also integrated qualitative observation and description with numerical data.

In the social sciences, the deliberate integration of qualitative interviewing and observation with survey data and social statistics dates at least from W. E. B. DuBois’s *The Philadelphia Negro* (1899), and such integration continued in classic sociological works such as *Middletown* (1929), *Yankee City* (1941), and many other studies through the 1960s, although becoming less common with the rise of statistical methods and quantitative research. In anthropology, quantitative data collection and analysis have frequently been integrated with ethnographic fieldwork since the 1920s, and psychologists such as Leon Festinger and Stanley Milgram also combined both methods in their research. There was a widespread recognition that both quantitative and qualitative methods had limitations and that combining the two could provide important benefits.

However, such combinations were not seen as a specific *type* of research, and within these studies, conflicts between proponents of the two approaches were largely absent; the lead researchers were typically involved in collecting and analyzing both types of data. Such conflicts became prominent after about 1970, due in part to the increasing dominance of quantitative research in prestige and funding, and led to what has been called the “paradigm wars” of the 1980s and 1990s.

Paradigm Issues

Paradigm Issues

The idea of a paradigm, popularized by Thomas Kuhn’s influential *The Structure of Scientific Revolutions*, became a key issue in the conflict between quantitative and qualitative research in the 1980s. In his 1969 postscript to this work, Kuhn described a paradigm as “the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community.” However, within the social sciences, this term came to refer mainly to the philosophical and ethical presuppositions of the different approaches, which were assumed to be foundational for each approach. Quantitative research was generally seen as based on positivism or postpositivism, which emphasized objective measurement and researcher neutrality. Qualitative research was claimed by many of its proponents to be based on constructivism (the view that reality was socially constructed, rather than being an objective entity), critical theory (incorporating ethical values and working against the oppression of disempowered groups), and/or postmodernism as alternative paradigms to positivism and postpositivism.

This entry cannot discuss in detail these paradigm debates, but they have played an important role in the development of mixed methods research since the 1980s. Prominent qualitative researchers such as Egon Guba and Yvonna Lincoln, adapting Kuhn’s idea of the “incommensurability” of paradigms, argued that qualitative and quantitative research, being based on different paradigms, were therefore incompatible and could never legitimately be combined in a single study.

Although earlier presentations of mixed methods focused mainly on combining data collection and analysis methods, the paradigm wars forced proponents of combining methods to address the broader issues involved in combining research approaches and not simply methods. In response, some mixed methods researchers chose to simply ignore philosophical debates and do whatever they believed worked to produce useful results, a stance that was variously labeled “pragmatic” or “a-paradigmatic.” Others claimed that *philosophical pragmatism*, as developed by John Dewey and others, resolved these issues and was thus the appropriate paradigm for mixed methods research. This led some proponents to claim that mixed methods research was itself a third paradigm, in addition to quantitative and qualitative research. Still others argued that multiple philosophical stances could be employed in mixed methods research and that paradigms, and not simply methods, could be mixed.

These different positions continued to be argued into the 20th century but generally less vituperatively. However, some quantitative researchers still treated qualitative research as less than fully scientific, a tendency that often characterized the movement promoting what has been called evidence-based research, with randomized controlled trials as the gold standard. Similarly, some qualitative researchers viewed mixed methods research with suspicion, seeing it as simply “positivism in drag.”

Current Recognition and Importance

Despite these disagreements, mixed methods research had become well established by 2000, with textbooks appearing as early as 1989 and proliferating after 2000. The first edition of the *SAGE Handbook of Mixed Methods in Social and Behavioral Research* was published in 2003 and the second (entirely new) edition in 2010; the *Journal of Mixed Methods Research* was founded in 2007, and the Mixed Methods International Research Association was established in 2014. Mixed methods studies are now commonly published in top-ranked peer-reviewed journals in the social sciences and have been funded by many governmental and nonprofit agencies. Courses in mixed methods research are now often given in universities and through various research organizations. However, there are still ongoing debates about important issues.

Recent Developments and Controversies

Research Design

Research design is the central issue for mixed methods research because the conception of design has been quite different in quantitative and qualitative research and even between different types of quantitative research. In experimental research, design has usually referred to particular types of research strategies, such as randomized experiments and different forms of quasi-experimental and single-subject research. Nonexperimental quantitative research has not usually been conceptualized in this way, although categorization in terms of the types of statistical analysis employed, such as structural equation modeling and hierarchical linear modeling, is common. Qualitative research, in contrast, lacks any explicit design categories; although works on qualitative research often distinguish between different approaches to research, such as grounded theory, phenomenology, and narrative research, these typically include grounded theory, phenomenology, and narrative research, these typically include philosophical and theoretical stances as well as methodological ones and aren't usually thought of as designs.

The dominant conceptualization of design in mixed methods research has been similar to that in experimental research, of defining specific types of mixed methods studies, based on criteria such as the order in which the methods are used, the relative dominance of the different methods, the degree of integration of the methods and results, and the purposes for which the methods are combined. However, this approach has been criticized by some mixed methods researchers, and alternative conceptions of design have gained increased recognition.

The most prominent of these alternatives has been termed an interactive, systemic, or dynamic approach that sees design as the relationships and mutual influences of the different components of a research study. These components include the study's goals, conceptual framework or theory, research questions, methods, and validity issues. The research questions are seen as the center or hub of the system, and influence and are influenced by all of the other components. This model is much closer to qualitative conceptions of research, in which design is an inductive and flexible aspect of a study that can adapt to unexpected developments or results.

Research Questions and Methods

The mixed methods community has been divided on the nature of appropriate research questions for mixed methods research. Some authors have argued that mixed methods studies require (in addition to possible quantitative and qualitative research questions) specifically mixed methods questions, ones that can be answered only by integrating qualitative and quantitative data. Others do not believe that this is necessary, seeing many questions as potentially or partially answerable by either qualitative or quantitative methods and arguing that the linking of qualitative and quantitative questions, in creating a broader and more inclusive understanding of the phenomena studied, is a legitimate goal of mixed methods research.

Closely connected to the different conceptions of design, there has also been disagreement over the relationship between research questions and research methods. For some researchers, the research questions are the primary and determining component, and the methods must follow from this (a view more characteristic of quantitative research). For others (and this is implied by the interactive concept of design), the research questions, though fundamentally important, need to be responsive to how the methods actually play out in practice, and to unexpected findings, validity threats, or theories that emerge during the study.

The Integration of Qualitative and Quantitative Methods and Data

As noted earlier, the goal of integrating qualitative and quantitative methods and data, rather than keeping them as separate strands of a study, has typically been seen as a defining feature of mixed methods research. However, the ways in which this integration can be accomplished have been inadequately studied and theorized, and researchers have received little guidance in how to do this.

Earlier mixed methods studies, those prior to the emergence of mixed methods as an explicit type of research, typically emphasized what later became known as *triangulation*—the use of one approach to test or confirm the results of the other. However, some of these studies also combined methods to provide a broader and more in-depth understanding of the phenomena studied, based on the complementarity of the two approaches in focusing on different aspects of these phenomena. For example, a quantitative approach could be used to rigorously measure the effect of a new educational program on student achievement, and a qualitative approach to understand how the program was perceived by teachers and students and the ways in which it was implemented in the settings studied.

The paradigm debates of the 1970s and 1980s problematized both of these approaches, raising the issue of whether such philosophically divergent approaches could in fact be combined in these ways. As this debate waned, however, additional purposes for combining methods emerged. In some studies, the two methods were not used concurrently but sequentially. This allowed the use of one method to develop the second method—for example, by using qualitative interviews or observations to develop a survey questionnaire or by further exploring the results of a quantitative survey through focus groups. Such sequential designs have become a prominent part of mixed methods. However, attempts to use the concurrent versus sequential distinction as a basis for a typology of mixed methods designs do not accommodate many studies in which the relationship is more complex than this—for example, iterative designs in the relationship is more complex than this—for example, iterative designs in which there is alternation or partial overlap in time between different methods or in which the sequencing of approaches in data collection and data analysis is different. Despite this proliferation of purposes and strategies for integrating methods, there has still been little explicit theorization of how such integration is done. Until this occurs, the most productive way to understand the integration of methods is to read accounts of exemplary mixed methods studies.

Joseph A. Maxwell

See also Constructivist Approach; Paradigm Shift; Positivism; Postpositivism; Qualitative Research Methods; Quantitative Research Methods; Triangulation

Further Readings

Bergman, M. (Ed.). (2008). Advances in mixed method research. Thousand Oaks, CA: SAGE.

Greene, J. (2007). Mixed methods in social inquiry. San Francisco, CA: Jossey-Bass.

Guba, E. (Ed.). (1990). The paradigm dialog. Thousand Oaks, CA: SAGE.

Irwin, S. (2008). Data analysis and interpretation: Emergent issues in linking qualitative and quantitative evidence. In S. N. Hesse-Biber & P. Leavy (Eds.), Handbook of emergent methods (pp. 415–435). New York, NY: Guilford.

Maxwell, J. A., Chmiel, M., & Rogers, S. (2015). Designing integration in multimethod and mixed methods studies. In S. Hesse-Biber & B. Johnson (Eds.), Oxford handbook of multimethod and mixed methods research inquiry (pp. 223–239). Oxford, UK: Oxford University Press.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). Handbook of mixed methods in social and behavioral research. Thousand Oaks, CA: SAGE.

Tashakkori, A., & Teddlie, C. (Eds.). (2010). Handbook of mixed methods in social and behavioral research (2nd ed.). Thousand Oaks, CA: SAGE.

Trend, M. (1979). On the reconciliation of qualitative and quantitative analyses: A case study. In T. Cook & C. Reichardt (Eds.), Qualitative and quantitative methods in program evaluation. Thousand Oaks, CA: SAGE.

Sohad Murrar Sohad Murrar Murrar, Sohad

Markus Brauer Markus Brauer Brauer, Markus

Mixed Model Analysis of Variance Mixed model analysis of variance

1074

1078

Mixed Model Analysis of Variance

The characteristics of the design and the variables in a research study determine the appropriate statistical analysis. A mixed model analysis of variance (or mixed model ANOVA) is the right data analytic approach for a study that contains (a) a continuous dependent variable, (b) two or more categorical independent variables, (c) at least one independent variable that varies between-units, and (d) at least one independent variable that varies within-units. “Units” refer to the unit of analysis, usually subjects. In other words, a mixed model ANOVA is used for studies in which independent units are “crossed with” at least one of the independent variables and are “nested under” at least one of the independent variables.

Mixed model ANOVAs are sometimes called split-plot ANOVAs, mixed factorial ANOVAs, and mixed design ANOVAs. They are often used in studies with repeated measures, hierarchical data, or longitudinal data. This entry begins by describing simple ANOVAs before moving on to mixed model ANOVAs. This entry focuses mostly on the simplest case of a mixed model ANOVA: one dichotomous between-subjects variable and one dichotomous within-subjects variable. Then, it briefly presents more complex mixed model ANOVAs and discusses these analyses in the context of linear mixed effects models.

Simple ANOVAs

Between-units (e.g., between-subjects) ANOVAs are characterized by units that are “nested under” one or more categorical independent variables. Between-subjects ANOVAs examine the differences between two or more independent groups. For example, a simple one-way between-subjects ANOVA may test whether girls or boys have better grades in school. Here, there is one dichotomous independent variable that varies between-subjects (gender). The goal of the ANOVA is to examine whether the mean scores for each group (boys vs. girls) are reliably different from each other. The statistical model can be described as

$$Y = b_0 + b_1 X + e,$$

where *Y* is the dependent variable (scores), *X* is the dichotomous independent variable (gender), and *e* refers to the residuals (the errors). If the coefficient b_1 is statistically significant, one would conclude that the data provide evidence for the idea that one of the two genders has better grades than the other. Between-subjects ANOVAs are more flexible than independent samples *t* tests because they allow for multiple independent variables with two or more levels each.

Within-units (e.g., within-subjects) ANOVAs are characterized by units that are “crossed with” one or more categorical independent variables. They frequently examine differences between one measurement of a particular variable and another measurement of the same variable for a given subject. In such cases, the observations are not independent of each other in that two data points from the same subject are likely to be more similar to each other than two data points from two different subjects. Within-subjects ANOVAs examine the differences between two or more dependent groups. Their goal is often to examine changes in an outcome variable over time. For example, a one-way, within-subjects ANOVA may test whether students have better grades in English or math. Here, there is one dichotomous independent variable that varies within-subjects (discipline: English vs. math). The statistical model can be described as

$$(Y_1 - Y_2) = b_0 + e,$$

where Y_1 is subjects’ English grade and Y_2 is subjects’ math grade. Like before, the *e* refers to the residuals (the errors). If the coefficient b_0 is statistically significant, one would conclude that the data provide evidence for the idea that students’ English and math grades differ from each other. Compared to paired samples *t* tests, within-subjects ANOVAs are more flexible because they allow for multiple independent variables with two or more levels each.

2 × 2 Mixed Model ANOVAs

A mixed model ANOVA is a combination of a between-unit ANOVA and a within-unit ANOVA. It requires a minimum of two categorical independent variables, sometimes called factors, and at least one of these variables has to vary between-units and at least one of them has to vary within-units. The explanations that follow focus on the simplest possible mixed model ANOVA, a so-called 2 × 2 mixed model ANOVA: one dichotomous between-subjects variable and one dichotomous within-subjects variable. To better understand mixed model ANOVAs, consider the following research study.

A group of researchers is interested in comparing boys’ and girls’ grades in English and math. Let’s assume they are predicting a gender difference (girls have better grades than boys) and they expect this gender difference to be greater in English than in math. In this example, there are two independent variables. The first is gender (boy vs. girl), a dichotomous between-subjects variable. The second is discipline (English vs. math), a dichotomous within-subjects variable. For ease of interpretation, let’s assume that the data confirm the researchers’ hypotheses.

The study just described has a classic 2 × 2 design, and its data can be analyzed with a two-way mixed model ANOVA. This data analytic approach allows researchers to test whether there are main effects for both gender and discipline. A main effect is the effect of a particular independent variable, averaging across all levels of the other independent variable(s). The data analytic approach also allows researchers to test whether there is an interaction between the two independent variables. An interaction is present when the effect of one independent variable is stronger at one level of the other independent variable than at the second level of that same independent variable. A mixed model ANOVA tests whether each of the three effects—the two main effects and the interaction effect—is statistically significant.

These three effects can be obtained with the following statistical models:

$$(Y_1 + Y_2) / 2 = b_0 + b_1 X + e,$$

$$(Y_1 - Y_2) = b_0 + b_1 X + e,$$

where Y_1 is subjects’ grades in English, Y_2 is subjects’ grades in math, X is the dichotomous between-subjects variable (gender), and e refers to the residuals (the error) in the model. The term on the left side of the equations is simply the average (Equation 3) or the difference (Equation 4) of the grades.

The coefficient b_1 in Equation 3 represents the main effect of gender, the between-subjects independent variable. If b_1 in Equation 3 is statistically significant, one would conclude that girls on average have reliably higher or reliably lower grades than boys, regardless of discipline. The coefficient b_0 in Equation 4 estimates the effect of discipline, the within-subjects independent variable, for students with a score of zero on X (gender). If X is coded 1 and 2, then this coefficient is rather meaningless. However, if X is “centered” (i.e., coded −.5 and +.5, or −1 and +1), then b_0 in Equation 4 represents the main effect of discipline. If it is statistically significant, one would conclude that the students, regardless of their gender, performed better in one of two disciplines. The coefficient b_1 in Equation 4 represents the interaction effect between gender and discipline. If this coefficient is statistically significant, one would conclude that the gender difference is greater for one of the two disciplines. The coefficient b_0 in Equation 3 is the grand mean (the average of all scores) and is usually not interpreted.

Note that many menu-based data analysis programs (like SPSS) will automatically center the dichotomous between-subjects variable (X) for the user when the appropriate module is chosen. When using other, more code-based programs, researchers may have to recode the between-subjects variable by hand to make sure it is centered prior to estimating the model described in Equation 4 if they want to interpret the main effect of the within-subject variable.

Advanced Mixed Model ANOVAs

Mixed model ANOVAs are not limited to dichotomous independent variables. For example, they can contain within-subjects independent variables with more than two levels. Imagine a group of researchers interested in comparing boys’ and girls’ grades in English, math, and biology. They are predicting a gender difference (girls have better grades than boys) and they expect this gender difference to be greater in English than in the other two disciplines (math and biology). Now, the within-subjects independent variable (discipline) has three levels (English, math, and biology).

In the case of such data, the study has a 2 × 3 factorial design that can also be analyzed with a mixed model ANOVA. The data analytic approach is the same as before examining two main effects and an interaction effect, but the within-subjects independent variable will most likely be examined with a specific contrast. Given that the researchers predict the gender difference to be greater in English than in the other two disciplines, the appropriate contrast would be 1, −.5, −.5 (which produces the same *F* and *p* values as the contrasts 2, −1, −1, and .67, −.33, −.33, respectively).

These main and interaction effects can be obtained with the following models:

$$(Y_1 + Y_2 + Y_3) / 3 = b_0 + b_1 X + e,$$
$$(Y_1 - ((Y_2 + Y_3) / 2) = b_0 + b_1 X + e,$$

where Y_1, Y_2, X, and e have the same meaning as in Equations 3 and 4. Y_3 is students’ grades in biology. The term on the left side of the equation is the average (Equation 5) or the weighted difference (Equation 6) of the grades.

The coefficient b_1 in Equation 5 represents the main effect of gender. It tests whether girls have on average reliably better or reliably worse grades than boys, regardless of discipline. The coefficient b_0 in Equation 6 corresponds to the effect of the within-subjects contrast. If this contrast is statistically significant, one would conclude that the students, regardless of their gender, have higher grades in English than in math and biology. The coefficient b_1 in Equation 6 describes the interaction between the within-subjects contrast and gender. If it is statistically significant, one would conclude that the gender difference is greater in English than in the other two disciplines. Like before, the coefficient b_0 in Equation 5, the grand mean, is usually not interpreted.

Researchers may also decide to include covariates—sometimes called confounding variables or concomitants—in their analyses. These are variables that are not of primary interest to the study but may affect the outcome variable. For example, students’ parental income can provide them with resources that may influence their grades. Thus, the researchers decide to measure parental income and to account for the effects of this variable in the statistical analysis. Here, a mixed model ANOVA with a covariate—called a mixed model analysis of covariance (or mixed model ANCOVA)—can be used to analyze the data. This approach allows researchers to examine the main effects of discipline and gender on grades, as well as the interaction between them, while statistically controlling for parental income.

The relevant effects can be obtained with the following statistical models:

The relevant effects can be obtained with the following statistical models:

$$(Y_1 + Y_2) / 2 = b_0 + b_1X + b_2Z + e,$$
$$(Y_1 - Y_2) = b_0 + b_1X + b_2Z + e,$$

where Y_1, Y_2, X, and e have the same meaning as in Equations 3 and 4, and Z is parental income.

The coefficient b_1 in Equation 7 represents the main effect of gender, over and above the effect of parental income. The coefficient b_0 in Equation 8 represents the effect of discipline for students who have a score of 0 on both X and Z. Like before, this coefficient is in most cases rather meaningless if X and Z are uncentered. In order to be able to interpret b_0, it is usually necessary to center both the dichotomous X (by recoding it into −.5 and +.5, or into −1 and +1) and the continuous Z (by “mean centering” it, i.e., by computing the average value of all scores and then subtracting this value from every participant’s score). When both X and Z are centered, the coefficient b_0 in Equation 8 represents the main effect of test discipline for the participant with an average score on the covariate. If this coefficient is statistically significant, one would conclude that the students with an average parental income, regardless of their gender, have better grades in one of the two disciplines. The coefficient b_1 in Equation 8 is the interaction effect between discipline and gender when controlling for the effect of parental income. A significant b_1 in Equation 8 suggests that the gender difference is greater for one of the two disciplines.

Note that none of the data analysis programs, not even the menu-based ones, will automatically mean center the covariate for the user. It is thus important to manually mean center the covariate before including it in the analysis. A failure to do so leads to an incorrect interpretation of the main effect of the within-subjects variable in a mixed model ANCOVA (unless a score of 0 on the covariate is a theoretically meaningful value).

The mixed model ANOVA is a powerful analytic approach for examining data from complex research designs. It is useful to note that one of the most common 2 × 2 mixed model ANOVAs contains one manipulated within-subjects variable, and the between-subjects variable is the order in which the levels of the within-subjects variable were administered (subjects in one order condition first did Level 1 and then Level 2 of the within-subjects variable, whereas participants in the other order condition started out with Level 2 and then did Level 1). For more information on this model, refer to the entry on *Repeated Measures Designs* in this encyclopedia.

Another important issue to note is that researchers sometimes conduct a 2 × 2 mixed model ANOVA with pretest and posttest as the within-subjects variable. One should know that this is not always the best test. Statistical power can be increased by including the pretest as a covariate (i.e., by regressing the posttest on both the between-subjects variable and the pretest). However, this data analytic approach can be chosen only if certain conditions are satisfied (see the entry *Repeated Measures Designs* in this encyclopedia for more details).

Model Assumptions for Mixed Model ANOVAs

Mixed model ANOVAs must meet certain assumptions in order to generate unbiased estimates of the main and interaction effects. As with usual ANOVAs, one assumption is that the residuals in both the between-subjects model (Equation 3) and the within-subjects model (Equation 4) must be normally distributed. The second assumption is homogeneity of variances or homoscedasticity. This assumption holds that the two groups defined by the between-subjects variable have approximately the same error variance. Applying transformations to the data may correct violations of these assumptions.

Mixed model ANOVAs also have several assumptions that are specific to them. One of these is “homogeneity of the variance-covariance matrices.” This assumption is satisfied when the pattern of intercorrelations among the various levels of the within-subjects independent variable(s) is consistent across groups of subjects defined by the levels of the between-subjects independent variable(s). The homogeneity of the variance-covariance matrices assumption is tested using Box’s M statistic. If Box’s M returns a *p* value that is less than .001, then the variance-covariance assumption is violated. Violations of this assumption can be corrected for with data transformations.

The final model assumption of mixed model ANOVAs is “sphericity” and applies only to models including within-subjects variables with three or more levels. This assumption is satisfied if the variance of the difference scores for any two levels of the within-subjects independent variable is similar to the variance of the difference scores for any other two levels. Mauchly’s test of sphericity can be used to evaluate this assumption, and if it is significant at $p < .05$, the F and p values of the coefficients in the mixed model ANOVA should be adjusted using the Greenhouse-Geisser or the Huynh-Feldt corrections.

Extending Mixed Model ANOVAs to Linear Mixed Effects Models

An increasing number of researchers are analyzing data from studies with both within-and between-subjects independent variables as linear mixed effects models. In this approach, the unit of analysis is the observation rather than the subject. As a result, the data have to be in "long format" (one line per observation) rather than in "wide format" (one line per subject). Data files in the traditional wide format have to be restructured into long format before they can be submitted to a linear mixed effects model analysis.

When specified correctly and with complete data, the linear mixed effects model yields the same results (i.e., the same coefficients, the same F and p values) as the mixed model ANOVA. And yet, linear mixed effects models have numerous advantages. They are more flexible in that they allow researchers to analyze the effects of continuous within-and between-subjects variables. They have the ability to incorporate missing data directly (i.e., there is no need to delete incomplete cases or impute for missing values). They can account for multiple sources of nonindependence (e.g., when subjects react to the same set of items). Finally, they allow researchers to relax the previously mentioned sphericity assumption under certain circumstances.

Sohad Murrar and Markus Brauer

See also Analysis of Covariance; Analysis of Variance; Multivariate Analysis of Variance; Repeated Measures Analysis of Variance; Repeated Measures Designs; Two-Way Analysis of Variance

Further Readings

Judd, C. M., McClelland, G. H., & Ryan, S. (2009). Repeated-measures ANOVA: Models with nonindependent errors. In C. M. Judd, G. H. McClelland, & S. Ryan (Eds.), Data analysis: A model comparison approach (pp. 247–276). New York, NY: Routledge.

Keppel, G., & Zedeck, S. (1989). The mixed two factor design. In R. Atkinson, G. Lindzey, & R. Thompson (Eds.), Data analysis for research designs (pp. 293–314). New York, NY: W. H. Freeman.

Laerd Statistics. (2013). Mixed ANOVA using SPSS. Retrieved from https://statistics.laerd.com/spss-tutorials/mixed-anova-using-spss-statistics.php

Rutherford, A. (2011). ANOVA and ANCOVA: A GLM approach. Hoboken, NJ: Wiley.

MMLE

See Marginal Maximum Likelihood Estimation

MMPI

See Minnesota Multiphasic Personality Inventory

Brian C. Wesolowski Brian C. Wesolowski Wesolowski, Brian C.

Model–Data Fit

Model–data fit

1078

1081

Model–Data Fit

Empirical models play an important role in bringing order, comprehension, and manageability to complex interrelationships among variables. They enhance researchers’ abilities to develop hypotheses and provide mechanisms to speculate about multifaceted processes. In educational and related psychological research, empirical models are most often developed in order to explain latent constructs and are therefore considered to be only approximations of reality. Latent constructs are inferred based upon observable (i.e., measured) indicators or behaviors, each subject to errors in measurement. Goodness-of-fit measures, used in the context of latent construct modeling, describe how well the observed data represents the latent constructs of interest. Inferences related to latent constructs are drawn from these observable occurrences; therefore, assessing the goodness-of-fit for a model is one of the most important aspects to the validity of interpretation in model building processes.

Beginning in the 1980s, methodologies for fit evaluation were rapidly and exhaustively conducted in educational and psychometric research, resulting in a multitude of approaches. First, this entry describes applications of basic goodness-of-fit tests. Second, this entry broadly surveys some of the more commonly used examples of absolute model fit indices that answer the question, “*Does the hypothesized model provide an overall fit to the observed data?*” Third, this entry broadly surveys some of the more commonly used examples of comparative model fit indices that answer the question, “*Which model most adequately replicates under different sample selections?*” Lastly, this entry provides a brief overview of parsimonious fit indices.

Goodness-of-Fit Statistics

Goodness-of-Fit Statistics

Historically, early applications of fit evaluation included goodness-of-fit tests for observed categorical frequencies placed within a contingency table, where adequacy of fit was based upon the error variance of the model. However, when applied to linear measures or, more specifically, linear model building (i.e., factor analytic models including path analysis or structural equation modeling), error variance for collected observations is unknown. Therefore, traditional methods for fit evaluation are rendered not suitable.

In the context of linear model building, two of the more commonly used methods for evaluating overall model fit is the likelihood ratio chi-square goodness-of-fit statistic and the Pearson chi-square goodness-of-fit statistic. Use of these goodness-of-fit statistics came at the advent of the maximum likelihood estimation for the multinomial distribution. Using the maximum likelihood estimation procedure, the sampling distributions are based upon asymptotic distributions and use a vector of frequencies from nongrouped, observed data. It is important to note that a goodness-of-fit *statistic* (as compared to a goodness-of-fit *index* [GFI]) is a type of GFI with a known sampling distribution. Use of a goodness-of-fit *statistic* allows the researcher to conduct hypothesis testing for overall model fit. In particular, the chi-square goodness-of-fit statistic tests the hypothesis that the obtained population covariance input matrix of the observed data matches the model-implied covariance input matrix expected by the hypothesized model.

Traditionally, larger chi-square statistics, in relation to their degrees of freedom, indicate a lack of model-data fit. Smaller chi-square statistics, in relation to their degrees of freedom, indicate good model-data fit. In the application chi-square statistics to linear model building processes, researchers are not interested in rejecting the null hypothesis; rather, they are interested in accepting the null hypothesis where insignificant differences are desirable. In these instances, smaller chi-square values indicate good model-data fit. Therefore, a significant chi-square statistic suggests that the model does not fit the data. Conversely, an insignificant chi-square statistic indicates adequate model-data fit. Furthermore, *p* values attached to the chi-squares with adequate model-data fit would be expected to demonstrate nonsignificance.

However, several weaknesses have been documented regarding the traditional chi-square statistic for use as a qualifier of true model adequacy. These weaknesses include violations to the assumption of multivariate normality and sensitivity to sample size and strength of correlations. This oversensitivity to model discrimination can often result in considerable Type I errors. Consequentially, the researcher may choose to move to an ad hoc measure of fit where transformations to the asymptotic chi-square statistic can provide more robust management of the observed data. These alternative measures can include but are not limited to the scaled chi-square statistic, the adjusted chi-square statistic, or the WLSMV chi-square estimator, for example. Furthermore, the chi-square is a measure of *exact* fit, which contradicts the conceptual notion that model building processes are based upon *approximations* of reality. Therefore, retaining the null hypothesis is never to be expected. As a result, the acceptance of the null hypothesis is not generally of interest to the researcher. Properties of goodness-of-fit *indices* are therefore more relevant and meaningful in the context of linear model building.

Goodness-of-Fit Indices

Goodness-of-fit indices can be classified into three broad categories of practical fit indices: (1) absolute fit indices, (2) comparative fit indices, and (3) parsimonious fit indices.

Absolute Fit Indices

Absolute fit indices determine the degree to which the hypothesized model predicts, or fits, to the observed data. These indices do not use an a priori baseline model for comparison. Rather, they provide a measure derived from the model fit of the observed and hypothesized covariance matrices. Absolute fit indices answer the question, “*Does the hypothesized model provide an overall fit to the observed data?*” Absolute fit indices assess the overall model fit of the hypothesized model using statistical hypothesis tests represented by one single statistical index. Absolute fit statistics answer the question, “*Overall, how well could the hypothesized model reproduce the observed data?*” The measure itself evaluates the magnitude of the discrepancy between the sample and model-estimated covariance input matrices.

In evaluating overall model fit, rejection of a null hypothesis is not necessarily informative. What is more interesting to the researcher is the magnitude and location of the misfit. One method of evaluating misfit is through an analysis of residuals. One example of an absolute fit measure that provides an index of residuals is the GFI and the closely related adjusted GFI. The GFI calculates the proportion of variance accounted for in comparing how much better the hypothesized model fits compared to no model. The GFI is calculated using the sum of squared residuals and sum of squared variances. The adjusted GFI is an adjustment to the GFI that uses the model's degrees of freedom. Conceptually, the relationship between the GFI and the adjusted GFI is similar to the relationship between R^2 and adjusted R^2 in the context of an ordinary least squares regression, where the model is adjusted based upon the amount of predictors in the model.

Another example of an absolute fit index that provides an index of residuals is the root mean square residual (RMR). The RMR is calculated as the square root of the difference between the residuals of the obtained population covariance input matrix and the residuals of the model-implied covariance input matrix. However, the RMR is problematic, as the maximum value is unbound, resulting in difficulty of interpreting the acceptability of model-data fit. The RMR is also problematic, as the reported calculations are based upon the specific scale categories. As an example, if a measure does not provide similar categories for every item (i.e., a partial credit-type measure), the interpretation of results are unclear. Therefore, in these instances, the standardized RMR provides a more meaningful and substantive interpretation. However, the RMR and standardized RMR still do not provide specific information on where the misfit occurs, only a single index of residuals. Furthermore, both indices confound the error of sampling with the error of approximation.

The root mean square error of approximation, unlike the RMR and standardized RMR, simultaneously takes into account two potential sources of misfit: (1) error of approximation and (2) error of sampling. In doing so, the index is more robust to the centrality of the chi-square distribution and is independent of sample size. The error of approximation refers to the lack of fit of the hypothesized model to the population covariance matrix. The error of estimation refers to the closeness between the model-data fit of the sample and the model-data fit of the population. The parsing of both sources provides an index that simultaneously offers a measure of discrepancy between the obtained population covariance input matrix of the observed data and the model-implied covariance input matrix expected by the hypothesized model. The root mean square error of approximation index answers the question, *"How much is the error of approximation discrepant from the error of estimation due to sampling error?"*

Comparative Fit Indices

Comparative fit indices, also referred to in the research literature as incremental or relative fit indices, are a category of fit indices that compare the hypothesized model to some type of restricted, nested baseline (i.e., null) model. The null hypothesis and expectation of models for these indices are that the observed variables are uncorrelated, thereby not inferring evidence of a latent variable. In most cases, covariances between all input indicators are fixed to 0 in the baseline model. As a result of the overly severe constrainment, the baseline model is expected to demonstrate poor fit with large chi-square statistics. Comparative fit indices answer the question, “*Which model most adequately replicates under different sample selections?*”

One example of a comparative fit index (CFI) is the normed fit index (NFI) or the Bentler-Bonett NFI. The NFI compares the chi-square value (or fit function value) of the hypothesized model to the chi-square value (or fit function value) of the null model. A drawback of the NFI is its sensitivity to sample size. Small sample sizes often underestimate fit of the hypothesized model.

The Tucker Lewis index (TLI) overcame some of the limitations of the NFI. The TLI compares the mean square of the hypothesized model to the mean square of the null model. In some research literature, the TLI is referred to as the nonnormed fit index when discussed in the context of covariance structure analysis. The index can be represented as a proportion between the discrepancy between the hypothesized and null model. Limitations of the TLI/nonnormed fit index include a negative bias to smaller sample sizes, sensitivity to models more complex (i.e., more parameter estimates) in nature, and difficulty in interpreting the indices due to their nonnormed nature.

The incremental fit index (IFI), also referred to as DELTA2 in the research literature, was proposed as an improvement to the NFI. Specifically, the IFI adjusts for the NFI’s sensitivity to small sample sizes by accounting for the hypothesized model’s degrees of freedom. However, some drawbacks of the IFI include a positive bias to small sample sizes and a penalty for parsimony in the model due to the inclusion of the degrees of freedom for the hypothesized model.

One of the most reported comparative fit indices in research literature is the comparative fit index (CFI). The CFI was developed as an improvement to the NFI and IFI in that it is robust to small sample sizes. Conceptually similar to the

NFI and IFI in that it is robust to small sample sizes. Conceptually similar to the logic of the root mean square error of approximation, the CFI measures improvements in noncentrality by fixing the noncentrality parameter to 0. As a result, estimation procedures are not affected by the sample size.

Parsimonious Fit Indices

In the context of linear model building, parsimony refers to the least amount of estimated parameters needed to achieve an adequate level of model-data fit. Conceptually, adding parameters to the model will improve model fit; however, adding the additional parameters may not be justified or warranted from a model fit perspective. Parsimonious fit indices provide a measure of discrepancy between the sample and model-estimated covariance input matrices while taking into consideration the complexity (i.e., the number of estimated parameters) of the model. Model parsimony favors more simple (i.e., less estimated parameters) hypothesized models over more complex (i.e., more estimated parameters) hypothesized models. Parsimony-corrected fit indices compare overidentified models with restricted modes and make adjustments to many of the indices previously described as a way to penalize for complexity of the model. Fit indices become lower the more complex the hypothesized model is, and generally, parsimonious fit indices have lower values of adequate model fit than other fit indices. Parsimony-corrected fit indices are proportion-based indices that are broadly calculated as the ratio of the number of degrees of freedom used by the model and the total number of degrees of freedom. Parsimonious fit indices adjust for losses in degrees of freedom by comparing an overfit model (i.e., excessive coefficients) with a restricted model. Examples of parsimonious fit indices include the parsimony GFU, parsimony NFI, Type 2 parsimonious NFI, the parsimonious CFI, and the Akaike information criterion.

Brian C. Wesolowski

See also Chi-Square Test; Goodness-of-Fit Tests

Further Readings

Balakrishnan, N., Voinov, V., & Nikulin, M. S. (2013). Chi-squared goodness of fit tests with applications. Oxford, UK: Elsevier.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238–246.

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88(3), 588–606.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. Structural Equation Modeling, 6(1), 56–83.

Hu, L. T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural equation modeling. Concepts, issues, and applications (pp. 76–99). London, UK: SAGE.

Mulaik, S. A., James, L. R., Alstine, J. V., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105(3), 430–445.

Sivo, S. A, Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for “optimal” cutoff properties: Fit index criteria in structural equation modeling. The Journal of Experimental Education, 74, 267–288.

Yuan, K. H. (2005). Fit indices versus test statistics. Multivariate Behavioral Research, 40(1), 115–148.

Aarti Bellara Aarti Bellara Bellara, Aarti

Modified Angoff Method Modified angoff method

1081

1083

Modified Angoff Method

The term *standard setting* refers to the process of recommending or establishing cut scores on examinations. The cut scores are meant to differentiate examinees into two (e.g., pass/fail) or more groups (e.g., below basic, basic, proficient, and advanced). While the definition of standard setting may seem simple, the process is anything but. The process involves experts in the field, often referred to as subject matter experts (SMEs) who participate in a standard setting panel that judges each item on the examination to collectively agree upon/recommend a cut score. There are several different standard setting models, with the modified Angoff standard setting procedure being the most commonly used in contemporary licensure and certification settings. This procedure was first briefly introduced by William Angoff in 1971, as a small section within a lengthy chapter on scaling, norming, and equating, and since then, has been referred to as the Angoff method. However, the method has almost never been used how it was originally described; rather, variations or modifications of the original Angoff method are the most commonly used procedure for setting cut scores in licensure and certification settings, hence the term *modified Angoff method*. This entry reviews the procedures of the modified Angoff method, describes the concept of a minimally competent examinee, and reveals important points to consider when employing the modified Angoff method.

Procedures

Two critical components are needed to conduct a modified standard setting procedure: (1) a completed test and (2) a panel of SMEs to judge the test. There are no strict guidelines as to how many SMEs are needed, but a general rule of thumb suggests somewhere between 10 and 20 SMEs should participant in a standard setting panel where they make judgments on the individual items on the test. The definition of SME can be considered subjective. Each organization should define a set of parameters to guide the selection of SMEs. For example, criteria may include a minimum number of years in practice, recognition of accomplishment, a postgraduate degree, or a leadership position.

The standard setting panel focuses on the SME providing ratings for each item, keeping in mind a subpopulation of examinees when providing the ratings. The subpopulation or referent group of interest are borderline/just passing examinees, or as originally described by Angoff, “minimally acceptable person,” now referred to as “minimally competent examinee” when applying any version of the Angoff procedure. Angoff’s original suggestion was for the SMEs to rate each item on the test, keeping in mind the minimally acceptable person. An item would receive a 1 if the minimally acceptable person would answer the item correctly, and a 0 if the respondent would answer the item incorrectly. The sum could represent the raw score of the minimally acceptable person and be used to represent the lowest acceptable passing score on the exam.

The most common variation, or the modified version of this, is to have SMEs state the probability that a minimally competent person would get each item correct. To conceptualize a probability, SMEs may be asked to consider how many out of 100 minimally competent examinees would get the item correct, and report this as a *p* value, or proportion. For example, if an SME felt 70 minimally competent examinees would get the item correct, the item would receive a score, now known as an “Angoff rating” of .70. Providing guidelines such as only using ratings that are multiples of 5 or 10 can be helpful. Often, multiple rounds or iterations (no more than three) are conducted in order to calibrate the SME’s ratings. In between rounds, real data may be presented, such as item difficulties (*p* values) for an entire sample of examinees to help SMEs to understand how the item has functioned in the past. If we consider the SME who provided an Angoff rating of .70 on an item and then reveal that in a previous examination, 50% of *all* examinees got the item correct (yielding a *p* value of .50), then that SME may want to consider lowering his or her Angoff rating. Additionally, impact information could also be provided, such as the percentage of previous examinees who would be classified as failing or passing using the first iteration of ratings.

These discussions with real data are meant to help raters converge toward consensus in their ratings. Creating a response sheet for each SME to record the Angoff rating for each item and each round helps to keep the new data organized and expedite the discussions between rounds. Finally, a recommended passing score is derived by taking the average of either the rater or item means from the final round of ratings. Oftentimes, the cut score derived from a standard setting procedure needs to be formally approved by a governing board for the organization; therefore, the goal of the standard setting procedure may be to “recommend a cut score.”

Minimally Competent Examinee

The concept of a minimally competent examinee is crucial because this is the key referent group for the entire standard setting process. A large portion of the training during the standard setting panel focuses on the SME’s collective, conceptual understanding of the minimally competent examinee. For example, the minimally competent examinee may be described as someone who should pass the exam but may not be stellar. The nature of the test, and its level of stakes, may play an important role in the conceptualization of such a hypothetical examinee. The discussions and iterations often center on reexamining the organization’s collective understanding of this examinee. Those leading the standard setting panel should not proceed or rush through these conversations, because if there is not a shared understanding of the nature of this subpopulation, the variability in the Angoff ratings across raters may be very large.

Important Points to Consider

The aforementioned modified Angoff procedure is ideal when there is enough time allotted, there are not a large number of items, and normative data are available. Variations when all these factors are not available include calibration and convergence of a sample of items, for example, discussing every *k*th item to avoid multiple iterations when there are a large number of items and limited time, and using overall previous cut scores in place of item-level statistics for newly created items or items that do not have data.

Further variations of the Angoff method include the extended Angoff method and the yes/no method. The extended Angoff method is used for constructed response items, where SMEs rate how the minimally competent examinee would score on the rubric for the item(s). The yes/no method is practically identical to the original Angoff method, and its advantages include requiring less judgment from the raters, yielding less variability in the ratings, avoiding the process of from the raters, yielding less variability in the ratings, avoiding the process of having to arbitrarily estimate a proportion, and creating a faster process. The modified Angoff standard setting procedure can be easily adjusted for tests using multiple item formats by simply combining multiple procedures and coming up with an overall total score and averaging or weighting the scores appropriately. It can also be used to recommend more than one score (e.g., basic, proficient, and advanced) by identifying the borderline examinees at each of these levels.

Although there are several limitations to the modified Angoff standard setting procedure and its variations, it is not without limitations. Empirical studies comparing different standard setting procedures exist, but the large number of Angoff studies does not help to provide an accurate and fair comparison. The arbitrary nature of the selection of the SMEs along with the nebulous conceptualization of the minimally competent examinee have also garnered criticisms, yet not much empirical work has been done to actually refute the Angoff method.

Aarti Bellara

See also Cut Scores; Standard Setting

Further Readings

Cizek, G. J., & Bunch, M. B. (2007). Standard setting. Thousand Oaks, CA: SAGE.

Hambleton, R. K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. N. Hansche (Ed.), Handbook for the development of performance standards (pp. 87–114). Washington, DC: Council of Chief State School Officers.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 89–116). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational measurement (4th ed.), pp. 433–470).

Westport, CT: American Council on Education/Praeger.

Zhiyong Zhang Zhiyong Zhang Zhang, Zhiyong

Moments of a Distribution Moments of a distribution

1083

1085

Moments of a Distribution

Moments are quantitative measures of a distribution function. Formally, the *n*th moment about a value *c* of a distribution *f(x)* is defined as

$$\mu_n = E[(x-c)^n] = \left\{ \begin{array}{cc} \sum (x-c)^n f(x) & \text{Discrete distibution} \\ \int (x-c)^n f(x) dx & \text{Continuous distribution} \end{array} \right\}.$$

When *c* = 0, they are called the *raw moments*, and when *c* is set at the mean of the distributions, they are called *central moments*. The first raw moment is the mean and the first central moment is 0. For the second and higher moments, the central moments are often used. For some distributions, their moments can be flexibly obtained through their moment-generating functions. Certain distributions can be uniquely determined by a few moments. For example, a normal distribution can be determined by its first two moments. Although higher moments of a distribution can be available, the first four moments are of great interest to researchers. The remainder of this entry defines and describes those first four moments.

The first raw moment $\mu_1 = E(x) = \mu$ is the mean of a distribution and the first central moment is equal to zero. Mean is a popular measure of the central tendency of a distribution, especially for symmetric distributions.

The second central moment $\mu_2 = E[(x-\mu)^2] = \sigma^2$ is the variance of a distribution and is often denoted by σ^2. Variance is a frequently used measure of deviation from the central tendency.

The third central moment $\mu_3 = E[(x-\mu)^3]$ is related to the skewness (γ_1) of a distribution:

$$\gamma_1 = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}.$$

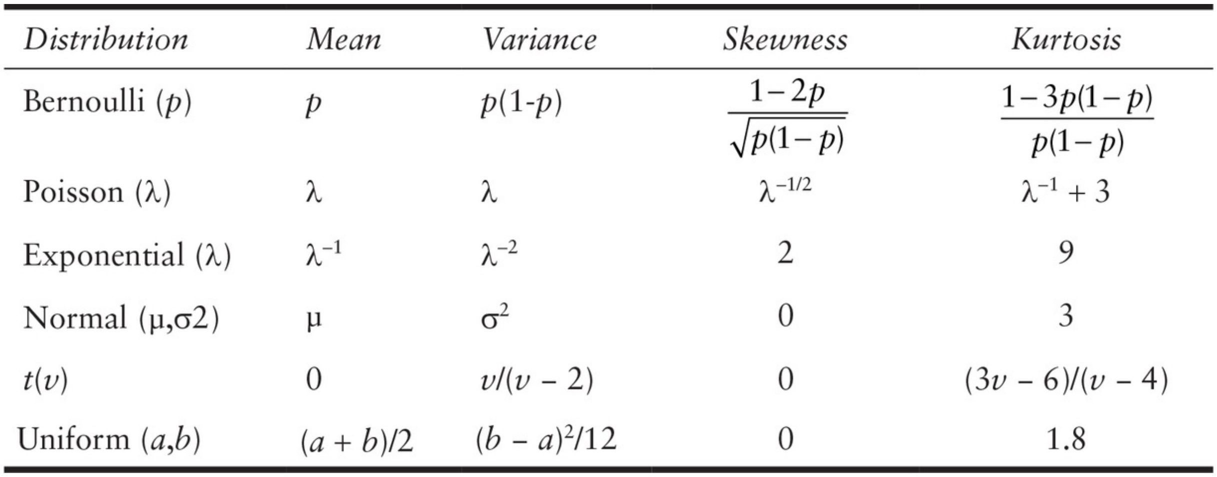
The skewness just defined is also called the third standardized moment and sometimes referred to as Pearson’s moment coefficient of skewness. Skewness measures the degree of asymmetry of a distribution. For symmetric distributions such as normal and Student’s *t* distributions, their skewness is 0. If the left tail of a distribution is longer than its right tail, the distribution has negative skew and the skewness is negative. If the right tail of a distribution is longer than its left tail, the distribution has positive skew and the skewness is greater than 0.

The fourth central moment $\mu_4 = E[(x-\mu)^4]$ is related to the kurtosis (γ_2) of a distribution:

$$\gamma_2 = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4},$$

which is also called the fourth standardized moment. Kurtosis is associated with the tail, shoulder, and peakedness of a distribution. Generally, kurtosis increases with peakedness and decreases with flatness, while many have argued that kurtosis has as much to do with the shoulder and tails of a distribution as it does with the peakedness. The kurtosis of a normal distribution is 3. Distributions with a kurtosis less than 3 are said to be platykurtic, whereas distributions with a kurtosis greater than 3 are said to be leptokurtic. Skewness and kurtosis are often used in testing the normality of a distribution.

Table 1 summarizes the first four moments for commonly used distributions.



Distribution	*Mean*	*Variance*	*Skewness*	*Kurtosis*
Bernoulli (p)	p	$p(1\text{-}p)$	$\frac{1-2p}{\sqrt{p(1-p)}}$	$\frac{1-3p(1-p)}{p(1-p)}$
Poisson (λ)	λ	λ	$\lambda^{-1/2}$	$\lambda^{-1}+3$
Exponential (λ)	λ^{-1}	λ^{-2}	2	9
Normal (μ,σ2)	μ	σ^2	0	3
$t(v)$	0	$v/(v-2)$	0	$(3v-6)/(v-4)$
Uniform (a,b)	$(a+b)/2$	$(b-a)^2/12$	0	1.8

Zhiyong Zhang

See also Distributions; Kurtosis; Skewness; Variance

Further Readings

Casella, G., & Berger, R. L. (2002). Statistical inference (2nd ed.). Pacific Grove, CA: Duxbury.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. Psychological Methods, 2(3), 292–307.

Groeneveld, R. A., & Meeden, G. (1984). Measuring skewness and kurtosis. The Statistician, 33(4), 391–399.

Kayla Tureson Kayla Tureson Tureson, Kayla

Anthony Odland Anthony Odland Odland, Anthony

Monte Carlo Simulation Studies Monte carlo simulation studies

1085

1089

Monte Carlo Simulation Studies

Monte Carlo simulation refers to a broad range of methods of evaluating statistical estimators through the use of computer algorithms. Monte Carlo methodology was developed by American physicist Stanislaw Ulam, who first conceptualized the method while attempting to determine the probability of winning a game of solitaire; he found that playing a number of games and determining the percentage of winning games was much simpler than attempting to calculate all possible card combinations. In the 1940s, Ulam and John von Neumann employed this method of developing the hydrogen bomb. The simulation is named after the famous Monte Carlo Casino in Monaco because the method is based on random chance.

The methodology generates a large number of occurrences based on a set of specified parameters, which can be used to estimate a given population. This method often utilizes randomly generated numbers to obtain a range of possible outcomes, the parameters of which are extrapolated from known populations or theory. The likelihood of the occurrence of a particular outcome can be determined by dividing the frequency that the outcome occurred by the total number of trials. As the number of trials increases, the accuracy of determining the likelihood of the particular outcome increases. Artificially generated data enable a researcher to evaluate a simulation that resembles the desired population and apply it in myriad ways, such as conducting hundreds or even hundreds of thousands of trials for a given pseudo-population. Given the broad-reaching utility of this methodology, it is employed in a variety of fields and subject matters, such as physics, engineering, biology, mathematics, finance, as well as behavioral sciences. This entry explores the methods and basic procedures of the Monte Carlo simulation, its application in social and procedures of the Monte Carlo simulation, its application in social and behavioral research, as well as its limitations.

Methods and Basic Procedures

The most commonly employed application of the Monte Carlo simulation is the examination of sampling distributions, although its application extends to a variety of scenarios for which a complete mathematical analysis is otherwise not feasible or is extremely difficult. Discussion of the enumerable number of applications is beyond the scope of the current entry.

Fundamentally, the methodology allows one to examine a very large number of observations that are created from a set of parameters. In part, the utility of the Monte Carlo methodology includes flexibility and the ability to generate a large number of observations based on existing parameters (e.g., summary data for a population of interest). The basic procedure of the Monte Carlo simulation is to first specify the pseudo-population through the development of a computer algorithm to generate data for the desired statistic. This computer algorithm generates artificial data to simulate a population. These data can then be used by the researcher to study and to better understand the behavior of the statistical estimates from the data. A commonly used algorithm is known as “middle-square digits.” For this algorithm, an arbitrary *n*-unit integer is squared, creating a 2*n*-digit product. A new integer is then created by removing the middle *n*-digits from the product, and then the process is repeated over and over, creating a long chain of integers that will eventually repeat itself. Observations are usually random or pseudorandom and are intended to generalize the population of interest. The extent to which the initial parameters are representative of the population of interest, the generated pseudo-population resembles a real-world population in all relevant aspects. The pseudo-population generally encompasses a very large number of observations, allowing for it to be analyzed with ancillary statistical techniques, which can be useful in better characterizing the actual population and generalizing appropriate inferences.

Applications of Monte Carlo Simulation

Monte Carlo methods have a variety of applications in social and behavioral research. In order to make standard parametric inferences, a high level of statistical theory about an estimator is required. The typical approach to inference requires an analytic proof regarding its sampling distribution and conditions of the data as well as formulas to estimate the parameters of the distribution of sample data. However, there are situations in which such an approach is not possible. Properties of estimators are uncertain when the conditions of a proof are not fully developed. Additionally, a method that lacks a well-developed statistical theory is not reliable for the typical approach to inference. Monte Carlo simulation can be particularly useful in situations in which there is no strong theory regarding the statistic of interest or the conditions needed for the statistical theory do not hold. If the components of a statistic are known or assumptions about its components can be made, then the components can be simulated using a pseudo-population.

A typical hypothesis test assesses the probability of a specific value of a statistic and assumes that the population value of the parameter has some null value. This can also be performed with a Monte Carlo simulation by setting the pseudo-population value of the parameter at hand to the null value and then calculating the percentage of trials that the parameter is above or below the value of the estimate observed in the real data. This procedure requires the assumption that everything else apart from the null parameter in the pseudo-population is exactly the same as it is in the true population.

Another way in which a Monte Carlo simulation can be utilized is testing a null hypothesis under various plausible conditions. If it is not possible to be certain that a specific distribution generated a variable, one can use a range of plausible distributions for the variable in question. The range of distributions can then be run in a series of experiments using pseudo-populations to evaluate the null hypothesis for each experiment. The results of these experiments can then be used to make inferences about the likelihood of the null hypotheses holding true in the population.

Occasionally a researcher may have a choice between two or more estimators of the same parameter in the population, for example, deciding between several measures of central tendency. By utilizing the Monte Carlo simulation, a researcher would be able to evaluate the different estimators with the same criteria in simulated situations with identical conditions. This simulation is conducted by first determining the desired criteria for comparison, then defining the pseudo-population that most closely approximates the population of interest. Next, the desired estimators for comparison are applied to the pseudo-population and a series of trials are conducted. After the trials are conducted, the estimators are compared with each other.

are compared with each other.

Statistical inference involves two types of error: Type I error, incorrectly rejecting a true null hypothesis, and Type II error, failing to reject a false null hypothesis. Since inference tests are conducted because the value of the parameter is not known, inferential errors can never be identified by solely looking at real-world data. Monte Carlo simulation provides a means of determining inferential errors because it has information regarding the value of the parameters of the population. Inferential tests should first be conducted and the findings from the actual data should be reported. Following the preliminary inferential tests, simulation experiments would be conducted on the estimated Type I and Type II error rates. Reporting real data results along with the simulated experimental results provide a larger scope of information from which to determine inferential error rates. An example of utilizing the Monte Carlo simulation in this manner is estimating the Type I error rate (i.e., incorrectly rejecting the null hypothesis).

A clinical example of Type I error would be a clinician interpreting a low score as evidence of a brain injury or other neurological disorder in a neurologically intact individual. The Monte Carlo method can be utilized to estimate the statistical likelihood of Type I error, which is partially determined by the base rate of normative individuals obtaining an atypical test performance. These base rates can be most accurately acquired through the use of conormed measures. Because conormed measures are rarely accomplished in the testing itself, statistical methods such as Monte Carlo can be used to estimate base rates for flexible test batteries.

The binomial model is another proposed statistical approach to estimate base rates, but the Monte Carlo method has several advantages over this model. In a binomial model, the probability of an event, usually described as success, is estimated given its probability and the number of trials in which the event can occur. The binomial model views each test as a trial in a series of trials, this series representing the test battery in which a person obtains a classification of either impairment or no impairment, to estimate the likelihood of obtaining an impaired test result for a number of confidence levels and number of tests administered. Unlike the Monte Carlo method, the binomial model does not take correlations among measures in a battery comprised of multiple tests (assumes independence of variables, which is rare in social sciences). The binomial model only takes the number of tests and probability of a low score into account, which does not change across age-groups. In contrast, Monte Carlo estimates as well as actual estimates do change across age-groups. Both binomial and Monte Carlo models are fairly accurate in estimating low-probability events, but in terms of estimating the frequency of test scores across the entire range of frequencies, the Monte Carlo method is more accurate. Although using a statistical method to control for variability and measurement error is not sufficient for use in identifying impairment in an individual, it may help improve specificity and sensitivity of test results by helping to rule out potentially misleading test results that were derived from random error or other types of variation.

Monte Carlo simulation can be useful in assessing the robustness of parametric inference to violations underlying assumptions. Parametric inference may be evaluated by testing a variety of conditions and variables such as error rate within the simulation framework. An example would include contrasting results for a variety of conditions (e.g., number of observations and distribution types) across appropriate samples. This can be a particularly useful application of the Monte Carlo simulation method, as assumption violations of statistical tests are a frequent occurrence.

John Crawford, Paul Garthwaite, and Catherine Gault (2007) utilized the Monte Carlo simulation method to estimate what percentage of the healthy population would be expected to exhibit one or more abnormally low test scores in a battery of neuropsychological test measures. If abnormally low scores on a given test are defined as scores that fall below the fifth percentile, then by definition, 5% of the population would then be expected to have a score that is lower. Because multiple tests are used in a neuropsychological assessment, several scores are taken into account for a patient’s profile, and it is not uncommon for someone to achieve at least one abnormally low score. That is to say, reference to the normal distribution assumes that one score is being interpreted, which is rarely the case in psychological or neuropsychological assessment.

This presents an important question as to what percentage of the healthy population would be expected to exhibit at least one abnormally low test score. The Monte Carlo method was used to estimate what percentage of the population would be expected to exhibit *j* or more abnormally low index scores on the Wechsler Adult Intelligence Scale–Third Edition as well as the Wechsler Intelligence Scale for Children–Fourth Edition. The simulation utilizes a *k* by *k* matrix between *k* components of the test battery. For this particular example, the *k* components would be the indices on the Wechsler Adult Intelligence Scale–Third Edition or Wechsler Intelligence Scale for Children–Fourth Edition, and the matrix is comprised of interscale correlations between each test scale. Once the matrix has been established, the Choleski decomposition of *R* should be obtained. The Choleski decomposition creates a lower triangular matrix by taking the square root of the correlation matrix *R*. Then, a random vector of *k* independent standard normal pseudorandom variates is generated. Vectors are multiplied a large number of times by the lower Choleski decomposition matrix.

In this particular study, this process was repeated one million times and the number of abnormal test scores obtained on each Monte Carlo trial was tabulated to create base rates. It was estimated that about 13% of the population is expected to exhibit one or more abnormally low scores on the Wechsler Adult Intelligence Scale–Third Edition, when abnormality for an individual scale is defined as 5%. For the Wechsler Intelligence Scale for Children–Fourth Edition, about 17% of the population is expected to exhibit one or more abnormally low scores. The percentage of the population expected to exhibit *j* or more abnormally low test scores increases considerably as the number of tests in the battery increases and the definition of significance is loosened. Although one index score may be classified as abnormally low, it is certainly not unusual for a member of the general population to obtain an abnormally low index score. Failure to consider the prevalence of low scores across a battery of tests could potentially lead to erroneous inferences of cognitive impairment. The Monte Carlo method could be incorporated into clinical practice by utilizing test scoring software or other computer software to help account for variation in test results across a multitest battery. Clinical interpretation of scores across the entire test battery should include reference of the prevalence of low scores in a normal population.

In addition to aiding interpretation of an individual profile as shown in the previous example, the Monte Carlo method can be applied to group-based research with specific patient populations (e.g., HIV, multiple sclerosis, diabetes) to estimate the prevalence of neuropsychological deficits within the specified population. The prevalence of deficits in these populations is estimated by calculating the percentage of cases that meet a predefined criterion of impairment. The typical criteria for such studies are that a patient should exhibit *j* or more abnormally low test scores on a given battery, for example, exhibiting at least three test scores that are 1 standard deviation below the mean. This particular application of the Monte Carlo method to estimate the prevalence of abnormal test scores in a population can be useful in reducing the risk of overinferring the presence of impairment.

Limitations and Considerations

The Monte Carlo method should be used with caution when the population means, standard deviations, and correlations have been estimated using a modestly sized standardization sample, generally when sample *n* is 300 or less. When examining test score differences using a modestly sized sample, there will be an overall trend of overestimation of the level of abnormality for each separate comparison. Any inferences drawn from a modest sample size should take into account the likelihood of overestimation or, in some cases, underestimation. Confidence in the Monte Carlo simulation program should be limited to studies of people with similar demographics to the mean of the normative sample. If certain factors known to moderate test performance, such as education, intelligence, or ethnicity, are not considered within the intercorrelation matrices or cut scores for individual scales, then Monte Carlo estimations will be less accurate for subgroups stratified based on those variables. Because clinical populations may be more likely to produce nonnormal score distributions, using the Monte Carlo method with clinical populations may produce less accurate estimates, if the parameters of the distribution are not fully understood and taken into consideration for a given simulation. Conversely, even though a construct is not normally distributed, it certainly does not mean that it cannot be simulated using the Monte Carlo method. When applied to a clinical population across a large sample size, Monte Carlo simulation has resulted in test performance base rates with 97% accuracy.

In addition to issues regarding the size and representativeness of the normative sample, insufficient attention may have been paid to normalizing test scores. Assumptions about multivariate normality include the assumption that scores are continuous. Consequently, the accuracy of the estimates will decrease if the tests in a battery have a limited number of possible raw or scaled scores, such as with Wechsler subtest scaled scores. When certain estimates of the population correlations are not available, such as a test battery comprised of measures from diverse sources, the Monte Carlo method can be used in a more exploratory manner to examine different assumptions for the population correlations. Any inferences drawn from exploratory applications should be carefully considered.

Kayla Tureson and Anthony Odland

See also Bootstrapping; Parameter Invariance; Parameter Mean Squared Error; Parameter Random Error; Sample Size; Stratified Random Sampling; Type I

Error; Type II Error

Further Readings

Brooks, B. L., & Iverson, G. L. (2010). Comparing actual to estimated base rates of “abnormal” scores on neuropsychological test batteries: Implications for interpretation. Archives of Clinical Neuropsychology, 25, 14–21. doi:10.1093/arclin/acp100

Brooks, B. L., Strauss, E., Sherman, E. M. S., & Iverson, G. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. Canadian Psychology, 50(3), 196–209. doi:10.1037/a0016066

Carrasco, R. M., Grups, J., Evans, B., Simco, E., & Mittenberg, W. (2013). Apparently abnormal Wechsler memory scale index score patterns in the normal population. Applied Neuropsychology: Adult, 22(1), 1–6. doi:10.1080/23279095.2013.816702

Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. Neuropsychology, 21(4), 419–430. doi:10.1037/0894-4105.21.4.419

Decker, S. L., Schneider, W. J., & Hale, J. B. (2012). Estimating base rates of impairment in neuropsychological test batteries: A comparison of quantitative methods. Archives of Clinical Neuropsychology, 27, 69–84. doi:10.1093/arclin/acr088

Eckhart, R. (1987). Stan Ulam, John Von Neumann, and the Monte Carlo method. Los Alamos Science [Special issue]. Los Alamos, NM: Los Alamos National Laboratory.

Generating individual samples from a pseudo-population. (1997). In Christopher

Z. Mooney (Ed.), Monte Carlo simulation (pp. 6–50). Thousand Oaks, CA: SAGE.

Gentle, J. E. (2003). Random number generation and Monte Carlo methods. New York, NY: Springer.

Lemieux, C. (2009). Monte Carlo and quasi-Monte Carlo sampling. New York, NY: Springer Science + Business Media.

Odland, A. P., Lammy, A. B., Martin, P. K., Grote, C. L., & Mittenberg, W. (2015). Advanced administration and interpretation of multiple validity tests. Psychological Injury and Law, 7(4).

Using Monte Carlo simulation in the social sciences. (1997). In Christopher Z. Mooney (Ed.), Monte Carlo simulation (pp. 66–93). Thousand Oaks, CA: SAGE.

Angela K. Murray Angela K. Murray Murray, Angela K.

Montessori Schools Montessori schools

1089

1092

Montessori Schools

Maria Montessori (1870–1952) developed an educational approach that emphasized treating each child as an individual with an innate ability to develop her own physical, social, emotional, and cognitive potential given an appropriate environment. More than 100 years later, an estimated 20,000 Montessori schools exist worldwide with over 4,000 in the United States including more than 500 in the public sector. The majority of Montessori schools serve children aged 3–6 years, but a large number of elementary programs exist with growing availability of options for younger and older students as well. Although the fundamentals of educational research apply to Montessori schools, unique aspects require researchers to address considerations in the areas of fidelity of implementation, types of research questions, and research design.

Fidelity of Implementation

The term *Montessori* is not legally protected, so any school can use the term in its name regardless of the degree to which it follows Montessori principles. Because this results in a great deal of variety in schools that label themselves Montessori, researchers wishing to draw conclusions about Montessori schools must establish the authenticity of the environments they study. National Montessori organizations agree on key elements that describe quality Montessori environments.

- Whole child-focused approach, allowing children to develop naturally and independently at their own pace and following their own individual interests.
- Montessori-trained teachers guiding children in learning environments fully equipped with specially designed hands-on materials for each age level.

Long periods of time (ideally 3 hours) allowing children freedom to work without interruption.

Limited whole-group instruction with focus on one-on-one instruction for early childhood and small groups for older students.

Multiage classrooms generally including 3-year age ranges (under 3, 3–6 years, 6–9 years, 9–12 years, 12–15 years, and 15–18 years).

Emphasis on formative assessment through teacher observation and detailed record keeping rather than traditional grading systems.

Although no one accepted measure exists to validate the quality of a Montessori environment, a number of researchers have created instruments designed to do so. Researchers can choose to use one of the observation tools or survey instruments developed by other investigators or create their own instruments based on key elements of Montessori education. Only after demonstrating fidelity of implementation can research questions about Montessori schools in general be addressed.

Research Questions

Academic Outcomes

Academic outcomes are the most obvious research questions for any educational approach, but accurately assessing academic progress in Montessori schools presents potential challenges. First, Montessori children often have less test-taking experience than other students because traditional tests are not a primary focus. Second, Montessori teachers introduce content at different times for individual students, depending on their interests and readiness, so students, may not cover material in the same order or at the same chronological age as their classmates or their traditionally educated counterparts. Even so, public Montessori schools may provide test-taking practice and require that students follow the state standard grade-level timetable for content coverage more closely and to maximize their performance on state assessments. Examples of academic achievement measures that have been used in Montessori school research include state assessments, Woodcock-Johnson Tests of Achievement, and TerraNova.

Nonacademic Outcomes

Although academic achievement is important and past studies suggest Montessori students tend to perform at least as well as and often better than students from traditional environments, Montessori education focuses on educating the whole child, which opens doors for exploring other areas of potential impact. Sometimes called *soft skills*, *socio-emotional learning*, or *executive functions*, nonacademic outcomes cover such things as self-regulation, persistence, focus, problem-solving skills, creativity, and intrinsic motivation, among others.

Although important to assess in Montessori environments, nonacademic outcomes are much more difficult to measure than academic achievement. Fortunately, a number of tools have been used to measure nonacademic dimensions in both Montessori and non-Montessori environments. The Head-Toes-Knees-Shoulders task, delay of gratification tests, and Dimensional Change Card Sort assess executive functions. The Experience Sampling Method gauges student affect during the course of the day. False belief tasks investigate theory of mind from the field of developmental psychology. Finally, social problem-solving tasks use reactions to stories to assess social competence.

Classroom Practices

Research questions can involve individual classroom practices rather than academic and nonacademic outcomes resulting from the implementation of a cohesive educational model. Montessori schools incorporate many techniques that are used in non-Montessori environments as well, so they provide settings for examining these practices. Examples include differentiated instruction, self-directed learning, mixed-age groupings, kinesthetic learning, nature-based learning, service learning, authentic assessment, peer tutoring, and many others. When these techniques are evaluated in non-Montessori settings, Montessori schools benefit from a much wider body of knowledge.

Research Design

Research design is driven by research questions, and the designs used in other educational environments apply to Montessori schools with a few additional considerations. The research design that provides the strongest evidence of effectiveness for any type of school is an experiment that involves random selection and random assignment. Of course, this ideal is challenging in practice when dealing with the education of real children, so researchers must find creative solutions. For example, one study evaluated the impact of Montessori education using test and control groups based on students' admission to a public Montessori program through a lottery system. The randomization in the lottery selection process served as a mechanism for randomly assigning students to either a Montessori school or some other type of educational environment. Because experiments are difficult to execute, other types of research such as quasi-experimental studies, survey research, qualitative research, and action research are more common in Montessori schools.

Quasi-Experimental Studies

Studies employing strategies for enhancing internal validity when experimental designs are impractical are referred to as quasi-experimental studies. Using naturally occurring groupings to compare students already enrolled in either Montessori or non-Montessori schools exposes research to questions about possible alternative explanations for results. To mitigate these potential confounds, researchers employ strategies such as demographically matching students in Montessori and non-Montessori schools or controlling for demographic differences using statistical procedures.

One study comparing Montessori preschools to conventional preschools used intact groups but asked parents to identify the school their child would have attended if not the Montessori school. The two most popular private preschools mentioned became the conventional comparison group for evaluating differences in school readiness skills while controlling for fall skill levels statistically. Such techniques help address concerns about the confounding effects of groups being inherently different in ways other than type of education, but they cannot eliminate them completely. The primary criticism that cannot be accounted for by these strategies is the likelihood that parents who ultimately choose Montessori schools for their children are inherently different from parents who choose other programs.

Survey Research

Surveys represent another popular research design used to gather information about Montessori education. Students are sometimes survey participants as are teachers and parents. Topics addressed by surveys of students are usually self-reported reactions to various aspects of the classroom experience. Teachers are often asked about their beliefs or classroom practices while parents’ attitudes about Montessori education or their parenting practices are of interest. Researchers sometimes use teacher and parent report surveys and checklists to evaluate such things as student behavior, social skills, or executive functions. Because one could argue that teachers and parents drawn to Montessori education may have very different perspectives on children, the appropriateness of using such tools to compare students from different educational environments may be called into question.

Qualitative Research

A great deal of the research that is done in Montessori schools is qualitative in nature, which means that it involves the collection, analysis, and interpretation of narrative and visual (nonnumerical) data through methods such as case study research, in-depth interviews, or focus groups. Instead of random sampling, large sample sizes, random assignment, and established instruments, credibility in qualitative research relies on extensive field notes and a paper trail authenticating findings. Methods used include in-depth dialogue and interaction with the participating teachers and students, multiple sources of data, member checks, and triangulation to ensure that the story told by the researcher matches the story perceived by the participants.

Because Montessori is an alternative educational approach with specialized schools and a limited number of teacher training centers, qualitative Montessori researchers often have strong personal ties to their research sites. These ties require researchers to acknowledge how their personal theories, preconceptions, or values may influence the conduct and conclusions of the study as well as the possible influence of the researcher on participants. Personal perspectives and the potential impact of the researcher in the school are not necessarily negative and, in fact, could be productive in qualitative research if sufficiently examined and explicated.

Action Research

Action research, sometimes called teacher research, is used to inform day-to-day classroom practice with educators gathering information in a structured manner to improve instruction and learning within their own classrooms or schools.

Information sources can include classroom observation, interviews or recorded conversations, questionnaires and attitude scales, and other naturally occurring data. The primary difference between action research and other types of research is that the goal of action research is to generate knowledge that is specifically relevant to the local setting, while the goal of traditional research is to produce knowledge that can be generalized to the field.

In the end, the quality of action research is judged by the extent to which it provides credible data used to successfully change practice. Action research aligns particularly well with Montessori education because it reflects the early approach Maria Montessori took in developing her method with underprivileged children in Rome, where she carefully observed students' responses to the various materials and practices she implemented. Because 21st-century Montessori teacher training often incorporates action research, examples of student-completed action research projects are available in the online repositories of universities offering Montessori teacher education programs.

Angela K. Murray

See also Active Learning; Experimental Designs; Kinesthetic Learning; Qualitative Research Methods; Quasi-Experimental Designs; Self-Directed Learning; Self-Regulation; Service-Learning; Socio-Emotional Learning

Further Readings

Besançon, M., & Lubart, T. (2008). Differences in the development of creative competencies in children schooled in diverse learning environments. Learning and Individual Differences, 18(4), 381–389.

Dohrmann, K. R., Nishida, T. K., Gartner, A., Lipsky, D. K., & Grimm, K. (2007). High school outcomes for students in a public Montessori program. Journal of Research in Childhood Education, 22(2), 205–217.

Laski, E. V., Jor'dan, J. R., Daoust, C., & Murray, A. K. (2015). What makes mathematics manipulatives effective? Lessons from cognitive science and Montessori education. SAGE Open, 5(2).

Lillard, A. (2012). Preschool children’s development in classic Montessori, supplemented Montessori, and conventional programs. Journal of School Psychology, 50(3), 379–401.

Lillard, A., & Else-Quest, N. (2006). Evaluating Montessori education. Science, 313, 1893–1894.

Rathunde, K., & Csikszentmihalyi, M. (2005). Middle school students’ motivation and quality of experience: A comparison of Montessori and traditional school environments. American Journal of Education, 111(3), 341–371.

Rathunde, K., & Csikszentmihalyi, M. (2005). The social context of middle school: Teachers, friends, and activities in Montessori and traditional school environments. The Elementary School Journal, 106(1), 59–79.

Gail Tiemann Gail Tiemann Tiemann, Gail

Mood Board Mood board

1092

1093

Mood Board

A mood board is a collage of images chosen to express ideas, sentiment, or emotion related to concepts or ideas. Mood boards can be created by researchers or by research participants by using art, photos, descriptive words, or other graphical elements from sources such as magazines, newspapers, or websites. The boards themselves can be physical or digital.

The underlying theory behind the mood board is that of stimulus and response. The researcher or research participant may select images that induce a complex emotional response to a certain stimulus such as a design, idea, or issue. The response to the mood board is then communicated to the researcher via rich discussion. Creating or responding to mood boards can be especially helpful for individuals who have difficulty with expressing emotions via words.

Mood boards are used in many areas of qualitative study including product design, industrial design, and market research. In product design, researchers can use mood boards to examine how potential consumers feel about certain product features. The mood board can be used to spur discussion. Information gleaned can be collected by the researcher.

Another potential application of mood boards is in focus group research. For example, in a small focus group, participants could be asked to construct a mood board around what educators would see as an ideal school environment. After creating boards, the discussion could revolve around why educators were motivated to select particular images. Analysis and interpretation of the conversation could then commence using other standard qualitative methods.

Gail Tiemann

See also Design-Based Research; Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Lucero, A. (2012, June). Framing, aligning, paradoxing, abstracting, and directing: How design mood boards work. In Proceedings of the Designing Interactive Systems Conference (pp. 438–447). ACM.

McDonagh, D., & Denton, H. (2005). Exploring the degree to which individual students share a common perception of specific mood boards: Observations relating to teaching, learning and team-based design. Design Studies, 26(1), 35–53.

McDonagh, D., & Storer, I. (2004). Mood boards as a design catalyst and resource: Researching an under-researched area. The Design Journal, 7(3), 16–31.

Meghan Ecker-Lyster Meghan Ecker-Lyster Ecker-Lyster, Meghan

Mortality

Mortality

1093

1094

Mortality

In educational research, mortality (also referred to as experimental mortality or attrition) is a metaphorical term that is used to describe the loss of participants from a study prior to completion. Mortality is among one of eight common threats to internal validity. Threats to validity can be troublesome for research, as these threats limit the conclusions that can be drawn from a study.

Threats to internal validity inhibit researchers’ confidence in reporting that a relationship exists between an independent variable and a dependent variable. To make valid conclusions about the results obtained from a research study, there must be sufficient evidence to substantiate the claim. Mortality threatens this assumption because it compromises the quality and quantity of data garnered from a study.

Mortality is particularly problematic for longitudinal research, as there is an increased potential for reasons a participant may drop out prior to completion (e.g., geographic move, apathy, changes in availability). Studies that employ rigorous or demanding conditions are more susceptible to mortality. For example, studies that require extensive time commitments, are physically or psychologically demanding, or place other stressors on participants may be more likely to experience higher rates of mortality than studies with less demanding conditions.

It is reasonable to assume that mortality is likely to occur across both experimental and nonexperimental research to some degree. Mortality rates become a concern and a threat to internal validity when mortality rates are significantly different between the study’s groups. However, mortality is exclusively a problem not only when differential loss occurs within a study but also when substantially high rates of dropout occur across all study participants. When either of these issues occurs within a study, the results can be dramatically impacted, making it more difficult to conclude that the outcomes obtained were the result of the treatment condition rather than mortality rates.

The underlying problem with differential loss and high mortality rates within a study is that participants who drop out of a study prior to completion, for whatever reason, are characteristically different from participants who complete the study. Differential loss and high mortality rates within a study can lead to relevant biases between groups that may inflate, obscure, or confuse the effects of interest being studied. Additionally, in experimental research, when mortality is systematically related to the study's design (e.g., treatment conditions are too demanding), it is unclear whether unintentional outcomes were produced by the research design rather than the manipulation of the independent variables.

Although there is no panacea for completely eliminating mortality, the best approach for dealing with this threat to internal validity is to employ randomization whenever possible. Using random assignment presumes that participants who are susceptible to dropout will be equally distributed across both groups.

Meghan Ecker-Lyster

See also Generalizability; Internal Validity; Random Assignment; Validity; Validity Generalization

Further Readings

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). Experimental and quais-experimental designs for generalized causal inference (2nd ed.). Boston, MA: Cengage Learning.

John Mark Froiland John Mark Froiland Froiland, John Mark

Motivation

Motivation

1094

1095

Motivation

Motivation is a crucial concept in education but is often either overlooked or misunderstood by educators. There are multiple types of motivation to consider. Also, there are many ways that teachers and parents can promote the development of students’ motivation to learn. For instance, parental autonomy support and teacher autonomy support can both help students develop a long-term love for learning, whereas controlling styles of communication can decrease intrinsic motivation and increase academic anxiety. This entry discusses why motivation is important in education and long-term learning and looks at the key elements of motivation.

Motivation is a significant predictor of long-term achievement, above and beyond many other important factors such as intelligence, gender, ethnicity, and parent education. Part of the way motivation affects achievement is via classroom engagement. Namely, students who are highly intrinsically motivated to learn often pay attention in class, raise their hands, and otherwise display studious behavior at school. Such students are also more likely than other students to read for pleasure outside school and engage with learning opportunities in multiple aspects of life, such as at museums and libraries. Another key aspect of motivation involves expectations. Students who expect to do well in school and complete college are more likely to work hard in school and achieve more than students who do not expect to accomplish as much academically.

Key Aspects of Motivation in Education

The average students lose intrinsic motivation to learn each year as they progress from kindergarten to high school. Intrinsic motivation involves seeing the beauty and purpose in learning or otherwise finding learning enjoyable. By contrast, Mark Lepper and colleagues discovered that the average student maintains similar levels of extrinsic motivation throughout the primary grades. Extrinsic motivation involves a desire for grades, money, or tangible rewards related to academic performance or behavioral compliance with teacher requests.

As John Mark Froiland and Emily Oros found, both intrinsic motivation to learn and extrinsic motivation for grades predict the development of achievement across the United States. However, intrinsic motivation is positively associated with numerous indicators of psychological well-being, such as happiness, low levels of anxiety, and sense of vitality. Students' intrinsic motivation can be effectively supported by autonomy supportive communication from parents and teachers, which includes the following: highlighting the interesting features of assignments, helping students see how skills gained in school can be applied to help others, and understanding and acknowledging students' vantage point.

Another key motivational force is student expectations, which are often measured in terms of the extent to which students believe they will graduate from high school, college, graduate school, or professional school versus dropping out of high school. Froiland, Aubrey Peterson, and Mark Davison found that parent expectations in kindergarten predicted student expectations in eighth grade, which, in turn, predicted student achievement in math, science, and reading. Higher parent expectations also predicted that parents would become more involved in their children's education (e.g., read more books with them and provide more children's books).

Froiland and Davison also found that parent expectations are related to student achievement, good behavior, and positive parent–school relationships, characterized by trust and satisfying interactions with educators. Students with higher expectations also have higher intrinsic motivation to learn, suggesting that these two aspects of motivation can work in concert to promote greater levels of achievement and engagement at school. Both teacher and parent expectations have been linked to student expectations in numerous studies. Although there are currently not enough rigorously tested interventions to increase student expectations, some intervention studies have suggested that promoting parent and student expectations leads to higher achievement, finishing high school, and college enrollment.

John Mark Froiland

See also Creativity; Dropouts; Educational Psychology; Intelligence Quotient; Mastery Learning; Parenting Styles; School Leadership

Further Readings

Brophy, J. E. (2013). Motivating students to learn. Oxford, UK: Routledge.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. Psychological Review, 95, 256–273.

Froiland, J. M. (2011). Parental autonomy support and student learning goals: A preliminary examination of an intrinsic motivation intervention. Child and Youth Care Forum, 40, 135–149. doi:10.1007/s10566-010-9126-2

Froiland, J. M. (2013). Homework. In J. Ainsworth (Ed.), Sociology of education: An A-to-Z guide (pp. 362–363). Thousand Oaks, CA: SAGE.

Froiland, J. M. (2014). Inspired childhood: Parents raising motivated, happy, and successful students from preschool to college. Seattle, WA: Amazon. Retrieved from http://www.amazon.com/dp/B00LT4OX5O

Froiland, J. M. (2015). Parents’ weekly descriptions of autonomy supportive communication: Promoting children’s motivation to learn and positive emotions. Journal of Child and Family Studies, 24, 117–226. doi:10.1007/s10826-013-9819-x

Froiland, J. M., & Davison, M. L. (2014). Parental expectations and school relationships as contributors to adolescents’ positive outcomes. Social Psychology of Education, 17, 1–17. doi:10.1007/s11218-013-9237-3

Froiland, J. M., Mayor, P., & Herlevi, M. (2015). Motives emanating from personality associated with achievement in a Finnish senior high school:

Physical activity, curiosity, and family motives. School Psychology International, 36, 207–221. doi:10.1177/0143034315573818

Froiland, J. M., & Oros, E. (2014). Intrinsic motivation, perceived competence and classroom engagement as longitudinal predictors of adolescent reading achievement. Educational Psychology, 34, 119–132. doi:10.1080/01443410.2013.822964

Froiland, J. M., Peterson, A., & Davison, M. L. (2013). The long-term effects of early parent involvement and parent expectation in the USA. School Psychology International, 34, 33–50. doi:10.1177/0143034312454361

Kover, D. J., & Worrell, F. C. (2010). The influence of instrumentality beliefs on intrinsic motivation: A study of high-achieving adolescents. Journal of Advanced Academics, 21, 470–498.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. Journal of educational psychology, 97(2), 184–196.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55, 68–78.

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. Intelligence, 34(4), 363–374.

Weiher, G. R., Hughes, C., Kaplan, N., & Howard, J. Y. (2006). Hispanic college attendance and the state of Texas GEAR UP program. Review of Policy Research, 23(5), 1035–1051. doi:10.1111/j.1541-1338.2006.00248.x

Dmitriy Poznyak Dmitriy Poznyak Poznyak, Dmitriy

Mplus

Mplus

1095

1099

Mplus

Mplus is a statistical software package that can implement a wide array of statistical models, but it is primarily known for its latent variable modeling capabilities. Latent variables are unobserved variables that are measured by multiple observed variables, also called items, indicators, or manifest variables, using a statistical model. Latent variables are typically used to summarize different measurements of the same unobserved characteristic that cannot be measured directly (e.g., student’s socioeconomic status) and represent the “true” outcomes as opposed to the observed variables, which are measured with an error.

Mplus is typically used by students, applied researchers, and practitioners interested in latent variables modeling, which is commonly used in the areas of education, psychology, and other social science disciplines. This entry first reviews improvements and add-on modules offered in updated versions of the statistical package. Next, applications for Mplus are considered. The entry also explains many of the functions of the Mplus interface and how to produce Mplus output files. The ability of Mplus to provide an evaluation of a statistical model is also discussed, and the entry concludes with a section about how to obtain Mplus and supporting materials.

Mplus Versions and Modules

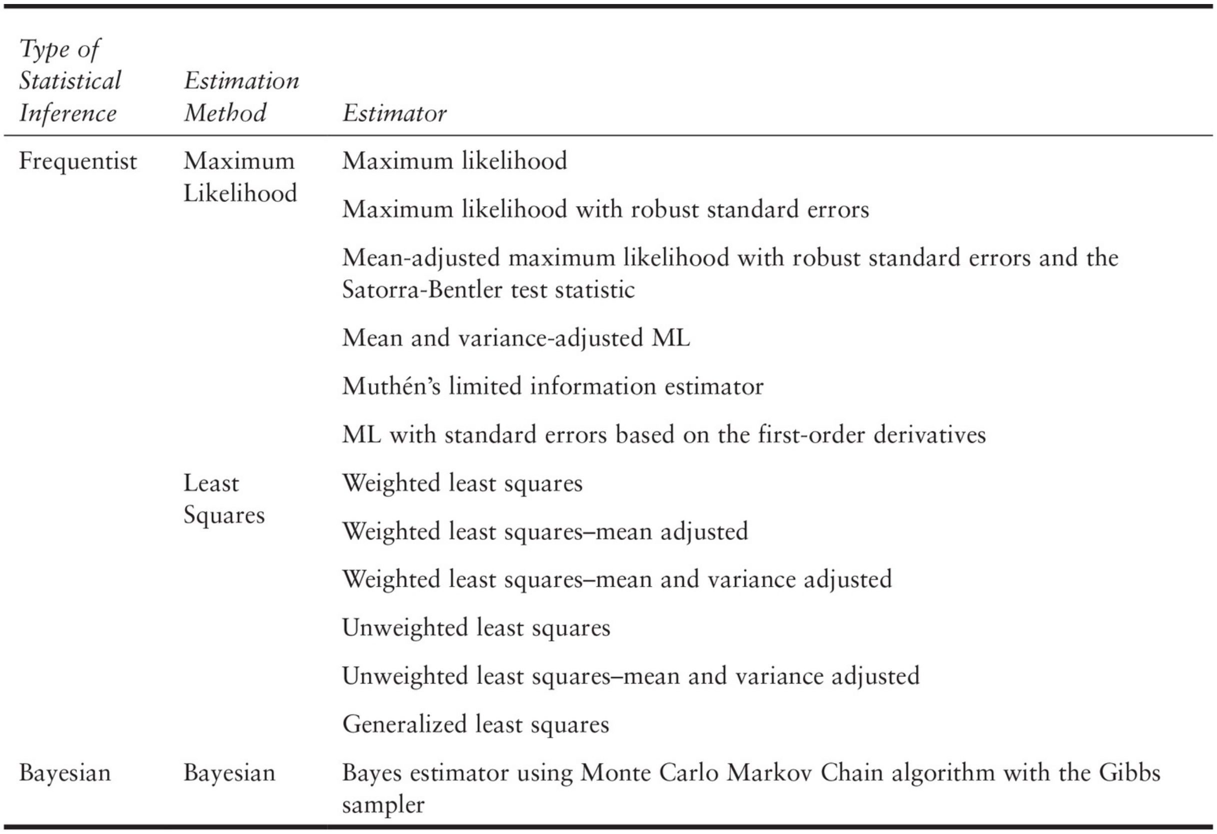
Between version 1.0, released in 1998, and version 7.4 8.0, released in 2015, Mplus has introduced numerous developments and improvements that make possible estimations of many different latent variable models with different data conditions, choosing from a wide number of estimators and algorithms for the analysis (an extensive review of the Mplus version history and features can be found on the program's website). The program is divided into the base program (Mplus Base) and three optional add-on modules. The base program allows the user to conduct exploratory factor analysis (EFA); confirmatory factor analysis (CFA); structural equation models; and regression, growth, and survival models with continuous, censored, binary, ordinal, nominal, and count variables or their combinations. The mixture add-on and multilevel add-on modules support a range of mixture models (such as latent class analysis) and multilevel models, respectively. The combination add-on module combines the features of the two individual add-ons and also enables estimation of several advanced models that combine the features of both modules (such as multilevel mixture models). Mplus can be installed on Windows, Mac OS, and Linux. The program is written in FORTRAN with graphical interface written in C and diagramming capabilities written in Java.

Applications

Mplus has flexible modeling capabilities, with the ability to estimate many different statistical models based on a wide variety of data types. The program can be applied to develop and validate scales (EFA and CFA), evaluate educational and psychological tests (categorical CFA and item response theory), discover unobserved groups in multivariate data (latent class and latent profile analyses), and estimate growth trajectory over time (latent growth analysis).

Mplus provides two approaches for analyzing complex survey data. The first approach takes into account stratification, nonindependence of observations due to cluster sampling, and/or unequal probability of selection when computing standard errors. The second approach, commonly referred to as multilevel or hierarchical modeling, models relationships between survey sampling and clustered standard errors by specifying a model for each level of the multilevel data.

Mplus supports both frequentist and Bayesian statistical analyses (see Table 1). The frequentist framework includes maximum likelihood and least squares estimators that can further be extended for a number of modeling situations.



Type of Statistical Inference	*Estimation Method*	*Estimator*
Frequentist	Maximum Likelihood	Maximum likelihood
		Maximum likelihood with robust standard errors
		Mean-adjusted maximum likelihood with robust standard errors and the Satorra-Bentler test statistic
		Mean and variance-adjusted ML
		Muthén's limited information estimator
		ML with standard errors based on the first-order derivatives
	Least Squares	Weighted least squares
		Weighted least squares–mean adjusted
		Weighted least squares–mean and variance adjusted
		Unweighted least squares
		Unweighted least squares–mean and variance adjusted
		Generalized least squares
Bayesian	Bayesian	Bayes estimator using Monte Carlo Markov Chain algorithm with the Gibbs sampler

Source: Muthén & Muthén (1998–2015).

Mplus applies default estimators based on the data type at hand to help ensure that the correct test statistics and standard errors are applied. However, users can override the defaults by specifying the ESTIMATOR option of the ANALYSIS command. The information about the default estimators and available options is referenced in the Mplus user guide written by Linda Muthén and Bengt Muthén. To estimate a model in which some of the variables have missing values, Mplus by default uses the full information maximum likelihood approach.

Mplus Interface

The data analysis workflow in Mplus involves three files: the file containing the data to be read into Mplus; the input file (.inp), which contains commands to read the data, estimate the model, and produce necessary output; and the output file (.out), which contains the model results. For the regression, path, CFA, structural equation models, growth, and survival analysis models, in addition to the output file, Mplus creates a diagram (.dmg file) that contains the graphical representation of the estimated model. The diagrams are not available for EFA models because the large number of possible paths between observed and latent variables would make it difficult to interpret.

Users have two options to enter data into Mplus. The first and most common option is to import a raw data set into Mplus. Mplus is only able to read ASCII (.dat) files, so data in other formats (such as Stata’s .dta, SPSS’s .sav or SAS’s .sas7bdat) must first be converted before they can be imported into Mplus. The Stat/Transfer software utility and Stata’s stata2mplus user-written program provide easy ways to convert files into ASCII format and also create the Mplus input (.inp) file with the commands needed to enter the data. If users do not have access to the individual-level data (e.g., students’ test scores) but have the estimates of the correlation between the variables (e.g., correlations between the college GPA and high school test scores), the second option—which may not be applicable for all types of analyses—is to only input the variable covariance or correlation matrix. To use this type of input, users need to create a free-format text file (e.g., CORRELATION_MATRIX.txt) that includes either the full or lower triangle of the correlation matrix without the variable names, for example.

Once the data have been entered, the statistical model can be specified either by writing code in the Mplus input file (the most common method), using a drop-down language generator, or through the graphical model specification.

Model specification using Mplus code involves a maximum of 10 language commands: TITLE, DATA, VARIABLE, DEFINE, ANALYSIS, MODEL, OUTPUT, SAVEDATA, PLOT, and MONTECARLO (see the Mplus website for additional information about the Mplus language). The TITLE command provides a title for the analysis. The DATA command provides information about the data set to be analyzed. The VARIABLE command provides information about the variables in the data set to be analyzed. The DEFINE command is used to transform existing variables and create new variables. The ANALYSIS command tells Mplus the technical details of the analysis. The MODEL command describes the model to be estimated in Mplus. The OUTPUT command requests additional output not included as the default. The SAVEDATA command is used to save the analysis data, auxiliary data, and a variety of analysis results as separate files. The PLOT command requests graphs of observed data and analysis results. Finally, the MONTECARLO command specifies the details of a Monte Carlo simulation study. Each of these 10 commands can be further expanded to refine the model specification and obtain the necessary output. For example, to specify a basic EFA model with continuous outcomes, one would use the following set of commands and options (subcommands):

```
FILE IS C:\Mplus\EFA.dat;
VARIABLE:
NAMES ARE ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6;
USEVARIABLES ARE ITEM1-ITEM6;
MISSING ARE ALL (-9999);
ANALYSIS:
TYPE IS EFA 1 2;
```

To input the correlation matrix in a .txt format, users will change the FILE statement to tell Mplus the data are correlations and specify the number of observations on which the correlations are estimated:

```
FILE IS CORRELATION_MATRIX.txt;
NOBSERVATIONS = 500;
TYPE=CORRELATION;
```

To estimate the EFA model with continuous outcomes, Mplus will use the default Geomin rotation and ML estimation methods. Users may change the rotation and estimation methods by using the corresponding options of the analysis command (e.g., ROTATION=VARIMAX; or ESTIMATOR=ULS).

An optional interactive language generator guides users through the series of steps that ask for information about the data and the desired analysis. The language generator is helpful for introducing new users to the commands needed to input data and estimate the most general models. However, additional commands and options (subcommands) may need to be manually added by the users. Mplus's website provides examples of code for estimating a variety of models. Graphical model specification in which a user-written model diagram will produce Mplus input statements is currently available for the regression, path, CFA, structural equation models, growth, and survival analysis models.

Mplus Output

The Mplus input file creates the output file that provides the information about the fit of the model (Model Fit Information section) and the model results (Model Results section). The output structure may vary depending on the model being estimated.

In the Model Fit Information section, Mplus reports fit statistics that show how well the tested model fits the data. For the EFA model previously presented, Mplus provides the relative fit indices used to compare the fit of nested models (which includes likelihood, χ^2, Akaike information criteria, and Bayesian information criteria) and absolute fit statistics used to compare the fit of the model to the established cutoff criteria (which includes root mean square error of approximation, comparative fit index, Tucker-Lewis index, and standardized root mean square residual). The model fit statistics provided by Mplus for the same model may vary depending on the type of the estimator selected. For instance, for Bayesian factor analytic models, Mplus does not provide the root mean square error of approximation, comparative fit index, and Tucker-Lewis index statistics. Instead, the program reports predictive posterior *p* value, Bayesian information criteria, and deviance information criteria, along with a variety of plots to check convergence of model parameters.

In the Model Results section of the example EFA analysis, Mplus provides unstandardized (default) and standardized (optional) point estimates (i.e., correlations between the variable and the factor) and their standard errors. Mplus does not, by default, compute the reliability of the estimated factors, but the model-based reliability coefficients (such as McDonald's Ω) can be requested with the additional code (subcommands). Note that the statistics in the model fit and model results sections will differ somewhat if the data file contains individual observations (data input Option 1) versus the correlation matrix only (data input Option 2).

Evaluation of a Statistical Model

A model's statistical adequacy can be determined by evaluating the global and local fit statistics provided by Mplus. The global fit statistics show how well the overall model fits the data. For factor analytic models in Mplus, global fit statistics include the aforementioned relative and absolute fit indices. The local fit indices indicate whether a model yields reasonable point estimates and standard errors. For instance, excessively large or small standard errors or negative error variances may suggest issues with model fit. For a variety of models, Mplus also provides model modification indices that indicate potential issues with the model and suggest changes that could improve its fit to the data. However, it is worth a reminder that, in addition to evaluating quantitative elements of Mplus output, a careful model-building and selection process should also rely on a review of qualitative elements such as a model's plausibility and also rely on a review of qualitative elements such as a model's plausibility and parsimony.

Availability: Download and Materials

Mplus can be purchased and downloaded from the company’s website. Mplus has different pricing for students, universities, and commercial/nonprofit/government users. A discount is available to those purchasing multiple copies of Mplus licenses as well. The Mplus website contains manuals, program updates, articles, training materials and presentations, video tutorials, and group discussions. Mplus also offers technical support to the program users.

Dmitriy Poznyak

See also LISREL; Path Analysis; Structural Equation Modeling

Further Readings

Byrne, B. (2011). Structural equation modeling with Mplus Basic concepts, applications, and programming. New York, NY: Routledge.

Geiser, C. (2013). Data analysis with Mplus. New York, NY: Guilford Press.

Hancock, G. R., & Mueller, R. O. (Eds.). (2013). Structural equation modeling: A second course (2nd ed.). Charlotte, NC: Information Age.

Hoyle, R. H. (Ed.). (2012). Handbook of structural equation modeling. New York, NY: Guilford Press.

Muthén, B. O., & Muthén, L. K. (n.d.). Mplus: A general latent variable modeling program. Retrieved from https://www.statmodel.com/download/Mplus-A%20General%20Latent%20Variable%20Modeling%20Program.pdf

Muthén, L. K., & Muthén, B. O. (1998–2015). Mplus user’s guide (7th ed.). Los

Angeles, CA: Author. Retrieved from http://www.statmodel.com/download/usersguide/MplusUserGuideVer_7.pdf

Wang, J., & Wang, X. (2012). Structural equation modeling. Chichester, UK: Wiley.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), Handbook of structural equation modeling. New York, NY: Guilford Press.

Websites

Mplus: http://www.statmodel.com

Kim Nimon Kim Nimon Nimon, Kim

Multicollinearity Multicollinearity

Multicollinearity

Multicollinearity refers to the linear dependence among two or more variables. Although collinearity technically refers to the linear dependence among two variables, multicollinearity and collinearity are often used interchangeably. When there is a perfect linear dependence among predictors, statistical analyses such as multiple linear regression cannot be conducted with all included variables as the regression equation becomes unsolvable.

Consider a scenario where a researcher wants to know whether variability in test scores is a function of hair color (i.e., brown, black, red, and blond). Although an analysis of variance would likely be the statistical analysis of choice, the general linear model indicates that multiple linear regression could also be used. Of course, the variable hair color could not be used in its original form given its categorical state. Therefore, the researcher would have to dummy code hair color into several variables.

Let’s pretend that the researcher created a dichotomous variable for each hair color (0 = *is not the color*, 1 = *is the color*). If a regression was conducted where brown, black, red, and blond were used as independent variables, the solution would be inadmissible, as there is a linear dependence among the independent variables (e.g., blond is known, given brown, black, and red). However, if one of the hair colors (e.g., black) was left out of the equation and interpreted as the intercept, the regression analysis would run just fine. Full multicollinearity can also occur when variables are perfectly corrected (e.g., $r_{X1.X2} = 1$). Although it is hard to imagine valid scenarios in the education field where predictors are perfectly correlated, it is much more likely that predictors are correlated but not to the point of achieving a perfect correlation (i.e., partial multicollinearity).

Imagine, for example, that a researcher wanted to determine how measures of engagement for first-year teachers related to their intent to stay. Measures of engagement could include vigor, dedication, and absorption, and while the researcher would not expect the correlations among the variables to be perfect, it would be very likely that there would be substantive correlations (e.g., $r > .5$) among the predictors. In this scenario, the regression analyses would run just fine. However, the results would be more difficult to interpret, as the regression coefficients would not indicate measures of relationship. Further, just because one variable (e.g., $X1$) had a low regression coefficient would not mean that it did not have a substantive relationship with the dependent variable. When a variable (e.g., $X1$) is correlated with another predictor (e.g., $X2$), some of one variable's (e.g., $X1$) credit to the regression effect may be captured by another variable's (e.g., $X2$) regression coefficient.

Situations like the one just described have led some researchers to suggest that multicollinearity is a problem in multiple regression and other general linear model analyses. However, if researchers analyze regression weights along with other measures of importance including structure coefficients (or bivariate correlations) and commonality analysis, multicollinearity is not a problem, as these techniques identify the presence, loci, and magnitude of multicollinearity. In the presence of multicollinearity, researchers should look beyond regression weights and fully interpret regression effects. Such results will ensure that regression results are not only properly interpreted but can also be used to inform theory and practice.

Kim Nimon

See also Analysis of Variance; General Linear Model; Multiple Linear Regression

Further Readings

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: ß is not enough. Educational and Psychological Measurement, 61, 229–248. doi:10.1177/0013164401612006

Nimon, K., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. Organizational Research Methods, 16, 650–674. doi:10.1177/1094428113493929

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. Educational Researcher, 38, 343–352. doi:10.3102/0013189X09339056

Kayla Tureson Kayla Tureson Tureson, Kayla

Anthony Odland Anthony Odland Odland, Anthony

Multicultural Validity Multicultural validity

1100

1102

Multicultural Validity

Multicultural validity is a construct used to describe the accuracy or authenticity of assessments and evaluative judgments across cultural differences. There is an emphasis on cultural awareness and attending to cultural issues and differences, specifically in terms of evaluation, in order to determine that valid assessments and judgments can be made. Validity lies at the core of evaluation to determine the extent to which empirical evidence and theoretical rationales support the accuracy of judgments or inferences made from test scores or other assessment methods. A key component in examining the validity of an assessment is determining the equivalence of the results, the extent to which a meaning of a score and its implications hold across different settings and contexts as well as across various populations. A number of test factors need to be considered during interpretation of assessment results, including the characteristics and test-taking ability of the person being assessed, as well as personal, situational, linguistic, and cultural differences that may influence evaluative judgments. This entry outlines the concept of multiculturalism, multiculturalism in terms of validity, as well as further considerations regarding multicultural validity in evaluation.

Basic Concepts of Multiculturalism and Validity

In order to better understand multiculturalism, it is important to first define the construct of culture. Culture encompasses a collective set of values, beliefs, knowledge, skills, and attributes that create a distinct identity. This identity shapes how people think and act, and cultural understandings may be shared among a large group of people and across generations. Karen Kirkhart identifies multiplicity, fluidity, and nonneutrality as three important facets of culture. Multiplicity refers to culture as a multidimensional construct of identity, a sense of interconnectedness shaped at the individual and collective group levels. In terms of fluidity, culture is not a fixed construct, as it changes with each situation, task, role, as well as with time. The third facet of culture, nonneutrality, is the acknowledgment that culture itself is not neutral. Although not always explicitly identified, power is attached in varying degrees to different aspects of culture, and the extent of its attachment can change from one context to another. The dominant majority perspective is the default societal viewpoint of culture. An example of the power expression of the dominant majority culture in the United States is the often limited availability of translations for medical documents or test measures in languages other than English. Analyses of culture and context should include power dynamics, and analysis should acknowledge that power dynamics privilege certain cultural identifications.

The concept of multiculturalism is the awareness, recognition, and support of cultural diversity. When applied to the construct of validity, multiculturalism brings awareness of cultural differences in determining the extent to which results of an assessment are valid. Multicultural validity can be categorized by three constructs of validity: internal, external, and construct. In terms of internal validity, cultural factors are important in understanding which variables influence results. External validity/ecological validity concerns the generalizability of results to other settings and times. Construct validity refers to the consequences and implications the effects have on higher order constructs. For example, use of the English version of the Minnesota Multiphasic Personality Inventory with bilingual patients should be carefully considered as to whether the test will accurately reflect the personality of the individual, as cultural differences and unfamiliarity with test content may affect item endorsement and score patterns. Failure to take cultural nuances into account when interpreting the validity of results may result in incorrectly diagnosing and stereotyping an entire cultural group. Multicultural validity imbues traditional concepts of validity by enhancing test selection, enriching interpretation, and strengthening the appropriateness and effectiveness of recommendations.

Considerations in Evaluation

Validity concerns the meaning of test scores, which includes the test items as well as the characteristics of the examinee and context of the evaluation. The evaluation parameters need to incorporate culturally relevant variables in order evaluation parameters need to incorporate culturally relevant variables in order to reach valid interpretations and recommendations. For example, evaluators may expand their understanding of pertinent cultural dimensions by researching a given culture's history.

Evaluators should be cognizant of shared values, beliefs, aspirations, and ideals that exist at the individual and group level. Particular attention should be paid as to how social problems are defined within the culture as well as who defines them. It is crucial that an evaluator is mindful of the cultural equivalence of the assessment as well. In terms of linguistic equivalence, if a test measure is translated, semantics must be evaluated to ensure similar validity and reliability. For example, the Chinese translation of the Minnesota Multiphasic Personality Inventory changed a test item referring to a popular piece of American literature to a more culturally relevant piece of Chinese literature. Although the test item was eventually removed, it helped to control for random responding due to unfamiliarity with the test item content. Functional equivalence should in part take into consideration whether behaviors serve similar functions in different cultures. Psychological concepts may differ in definitions, applicability, acceptability, and presentation from one culture to another. Multicultural validity requires attention to cultural differences in assessments and evaluative judgments, as well as consideration for generalizability and potential consequences.

Kayla Tureson and Anthony Odland

See also African Americans and Testing; Cross-Cultural Research; Demographics; Diagnostic Tests; Gender and Testing; Generalizability; Predictive Validity; Second Language Learners, Assessment of; Threats to Research Validity; Validity Generalization

Further Readings

Conner, R. F. (2004). Developing and implementing culturally competent evaluation: A discussion of multicultural validity in two HIV prevention programs for Latinos. New Directions for Evaluation, 2004(102), 51–65.

Hanberger, A. A., & Umeå universitet, S. O. (2010). Multicultural awareness in evaluation: Dilemmas and challenges. Evaluation, 16(2), 177–191. doi:10.1177/1356389010361561

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. Evaluation Practice, 16(1), 1–12.

Kirkhart, K. E. (2010). Eyes on the prize: Multicultural validity and the evaluation theory. American Journal of Evaluation, 31(3), 400–413. doi:10.1177/1098214010373645

Kwan, K. K., Gong, Y., & Maestas, M. (2010). Language, translation, and validity in the adaptation of psychological tests for multicultural counseling. In Handbook of multicultural counseling (3rd ed., pp. 397–404). Thousand Oaks, CA: SAGE.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.

Ridley, C. R., Tracy, M. L., Pruitt-Stephens, L., Wimsatt, M. K., & Beard, J. (2008). Multicultural assessment validity: The preeminent ethical issue in psychological assessment. In L. A. Suzuki & J. G. Ponterotto (Eds.), Handbook of multicultural assessment: Clinical, psychological, and educational applications (pp. 22–33). San Francisco, CA: Jossey-Bass.

W. Holmes Finch W. Holmes Finch Finch, W. Holmes

Multidimensional Item Response Theory Multidimensional item response theory

1102

1105

Multidimensional Item Response Theory

Item response theory (IRT) is a core analysis for test developers and researchers working with questionnaires, surveys, psychological measures of mood and cognition, and educators interested in academic achievement. It serves to provide information about both items and individuals in a comprehensive, connected framework. Basic IRT models make an assumption that the set of items measures a single common construct, such as intelligence. However, in reality, many constructs are multidimensional in nature, meaning that they consist of what can be thought of as several subconstructs. Standard IRT techniques are not able to accurately model such item responses because of their assumption of unidimensionality. For this reason, researchers working with multidimensional constructs need an alternative modeling paradigm, which comes in the form of multidimensional IRT (MIRT). This entry first briefly review unidimensional IRT models and then extend those to the multidimensional context by focusing on two popular ways in which these models can be viewed. The entry concludes by describing software options that researchers have when trying to fit MIRT models.

Unidimensional IRT Models

Unidimensional IRT refers to a set of statistical models designed for use with responses to items on tests, questionnaires, and other such instruments in order to obtain estimates of individuals’ levels on the construct measured by the scale as a whole. For example, a commonly used model is the three-parameter logistic (3PL) model, which contains a parameter specific to respondents (person parameter) and three parameters specific to the items measuring the construct of interest. The person parameter is the estimate of the latent trait being measured by the scale (e.g., reading ability and depression) and is referred to as θ. The item parameters include (a) location on the latent trait scale, (b) the item's ability to differentiate among individuals with different levels of the construct, and (c) the likelihood that an individual will endorse the item due solely to chance. The construct itself is a latent variable that is measured by the set of items, which serve the role of indicators much as do the observed variables in factor analysis models. Indeed, there is a direct relationship between IRT and factor analysis models and their parameters, so that one can be easily converted to the other. IRT models can be divided into two broad families based on whether they model dichotomous item responses with two categories or polytomous items with three or more categories.

The simplest unidimensional IRT model is the Rasch model, which is expressed as

$$Px_j = 1 \mid \theta,\ b_j = e\theta, -b_j 1 + e\theta, -b_j.$$

In Equation 1, x_j is the response to item *j* with 1 being correct in the context of an achievement test and 0 being incorrect. An individual's level of the latent trait being measured by the set of items is represented by θ, and the item location is b_j. For a math test, θ corresponds to an examinee's math ability while b_j is the difficulty of item *j*. For a depression inventory, θ would be the patient's level of depression and b_j the likelihood of an individual endorsing the behavior measured by item *j*. An important strength of all IRT models is that item difficulty and examinee ability are placed on the same scale. Therefore, it is possible to directly compare where an individual lies on the latent trait scale with the location of any item on the instrument. In addition, *b* and θ are both centered at 0, which represents average or typical location for both.

The Rasch model is a special case of what is known as the one-parameter logistic (1PL) model. What distinguishes these two models is that the item discrimination value (*a*) is set equal to 1 for the Rasch and freely estimated for the 1PL. Discrimination refers to the ability of an item to differentiate among individuals with different levels of the latent trait of interest, so that items with larger discrimination values are better able to do this than items with smaller values. All 1PL models hold discrimination equal across items but do estimate this common value. The 1PL model can be expressed as

$$Px_j = 1 \mid \theta,\ b_j = ea\theta, - b_j 1 + ea\theta, - b_j.$$

As just noted, for both the Rasch and 1PL models, a single item discrimination value exists for all of the items, either set at 1 in the case of Rasch or a single estimated value for the 1PL. However, this tacit assumption of equal discrimination across items may not be reasonable in many cases, leading researchers to investigate the fit of the two-parameter logistic (2PL) model, which allows for unique discrimination values for each item. The 2PL model takes the following form:

$$Px_j = 1 \mid \theta,\ b_j = ea_j\theta, - b_j 1 + ea_j\theta, - b_j.$$

The parameters in Equation 3 are the same as in Equations 1 and 2, with the exception that there are unique discrimination values for each item.

For some types of items, respondents may be able to obtain a correct response or endorse the item due solely to chance. Consider, for example, a multiple-choice math test in which examinees can make a guess at the correct answer if they don’t know what it is. Therefore, the key addition to the 3PL model, beyond the 2PL, is the inclusion of a pseudo-chance parameter (*c*). This value estimates the likelihood that an individual will endorse an item due solely to chance.

Returning to our math test example, it is possible that an examinee could answer an item correctly by guessing. Naively, we may assume that on a multiple-choice exam with five options per item, the probability of correctly guessing would be 1/5 or 0.2. However, for most multiple-choice items, some incorrect response options are more appealing compared to others, meaning that examinees who do not know the correct answer may be able to eliminate some options and thereby increase the likelihood of a correct guess. Conversely, some response options might be so attractive that they are more likely to be selected by individuals who do not know the correct answer than would be expected in the case of simple random guessing. Therefore, *c* will not always equal 1 divided by the number of available response options. As a separate matter, including *c* is not appropriate for every situation involving dichotomous items. Indeed, only when an individual could realistically endorse an item solely due to chance would researchers want to use the 3PL model. Thus, whereas the 3PL model may be useful for a multiple-choice math test, it will probably not be helpful in modeling a depression inventory asking whether individuals engaged in specific behaviors or had specific thoughts over the last month. The 3PL model is expressed as

$$Px_j = 1 \mid \theta,\ b_j = ci + 1 - ciea_j\theta, -b_j 1 + ea_j\theta, -b_j.$$

Assumptions Underlying Unidimensional IRT Models

The unidimensional IRT models described in the previous section rest on three foundational assumptions: (1) monotonicity: the relationship between the latent trait and the probability of item endorsement is monotonically increasing; (2) local independence: when the latent trait is controlled for, there is no correlation between item responses; and (3) unidimensionality: only a single latent trait is being measured by the set of items. Our interest in this section is on the last assumption, which essentially means that the set of items have a single common construct underlying them. If this assumption is violated, research has shown that estimates of both person and item parameters are compromised. Therefore, it is important both that researchers assess the dimensionality of their scales and that when unidimensionality is violated, they use an appropriate multidimensional model. Prior to describing these models, it should be noted that there exists a variety of methods for assessing whether a set of items are in fact unidimensional, including factor analysis models for IRT data, conditional covariance approaches to identify dimensions in the data, and hypothesis tests for unidimensionality. These methods are not the focus of the current discussion but should be used to determine whether a unidimensional or multidimensional model is appropriate for a set of data.

Compensatory MIRT Models

When data are known to be multidimensional in nature, such that the items measure more than one latent trait, researchers will want to fit a model appropriate to this reality. In this section, we consider two common MIRT models. In many respects, these MIRT models are similar in form to confirmatory factor analysis models, with multiple indicators (items in the case of MIRT) associated with each latent trait and typically no indicator associated with more than one trait. The first model that we will examine is the standard compensatory 2PL MIRT model, which can be expressed as

$$Pu_{i}=1\,|\,\theta_{i}=ek=1Kaik\theta jk+di1+ek$$

$$=1Kaik\theta jk+di,$$

where *aik* = discrimination for item *i* on latent trait *k*, θ*jk* = level of latent trait *k* for person *j*, and *di* = difficulty for Item *i*.

The interpretation of the item and person parameters is similar as in the unidimensional case, with the exception that there are multiple latent traits of interest so that we have more than one item discrimination value and more than one θ estimate for each individual in the sample. This model is known as compensatory MIRT because the impact of the latent traits is summative, so that deficits on one can be at least partially offset by higher values on the other. There also exist noncompensatory models, which multiply rather than sum the latent traits together. However, these models are much more difficult to fit than the compensatory ones and are also much less widely used in practice.

In addition to the standard item discrimination and difficulty parameters, it is also possible to obtain multidimensional discrimination and difficulty parameter estimates from the MIRT model. These statistics provide overall measures of item characteristics, in contrast to the unidimensional information contained in the *aik* and *di* parameters of Equation 5. These multidimensional item parameters are calculated as follows:

$$\mathrm{MDISC} = K = 1kaik2\,\mathrm{MDIFF} = -di\mathrm{MDISC}.$$

These values can be used to describe the overall discrimination and difficulty of items, across all dimensions that they measure.

Bifactor Model

A second very common way in which MIRT models can be specified is as a bifactor model. With the bifactor model, each item is associated with a general factor and with what is called a specific factor. There is only a single general factor with which all items are associated and then multiple general factors with which subsets of the items are associated. As an example, rather than conceive of each item on an IQ test as being uniquely associated with a single latent trait (IQ), we might think of each item as being associated with the single general trait of IQ and also associated with a unique component of IQ such as cognitive processing speed. Thus, the item would have two factor loadings/discrimination parameters, one for the general latent trait and one for the specific latent trait. This model differs from the general compensatory MIRT model in that for the bifactor model, each item is associated with two dimensions. However, at a more basic level, it also represents a different way in which we might view the mechanism underlying item responses. Researchers deciding which approach to use should consider what they believe to be the underlying item response mechanism and make decisions regarding which is optimal accordingly.

mechanism and make decisions regarding which is optimal accordingly.

Software for MIRT Models

Although there is a great deal of software available for fitting unidimensional IRT models, there are fewer such options for fitting MIRT models. One worthwhile option is the R software package, which features the MIRT library of functions for this purpose. Another possibility is the FlexMIRT software package, designed specifically for this purpose. IRTPRO is a general use IRT modeling package that has some capabilities for fitting MIRT models, and the latent variable modeling software packages Mplus and EQSIRT are also viable options for this purpose.

W. Holmes Finch

See also Confirmatory Factor Analysis; FlexMIRT; IRTPRO; Item Response Theory

Further Readings

de Ayala, R. J. (2009). The theory and practice of item response theory. New York, NY: Guilford Press.

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. Applied Measurement in Education, 21, 119–140.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

Reckase, M. D. (2009). Multidimensional item response theory. London, UK: Springer.

Patrick Mair Patrick Mair Mair, Patrick

Multidimensional Scaling Multidimensional scaling

1105

1107

Multidimensional Scaling

Multidimensional scaling (MDS) is a technique that represents proximities among objects as distances among points in a low-dimensional space. It allows researchers to explore similarity structures among objects (e.g., persons and variables) in a multivariate data set. Early MDS developments can be traced back to the late 1950s and the 1960s. In the 1970s, technical MDS details were worked out and important MDS extensions were proposed. During that time, MDS software was developed and a first peak of MDS applications was reached. Since then, MDS has been widely applied in fields like psychology, marketing, political sciences, ecology, and several others. In this entry, the basic principles of MDS are highlighted and extensions of MDS are examined.

Basic Principles

An easy way to explain the basic principles of MDS is to consider a simple example involving 20 cities (*objects*). The data consist of distances (as the crow flies) in kilometers between each pair of cities. These distances can be organized in a 20 × 20 symmetric matrix. With this distance matrix as the input, MDS produces a geographic map by computing the coordinates (basically, longitude and latitude) for each city. In MDS terminology, such a representation is called a *configuration*.

Applications in the social and behavioral sciences are typically not based on geographic input distances. Instead of distances, proximities (i.e., similarities or dissimilarities) are used as the input, which can be computed easily from an ordinary person × variables matrix. Popular proximity measures are correlation coefficients, a Euclidean distance measure, a Jaccard coefficient, and so on. Depending on whether the rows or the columns of the data matrix are subject to scaling, the proximity measure of choice is applied to either the rows or the columns. From this point in this entry, all explanations are limited to dissimilarities because most MDS software packages require dissimilarities as input. Note that similarities can be easily converted into dissimilarities (and vice versa).

Formally, MDS takes a symmetric input dissimilarity matrix Δ of dimension *n* × *n* and computes a configuration *X* in a space of dimension *p*. In the cities example provided, it was quite natural to choose two dimensions (map), even though three dimensions would have made sense as well (globe). Similar to principal components analysis, *p* needs to be fixed a priori. In MDS, researchers typically aim for a small *p* (e.g., *p = 2* or *p = 3*) so that the configuration can be easily plotted. In order to compute *X* on the base of Δ, a target function needs to be formulated and solved (i.e., minimized). The most popular target function in MDS is Kruskal’s *stress*, which, in its simplest form, can be expressed as

$$\text{stress} = \sum_{i<j} \left(\hat{d}_{ij} - d_{ij}(X) \right)^2 .$$

Minimizing this function implies that the distances among the points in the MDS space should be as close as possible to the input dissimilarities δ_{ij}, or, to be more precise, a transformed version of them: . The resulting outcomes are called *disparities* or, simply, *d-hats*. The most popular transformation functions (*optimal scaling*) are a monotone step function (which leads to Kruskal’s *nonmetric* or *ordinal MDS*) or a linear regression function of the form (called *interval MDS*, or, if *a* = 0, *ratio MDS*). The choice of the transformation function gives the user some modeling flexibility (e.g., considering the input dissimilarities to be on an ordinal or on a metric scale level).

The smaller the stress, the better the MDS fit (i.e., the better the fitted distances approximate the transformed observed dissimilarities). Finding the solution that leads to the smallest possible stress value is not a trivial task. It can happen that the algorithm does not necessarily end up in the global minimum but rather gets stuck in a local minimum. This behavior can depend on the starting solution for the configuration. MDS software tools typically provide a good initial guess but, in practice, it is suggested that the user tries out different initial (random) configurations and explores how the resulting stress values behave.

Goodness-of-fit assessment in MDS is a task that needs special consideration, and researchers should not rely on stress rules of thumb. In general, it holds that the larger *p*, the smaller the stress. The choice of *p* is one facet of goodness-of-fit assessment; however, other aspects to consider are the interpretability/theoretical relevance of a configuration (i.e., in case there is a substantive theory in the background) and the stability/replicability of a solution across multiple samples. Graphical tools that support researchers with goodness-of-fit assessment are the scree plot (using the elbow criterion just as in principal components analysis) and the Shepard diagram (displays the transformation function). Other diagnostics include the *stress per point* for detecting points that (heavily) influence the fit of an MDS solution.

Once a satisfactory MDS model is chosen, the main output is the configuration plot, which allows researchers to explore structural properties of the data. Sometimes it is possible to find a meaningful/relevant interpretation of the dimensions; in other applications, geometric patterns of the MDS space (clusters, regions, etc.) are subject to interpretation. A normalized version of the stress value (called *Stress 1*) is typically reported as a global fit measure.

MDS Extensions

An important, early extension of MDS are *three-way MDS* models. A “way” in MDS terminology denotes the dimensionality of the input matrix. So far, only two-way models have been considered. Three-way models take multiple dissimilarity matrices $\Delta_1, \cdots, \Delta_K$ as input; each of these is of dimension $n \times n$, which makes it possible to arrange them in a three-dimensional array structure. Such data structures are often collected in psychological experiments in which each individual rates multiple proximities among objects. The stress equation previously provided can be extended by incorporating *K* individual configurations $X_1, \cdots, X_K$. It is assumed that each individual configuration is generated from a common group space by individually weighting a common set of dimensions of the group space (INDSCAL) or an individually chosen set of dimensions of the group space (IDIOSCAL). The group configuration is typically subject to plotting, unless researchers are particularly interested in (some of) the individual configurations.

The MDS models discussed thus far were unrestricted and of exploratory nature. Now, let us consider restricted, *confirmatory MDS* models. One way of posing restrictions on the configuration is through geometric shapes: spherical restrictions (i.e., a circle in two dimensions) are the most popular shape because in some applications (e.g., color perception and personal values) there are underlying theories stating that the resulting points can be arranged on a circle. In order to examine such theories, researchers can fit an unrestricted as well as a restricted solution and compare the stress values. If the restricted stress value is not much larger than the unrestricted stress value, there is evidence in favor of the theory.

Another way of restricting MDS solutions is through external variables. Such variables can be metric covariates, ANOVA-like designs, regional constraints, and so on. Again, an externally restricted solution can be compared to an unrestricted solution in order to explore to which degree such external variables influence the solution.

The final model family considered here are unfolding models. Compared to basic MDS, the input structure is slightly different: Δ is not square and its elements are typically preference scores (such as rank orders of preference) of different individuals for a set of choice objects. These preference values can be converted into dissimilarities, which makes unfolding a special case of MDS. In MDS terminology, where each set of objects to be scaled is called a *mode*, unfolding models are considered as two-mode models. A stress expression can be established that involves two configuration matrices: one for the rows (X_1) and one for the columns (X_2). Thus, unfolding scales the rows and columns of the input matrix and is therefore called a *dual scaling* method. Individuals are represented as “ideal points” so that the distances from each ideal point to the object points (columns) correspond to the preference scores.

All these extended MDS models can be estimated in an ordinal (nonmetric) fashion as well as in a ratio/interval (metric) fashion.

Patrick Mair

See also Exploratory Factor Analysis; Goodness-of-Fit Tests

Further Readings

Borg, I., & Groenen, P. J. F. (2005). Modern multidimensional scaling: Theory and applications (2nd ed.). New York, NY: Springer.

Borg I., Groenen, P. J. F., & Mair, P. (2013). Applied multidimensional scaling. New York, NY: Springer.

Busing, F. M. T. A., Groenen, P. J. F., & Heiser, W. J. (2005). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika, 70, 71–98.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. Annual Review of Psychology, 31, 607–649.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. Psychometrika, 35, 283–320.

Coombs, C. H. (1964). A theory of data. New York, NY: Wiley.

Cox, T. F., & Cox, M. A. A. (2001). Multidimensional scaling (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. Journal of Statistical Software, 31(3), 1–30.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, 1–27.

Mair, P., Borg, I., & Rusch, T. (2016). Goodness-of-fit assessment in multidimensional scaling and unfolding. Multivariate Behavioral Research, 51(6), 772–789.

Kimberly Ethridge Kimberly Ethridge Ethridge, Kimberly

Anthony Odland Anthony Odland Odland, Anthony

Multiple Intelligences, Theory of Multiple intelligences, theory of

1108

1110

Multiple Intelligences, Theory of

The theory of multiple intelligences is a multidimensional view on intelligence and focuses on multiple abilities and domains, rather than a single mode or type of intelligence. Psychologist Howard Gardner first put this theory forth in his book *Frames of Mind: The Theory of Multiple Intelligences*. Because the definition of intelligence is abstract and open to many different interpretations, Gardner made a list of eight criteria that a behavior had to meet in order to be considered an intelligence. Additionally, he chose eight areas of intelligence that met his criteria and that he considered the most important pieces of intellect. This entry discusses the criteria, the eight modes of intelligence, the limitations of Gardner's theory, and critiques of the theory.

Criteria for Intelligence

Gardner established eight criteria that must be met in order for a behavior or action to be considered an intelligence. The criteria sought to establish an empirical basis not only for this theory but also for the study of intelligence in general. Gardner based his modalities of intelligence on the following criteria: psychometric validation; experimental psychology research; the possibility and the presence of individuals who would be considered geniuses, prodigies, or savants; a developmental time frame that could be analyzed; ability for the behavior to be encoded through the use of symbols or symbolic expression; an evolutionary importance; an identifiable core operation; and the ability for the behavior to be isolated in the brain due to brain damage.

Eight Types of Intelligence

Eight Types of Intelligence

Gardner used the criteria discussed in the previous section to establish eight domains of intelligence, pushing back against the convention of a single type of intelligence that forms the basis of the most widely accepted ideas about intelligence in the academic literature. Gardner’s goal was to move away from the restricted view of intelligence in order to account for more behaviors and types of people. However, Gardner’s theory also accepted the idea of a *g* factor, which refers to a person’s general mental ability as reflected by his or her IQ score.

Many of Gardner’s modes of intelligence correlate with standard intelligence as measured by an IQ test. This was backed by empirical research suggesting there is a single dominant or stronger type of intelligence and that what Gardner identifies as intelligences are made up of many factors, namely *g* factor, other cognitive abilities, and noncognitive factors such as personality. This was very similar to the research suggesting that typically thought of intelligence and IQ scores can be a combination of environmental factors and cognitive ability. The eight modalities are described in greater detail in the remainder of this section.

Musical—Rhythmic and Harmonic

Musical intelligence is comprised of the ability to sing, play musical instruments, and compose pieces of music. Having musical intelligence also suggests the individual has good or perfect pitch, can differentiate sounds and tones better than others, and has more natural rhythm. Furthermore, musical intelligence according to Gardner means the person is sensitive to various components of music and sound such as pitch, tone, meter, and rhythm and can better discern melodies and musical counts.

Visual–Spatial

This domain is concerned with spatial relationships, the ability to visualize images and scenarios in one’s mind, and being adept at perceptual reasoning. Examples of this intelligence would be an increased ability to read maps, visualize objects in abstract space, or reason through puzzles. Visual–spatial ability is oftentimes tested on traditional intelligence tests, such as the Wechsler Adult Intelligence Scale–Fourth Edition, in which there is an entire index devoted to visual–spatial functions. Visual–spatial intelligence is a component of the *g* factor in overall intelligence that is typically talked about in academic settings.

Logical–Mathematical

Logical–mathematical intelligence is involved in procedures of critical thinking. This involves the skills of abstraction, reasoning, use of numbers, and logic. Additionally, logical–mathematical intelligence encompasses the ability to determine cause and effect or casual relationships of behaviors, events, systems, and situations. This type of intelligence is also part of the traditional view of intellect and the overall *g* factor.

Verbal–Linguistic

Verbal intelligence is an area that is most often associated with general overall intellect and is highly correlated with the *g* factor associated with such intelligence. This modality is associated with proficiency for words and languages as well as increased ability for reading, comprehension, writing, memorizing words, and expressing ideas. As with the visual–spatial modality, this area is measured on conventional intelligence tests, particularly in the verbal index score from the Wechsler Adult Intelligence Scale–Fourth Edition.

Bodily–Kinesthetic

This modality of intelligence has to do with physical motor function and the ability to control those functions. Gardner stated that the bodily–kinesthetic intelligence encompassed control of bodily movements, awareness of one’s body in relation to the environment, and ability to maintain coordination when handling objects, mainly involving fine and gross motor skills. Skills in this area include a sense of timing for movements and the capability to train physical movements and responses to various actions. Many individuals with high levels of this type of intelligence are athletes, dancers, actors, or soldiers, as they require the ability to train and coordinate complex movements.

Interpersonal

Gardner described this intelligence as the ability to understand others’ thoughts, feelings, mood, and personality. Interpersonal intelligence can also factor into how an individual works as part of a team and suggests that these individuals may work best in either a leadership or follower role because of their emotional aptitude. Effective communication and empathy are also components of this modality, and Gardner likened interpersonal intelligence to the concept of emotional intelligence that was already existent in the literature.

Intrapersonal

Intrapersonal intelligence deals with the interaction with oneself, the opposite of interpersonal intelligence. Intrapersonal intelligence involves self-reflection and introspection, the ability to look inward and understand one’s self in regard to strengths and weaknesses. Additionally, this involves the skill to be able to understand and eventually predict one’s mood, reactions, emotions, and behaviors when faced with certain situations. It is the ability to know and understand oneself.

Naturalistic

Naturalistic intelligence was not in the original theory proposed by Gardner but was eventually included in his eight main intellectual modalities. Naturalistic intelligence encompasses an individual’s ability to be able to relate to the surrounding natural environment; make differentiations and evaluations about animals, plants, and geography; and understand the role that humans have in the environment. This type of intelligence was postulated to have the most evolutionary significance for earlier generations but still holds important relevance for individuals who are farmers, botanists, scientists, and other professions that closely interact with the natural world.

Limitations and Critiques of the Multiple Intelligences Theory

Although Gardner’s theory of multiple intelligences is accepted by many educators, it has faced a significant amount of criticism among researchers. Much of the criticism surrounding this theory has been from the proponents of the prevailing intelligence theory, that of a general intelligence with strengths and weaknesses within that one construct. Existing intelligence tests such as the Wechsler Adult Intelligence Scale–Fourth Edition have not provided empirical support for Gardner’s theory, and correlational research has suggested that different modalities of intelligence are highly correlated with one another, rather than having low correlations as Gardner’s multiple intelligences theory would indicate. High intercorrelations among various aspects supports the mainstream, accepted theory that intelligence is one composite score made up of many factors, but all related to one another as one complex construct. Gardner’s theory would support a definition of intelligence that would be more splintered and separate, rather than one cohesive unit.

The lack of empirical evidence and the evidence suggesting correlational relationships makes it difficult for the scientific community to accept Gardner’s theory. Additionally, the way in which Gardner defines intelligence and the construct of intelligence makes it very difficult to measure each of the eight modalities that he describes. Individuals’ strengths in many of the areas are measured through subjective reasoning and judgment, so that these measurements lack the validity and reliability necessary for a scientific theory. Finally, since the definition of intelligence is so vague and the areas are so subjective, there are opponents to the theory who suggest that this cannot be considered a theory, as it is unfalsifiable.

Kimberly Ethridge and Anthony Odland

See also Cattell–Horn–Carroll Theory of Intelligence; Emotional Intelligence; *g* Theory of Intelligence; Intelligence Quotient; Intelligence Tests; Wechsler Intelligence Scales

Further Readings

Campbell, L., Campbell, B., & Dickinson, D. (1999). Through multiple intelligences. Needham Heights, MA: Allyn & Bacon.

Chen, J. Q. (2004). Theory of multiple intelligences: Is it a scientific theory? Teachers College Record, 106(1), 17–23.

Gardner, H. (2011). Frames of mind: The theory of multiple intelligences. New York, NY: Basic Books.

Gardner, H., & Hatch, T. (1989). Educational implications of the theory of multiple intelligences. Educational Researcher, 18(8), 4–10.

Gray, J. H., & Viens, J. T. (1994). The theory of multiple intelligences: Understanding cognitive diversity in school. National Forum: Phi Kappa Phi Journal, 74(1), 22–25.

Morgan, H. (1996). An analysis of Gardner's theory of multiple intelligence. Roeper Review, 18(4), 263–269.

Wechsler, D. (2014). Wechsler adult intelligence scale–Fourth Edition (WAIS–IV). New York, NY: Pearson.

Julius Sim Julius Sim Sim, Julius

Multiple Linear Regression Multiple linear regression

1110

1113

Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression in which values on an outcome (*Y*) variable are predicted from two or more predictor (*X*) variables. There are three principal objectives of multiple linear regression: (1) to obtain specific predicted values on *Y* corresponding to specific observed values on the *X* variables; (2) to determine how well a predetermined set of *X* variables predict values on *Y* (i.e., to gauge the predictive strength of this set of predictors, taken together); and (3) to select from a group of *X* variables a subset that are the “best” predictors of *Y*. This entry reviews the form of the multiple regression model, assumptions of the analysis, and how to go about selecting and validating a model.

Form of the Multiple Regression Model

The form of the regression model, in the case where there are, for example, three predictor variables, is given by the following equation:

$$\widehat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

Here, is the predicted value of *Y*, α is the intercept, and β is the slope coefficient. The intercept is a constant that represents the predicted value of *Y* when each of the *X* variables has the value 0 (this parameter is not normally of substantive interest). The slope coefficient, which may be positive or negative, is the change in the predicted value of *Y* for a 1-unit increase in the *X* variable concerned. Alternatively, the equation can be represented as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Here, Y is the observed value of the outcome variable and ε is the residual—the difference between the observed and the predicted value of the outcome variable ($Y - \hat{Y}$). The residuals will reflect measurement error and the influence of all potential predictors of Y not included in the model.

As an example of a multiple regression model, assume that students' examination scores (on a 0–100 scale) are to be predicted from a scale representing their attitudes to schooling (0–30, higher scores more positive), their age (in months), and their sex (0 = *male*, 1 = *female*). The intercept for this model is 131.85 and the regression coefficients for β_1, β_2, and β_3 (the attitude scale, age, and sex, respectively) are 0.23, –0.45, and 0.19. For a female student aged 126 months with an attitude score of 22, the predicted exam score would therefore be $131.85 + (0.23 \times 22) + (-0.45 \times 126) + (0.19 \times 1) = 80.40$. When there is more than one predictor variable in a regression model, each slope coefficient is adjusted for the others; hence, if age is removed from the model, the coefficient for the attitude scale changes to 0.26 and that for sex to –0.07. The regression coefficient in multiple regression is therefore not simply "the change in $\hat{Y}$ for a 1-unit change in X," but "the change in $\hat{Y}$ for a 1-unit change in X, with the other X variables held constant."

Because regression coefficients are often expressed in terms of different scales, they cannot be compared for their magnitude. So, a coefficient given in terms of points on a 0–30 scale cannot meaningfully be compared with one given in terms of months; this becomes clear when we consider that if age had been given in years, the coefficient would increase 12-fold, but its predictive strength would be the same. However, statistical software output normally includes standardized coefficients, which are expressed in standard deviation units; the coefficient represents the change in Ŷ in standard deviation units for a 1 standard deviation increase in X. As these standardized coefficients are on the same scale, their relative magnitude can be assessed within a model (though their comparison *across* models is not recommended).

Assumptions of the Analysis

The assumptions of simple (bivariate) linear regression apply equally to multiple linear regression: 1. The predictive relationship between each X and Y is linear. The crucial consideration here is linearity within the model, with the other X variables. This is most effectively assessed by a special form of scatterplot called a partial plot.

2. The level of measurement of *Y* is interval or ratio and that of *X* is either interval, ratio, or binary. *X* variables that are ordinal or are nominal with more than two levels must first be converted to binary dummy variables. If the *Y* variable is ordinal rather than interval or ratio, an ordinal regression model should be used.

3. The *X* variables are fixed and measured without error. An *X* variable that is random can normally be used provided that Assumption 7 is satisfied. As measurement error in the *X* variables leads to biased coefficients (coefficients are underestimated in bivariate regression but may be either under-or overestimated in multiple regression), it should where possible be minimized.

4. The residuals are independent (i.e., the value of one residual does not influence, and is not influenced by, the value of any other residual).

5. The residuals have homogeneity of variance (homoscedasticity).

6. The residuals are (approximately) normally distributed—this assumption, which only applies to the residuals, not to *Y* or the *X*s, is required for hypothesis tests and confidence intervals. The normality assumption becomes less stringent as sample size increases.

7. The residuals are not correlated with the *X*s. If they are correlated, this may indicate that the model has not been correctly specified, and regression coefficients may be biased (omitted variable bias). The model being correctly specified means that all relevant *X* variables have been included.

Collinearity

An additional consideration in multiple regression is (multi)collinearity. This occurs if two or more predictors are highly correlated, such that they are each explaining much the same variance in the outcome variable, making it hard to obtain separate estimates for these predictors. The consequences are that the overall model may be statistically significant, but paradoxically none of the individual regression coefficients are, owing to inflated standard errors, or that coefficients may have the “wrong” sign (an *X* that should theoretically have a positive relationship with *Y* has a negative coefficient or vice versa). Collinearity does not, however, bias the regression coefficients. High pairwise correlations (greater than ±0.80 is sometimes suggested as a criterion) among the *X* variables provide an initial indication of collinearity, but it is important to test for it more fully within the model. As well as looking for the consequences of collinearity just outlined, one can generate specific statistics that help diagnose collinearity such as the variance inflation factor.

Various steps can be taken to address collinearity, if it is found: An increase in sample size will reduce standard errors; one or more of the collinear variables could be omitted, if this does not adversely affect the information provided by the model; two or more collinear variables could be combined into a smaller number of variables such as through principal components analysis; collinear variables can be subjected to joint, rather than individual, hypothesis tests; and one could use a special form of regression—ridge regression—that produces more stable, albeit biased, coefficients.

Selecting a Model

Sometimes, the predictors to be used in the regression model have been predetermined, particularly if the goal is to assess the predictive power of these variables collectively. In such a case, the predictors are often simply entered together. Their predictive power can be assessed by looking at the goodness of fit of the model, as indicated by the R^2 statistic; this is a squared multiple correlation coefficient that expresses, on a 0–1 scale, the proportion of variance in the outcome variable that is explained by the predictors in the model. As the number of predictors in the model will in itself affect its predictive power, an adjusted R^2 that takes into account the number of predictors and the sample size is often preferred over the crude R^2 statistic. An alternative to entering all the predetermined *X* variables at once is to enter them in blocks (hierarchical entry). This allows the goodness of fit of the model to be assessed incrementally, such that the additional contribution of a new predictor, or set of predictors, within each block can be assessed in terms of the change produced in R^2.

Another objective that was identified for multiple regression was that of determining a “best” set of predictors, usually interpreted in terms of those predictors that are statistically significant. There are three principal ways in which such a model can be identified. First, all candidate predictors can be entered and the computer can be asked to perform *backward selection*. This is an iterative process whereby the variable with the smallest partial correlation with the outcome variable is removed (provided the diminution in model fit is not statistically significant), the model is refitted, and a variable is again removed in the same way. Once a variable has been removed, it remains excluded from the model. If at any stage no variables within the model meet the criterion for elimination, the analysis is terminated. The final model will consist of those variables that have made a significant additional contribution to the model. The second alternative is *forward selection*, which is essentially the reverse of backward selection. Here, the initial model just contains the intercept. At the first step, the variable with the greatest zero-order correlation with the outcome variable is entered (provided that this correlation is statistically significant). At each successive step, the variable that has the greatest partial correlation with the outcome variable is added (provided that the contribution to model fit is statistically significant). Once a variable has entered the model, it stays in. If at any stage no variables outside the model meet the criterion for entry, the analysis is terminated.

Finally, there is *stepwise selection*. This procedure moves variables in and out of the model according to their associated partial correlations. At any stage, either the variable with the largest partial correlation that is not already in the model is moved in (provided that its *p* value is below a threshold) or the variable with the lowest partial correlation already in the model is moved out (provided that its *p* value is above a threshold, which must be higher than that for moving variables in). When there are no variables in the model that are eligible for exclusion and no variables outside it eligible for inclusion, this is the final model. Variables can move in or out of the model at any stage, though only one variable is moved in or out at each stage.

These automated methods of variable selection are often criticized for allowing a model to be constructed on purely statistical grounds and also for the use of multiple hypothesis tests. Furthermore, there are situations in which they are inappropriate: If a categorical predictor has been transformed to two or more dummy variables, it would not be appropriate to allow one to be included in the model and another not; similarly, if a polynomial model has been used, lower order terms should not be separated from higher order terms for the predictor(s) concerned; and if a predictor has been identified as an important control variable, it should be included regardless of its statistical significance. A way of accommodating these concerns when using automated variable selection is to use hierarchical entry. This would involve placing predictors that must be retained in the model, irrespective of statistical significance, in the first block, and then creating a second block containing the predictors that are to be subjected to automated selection.

The overall principle underlining model selection is one of the compromise between goodness of fit and parsimony; the resulting model should include sufficient predictors to ensure a good fit, but not so many as to make the model unwieldy and difficult to interpret in practical situations.

Validating a Model

When a regression model is estimated from a particular sample, it will be the “best” model for that particular sample—the regression coefficients will have the optimum predictive power for the sample from which they were estimated. However, they may not have the same predictive power in a different sample from the same population of interest. Cross-validation is a means of examining this issue. It involves applying the coefficients estimated in the original (screening or training) sample to a new (calibration or validation) sample, so as to produce predicted values in this second sample. The squared correlation of these predicted values and the observed values in the calibration sample will give an R^2 value. Subtracting this R^2 from the R^2 obtained in the original screening sample will give the cross-validation “shrinkage”—the degree to which the fit of the model is lower when it is applied to another sample. The smaller the shrinkage, the more reliably the regression coefficients can be used for prediction in other samples.

Julius Sim

See also Goodness-of-Fit Tests; Hierarchical Regression; Multicollinearity; Multiple Linear Regression; Partial Correlations; Residuals; Simple Linear Regression; Stepwise Regression

Further Readings

Aiken, L. S., West, S. G., Pitts, S. C., Baraldi, A. N., & Wurpts, I. C. (2013). Multiple linear regression. In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology. Volume 2: Research methods in psychology (2nd ed.), pp. 509–542). Hoboken, NJ: Wiley.

Allison, P. D. (1999). Multiple regression: A primer. Thousand Oaks, CA: Pine Forge Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Kahane, L. H. (2008). Regression basics (2nd ed.). Thousand Oaks, CA: SAGE.

Lewis-Beck, M. S. (Ed.). (1993). Regression analysis. London, UK: SAGE.

Mendenhall, W., & Sincich, T. (2014). A second course in statistics: Regression analysis (7th ed.). Harlow, UK: Pearson.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (4th ed.). New York, NY: Wiley.

Pedhazur, E. J. (1997). Multiple regression in behavioral research (3rd ed.). Melbourne, Australia: Wadsworth.

Laura M. B. Kramer Laura M. B. Kramer Kramer, Laura M. B.

Multiple-Choice Items

Multiple-choice items

1113

1116

Multiple-Choice Items

The proliferation of multiple-choice assessments in educational measurement is likely due to the relative ease, objectivity, and cost-efficiency of scoring, particularly when assessment is conducted on a large scale, with a short timeline for reporting scores, or on a tight budget. Although assessment has been occurring for millennia, the multiple-choice question is a relative newcomer; its first large-scale use is generally regarded as the Army Alpha, an aptitude test used to screen military recruits in World War I and assign them to military jobs. Multiple-choice items are common in educational assessments used to meet the requirements of the Elementary and Secondary Education Act, originally passed in 1965, which requires assessment every year in multiple subjects for students from Grade 3 to high school. Multiple-choice tests can be scored quickly and inexpensively in order to produce timely reporting as required by the Elementary and Secondary Education Act. After providing an expansive review of the makeup of traditional multiple-choice items, this entry examines how to construct and then score multiple-choice questions, highlights their advantages and disadvantages, and reveals some nontraditional multiple-choice item types.

Traditional Multiple-Choice Items

Multiple-choice questions belong in a larger item category sometimes referred to as “selected response.” The hallmark of a selected-response item type is that the examinees must choose their answer from a provided list of possible answers, as opposed to generating an answer on their own (constructed response) or carrying out an activity (performance task). Selected-response items can have as few as two answer choices (e.g., a true/false question) or many (e.g., an item utilizing a word bank from which the examinee chooses the correct word from the bank to complete a sentence, label a diagram). The format with which most people are familiar is the four-option multiple-choice question, such as those encountered in the SAT or ACT standardized tests.

The traditional multiple-choice item consists of a stem and answer choices. The stem sets up the problem to be addressed and asks the question, and it may contain stimulus material, such as a graph or a text that the examinee must use to respond. The answer choices are plausible responses to the question posed in the stem. In a traditional multiple-choice item, there is only one correct answer, called the key; the incorrect responses are called distractors.

Answer choices in a four-option multiple-choice item are most commonly labeled A, B, C, and D to correspond to the bubbles on a scannable answer document (bubble sheet). Longer assessments may alternate a series of letters for each question to help prevent students from responding on the incorrect line of a scannable answer document. Odd-numbered questions would be labeled A, B, C, and D; even-numbered questions are then labeled E, F, G, and H. With computer-delivered assessments becoming more commonplace, answer choices presented on-screen may not be labeled for the examinee, but labeling is still a useful conceptual tool when developing a complete test form.

Construction

Questions used in large-scale assessment programs, especially those with high stakes attached, go through a lengthy development process, including content review by subject matter experts, editorial review, and bias review. Although a classroom assessment may not need to stand up to the same level of scrutiny, care should be taken to make sure that test questions are fair and are measuring what they are intended to measure.

The first step in creating a multiple-choice test question is to clearly define what is to be assessed. An evidence-centered design approach is often used to guide item developers in identifying the important content, differentiating between the performance of a master and a nonmaster of that content, and devising questions that will elicit a response from the examinees that will provide evidence of their mastery of the content.

Stems should generally be presented in the form of a question, rather than a fill-in-the-blank format; the examinee should be able to know what is being asked and be able to formulate a response as if the answer choices were not present. The answer choices should be plausible, and whenever possible, distractors should be based on the common errors or misconceptions of nonmasters of the content. The use of common error distractors, sometimes also known as diagnostic foils, can provide educators and policy makers with information about overall areas for additional instruction or professional development based on the types of errors or misconceptions commonly held by students.

The answer choices should be parallel in construction, to the extent possible; for example, all four answer choices are complete sentences or all fractions. It should not be that three of the responses are sentences and the last is a single word or that three of the responses are fractions and the last is a decimal. If all four answer choices being parallel is not possible, then two sentences and two single words, or two fractions and two decimals, would be appropriate. Having one answer choice that is noticeably different from the others in structure or format is to be avoided.

Also to be avoided is any type of clueing or “clang.” The most common type of clueing is when a word in the stem is repeated in only the correct answer choice. More subtle clueing can come from nonparallel answer choices, such as the correct answer being a verb and the distractors all being nouns, or the correct answer is a range of values and the distractors are all single numbers. Clueing can also happen across multiple items, so when constructing a test, care should be taken that an earlier question does not provide an answer to a later question in the same test form.

Scoring

Traditional multiple-choice questions are scored such that a correct answer earns the examinee one point, and an incorrect response earns no points. Unanswered questions are treated as incorrect and earn the examinee no points. At one time, the SAT included a penalty for guessing incorrectly by subtracting one-fourth of a point for an incorrect answer, but that is no longer the case.

Single items or groups of items measuring a domain of particular importance can be emphasized in scoring by weighting them to be worth more than one point each. Nontraditional multiple-choice questions may be worth multiple points or offer partial credit, depending on the content being measured and the data analysis model being used.

The total number of points earned on the entire multiple-choice test is a raw score, which can be converted to a variety of other scores such as a percent correct, percentile rank, or scaled score using either classical or item response theory methods.

Advantages

Objectivity

Because test takers are given a finite list of possible responses, and the test makers should have gone to great lengths to ensure that there is in fact only one correct answer, scoring is completely objective. If the key is “C,” then test takers who chose “C” are correct and test takers who chose any other response are incorrect. This removes any judgment or bias in scoring.

Efficient Scoring

For small-scale assessment applications, such as a classroom assessment, multiple-choice items can be quickly graded by a teacher using a paper answer key. Large-scale assessments, such as statewide testing programs to meet the requirements of the Elementary and Secondary Education Act, require rapid scoring of a few hundred thousand tests for a small state quickly rising to millions of answer documents for a medium-sized state. Scannable answer documents, as the name implies, can be fed into an optical scanning machine. Simple scanners that are marketed to schools and districts can scan and score 60 or more answer documents a minute. Major testing companies have more sophisticated scanners that can score as many as 250 answer documents a minute.

Ease of Construction

There are a limited number of ways a question can be asked. For example, in a reading comprehension test, “What is the main idea?” is a question that can be applied to just about any literary text. In a math test, the situations and scenarios can change, but there are often some common elements that can be changed to create a “clone” of the item. For example, “A factory produces widgets at the rate of 15 widgets per hour. How long will it take for the factory to produce 300 widgets?,” has two numeric elements that can be varied. The rate of widget production and the final number of widgets needed can be changed to create a variety of similar items. Even the surrounding scenario of a factory setting and widgets could be changed to a bakery making cakes. Some researchers have investigated automated cloning of items, including predicting the psychometric properties of the cloned item based on the properties of the parent item.

Concerns

The main criticism of multiple-choice tests is that they can only be used to measure how well an examinee has memorized discrete facts or rote information and can recognize the correct answer. Although poorly constructed tests of only low-level recall questions have given multiple-choice tests a bad reputation, the ability to quickly assess factual knowledge is actually one of the benefits of a multiple-choice test. The criticism that some procedural knowledge and actual performance of certain tasks cannot be measured adequately by a multiple-choice test is a fair one. Although an assessment for a welding class could include multiple-choice questions about the methods, types of materials, and kinds of energy sources used in welding, the real test is whether a student can successfully complete welds on different materials using the appropriate processes and techniques. However, a multiple-choice test evaluating the examinee’s factual knowledge and safety techniques could be a worthy precursor to the actual performance of the welds in order to reduce wasted material and accidents.

Deeper levels of thinking, reasoning, analyzing, and problem solving can be measured in a multiple-choice test; however, these questions generally require more information in the stem and stimulus materials, longer and more descriptive answer choices, or both. Lengthy items increase the reading burden for examinees, which can threaten validity. For example, a very wordy math test may disadvantage mathematically proficient students who lack the verbal skills to decode the text in a math problem.

More complex multiple-choice questions that measure more in-depth knowledge take more time and skill to develop and lend themselves less to automated item generation. However, although they take more time to create and more time for test takers to respond to, they still permit objective, efficient machine scoring.

Nontraditional Multiple-Choice Item Types

There are other selected-response item types that may have more than one correct answer or answers with varying degrees of “correctness.”

Multiselect Multiple-Choice Item Type

This multiple-choice item type has more than one correct answer and often has more than the typical number of answer choices. Items may be identified with a phrase such as “choose all that apply” in the stem. Scoring may be all-or-none, partial credit may be awarded for correctly selecting some of the correct answers, or points can be taken away for selecting too many of the distractors. The choice of scoring model should be dictated by the content. For example, in an assessment of medical knowledge, choosing a procedure or medication that would kill the patient might be more heavily penalized than choosing an expensive but not diagnostically useful procedure.

Situational Judgment Task

Although this item type is most prevalent in workforce assessments, it is making inroads in educational measurement for measuring “21st-century” skills such as problem solving, leadership, and teamwork. In this multiple-choice item type, all of the answers are correct but to varying degrees. The stem provides a scenario and poses a problem, and the answer choices are possible resolutions to the problem. The answer selected by the examinees can provide insight into the depth and sophistication of their level of knowledge or ability. Development of situational judgment tasks requires an additional adjudication step in which a panel of experts provides judgments on the relative desirability of the proposed solutions used as answer choices. The expert rankings then inform the scoring, with more points being awarded for the most desirable response.

Laura M. B. Kramer

See also Constructed-Response Items; Evidence-Centered Design; Performance-Based Assessment; Scales; Score Reporting

Further Readi

Further Readings

Gierl, M. J., & Haladyna, T. M. (2013). Automatic item generation: Theory and practice. New York, NY: Routledge.

Haladyna, T. M. (2004). Developing and validating multiple-choice test items. New York, NY: Routledge.

Weekley, J. A., & Ployhart, R. E. (Eds.). (2006). Situational judgment tests: Theory, measurement, and application. Mahwah, NJ: Erlbaum.

Yi-Fang Wu Yi-Fang Wu Wu, Yi-Fang

Multitrait–Multimethod Matrix Multitrait–multimethod matrix

1117

1119

Multitrait–Multimethod Matrix

The term *multitrait–multimethod matrix* refers to a practical approach to evaluate construct validity evidence of an intended measure based on relations among a set of measures. This entry first introduces a brief history of the multitrait–multimethod matrix and describes the purpose of the approach. It then describes how the various elements in the matrix support the construct validity of the intended measure. The entry provides an example of multitrait–multimethod matrix in achievement testing and concludes with pros and cons of this approach.

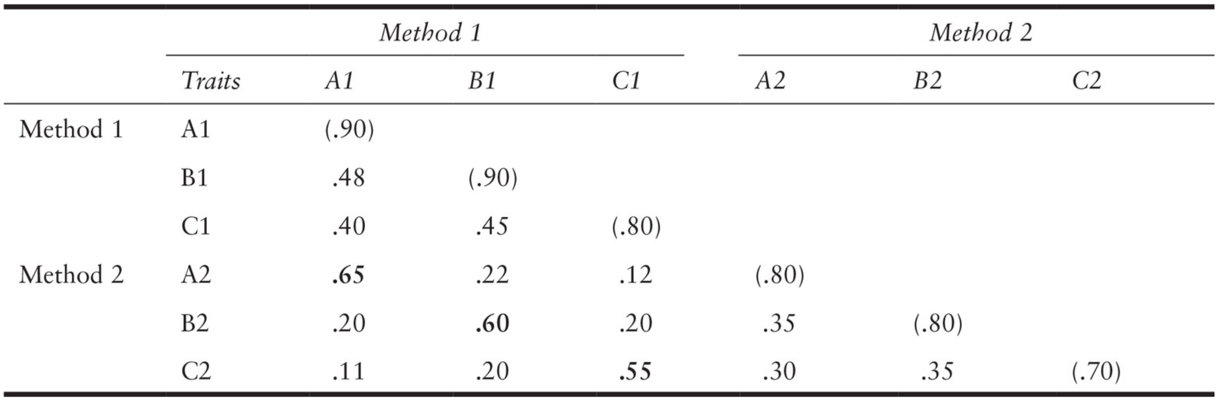
Brief History and Introduction

The multitrait–multimethod matrix was established in 1959 by Donald T. Campbell from the Graduate School of Northwestern University and Donald W. Fiske from the Department of Psychology at the University of Chicago. A multitrait–multimethod matrix is used for validation, the process involving accumulating relevant evidence in order to justify measures, to provide a sound scientific basis for interpretations and uses of the results (e.g., test scores), and to establish evidence for construct validity. In this entry, traits and constructs are used interchangeably.

Specifically, a multitrait–multimethod matrix aims to provide convergent and discriminant evidence that demonstrates the relationships of the measure of interest to other measures. Convergent evidence is the degree to which the measure of interest and other measures are intended to assess the same or similar constructs. Discriminant evidence is the degree to which the measures of different constructs are not related in reality. Convergent and discriminant evidence are both required to demonstrate that the measure of interest supports the construct underlying the proposed interpretations and/or uses of the measure. They can both be assessed using a multitrait–multimethod matrix.

Matrix Elements and Validity Requirements

A multitrait–multimethod matrix consists of the correlations when each of several traits is measured by each of several methods. Table 1 presents a synthetic example of the multitrait–multimethod matrix for three traits (i.e., A, B, and C) measured by two methods (i.e., Method 1 and Method 2). In the table, A1 denotes Trait A measured by Method 1, B1 denotes Trait B measured by Method 1, and so on. The matrix has four blocks. The *monomethod blocks* are at the top left and bottom right, each of which consists of the reliability diagonal with values enclosed by parentheses and the adjacent heterotrait–monomethod triangle enclosed by solid lines. The bottom left is the heteromethod block, which has the validity diagonal with values in bold face, and two heterotrait–heteromethod triangles bordered with the dashed lines. Note that the two heterotrait–heteromethod triangles are not identical.



		Method 1			*Method 2*		
	Traits	*A1*	*B1*	*C1*	*A2*	*B2*	*C2*
Method 1	A1	(.90)					
	B1	.48	(.90)				
	C1	.40	.45	(.80)			
Method 2	A2	**.65**	.22	.12	(.80)		
	B2	.20	**.60**	.20	.35	(.80)	
	C2	.11	.20	**.55**	.30	.35	(.70)

A few observations need to be addressed based on this table. First, reliability is the agreement between two measures for the same trait through similar or same methods. The reliabilities could also be viewed as the monotrait–monomethod values. In this synthetic scenario, for example, the reliabilities for the three traits are higher for Method 1 and lower for Method 2.

Second, validity is represented by agreement between two attempts to measure the same trait through different methods. Convergence evidence across methods for each of the three traits can be evaluated by the validity diagonal resulting from the monotrait–heteromethod. Each entry is the correlation between measures of a single trait using the two different methods. As Campbell and

Fiske pointed out, the entries in the validity diagonal should be significantly different from zero and sufficiently large.

[a] Heterotrait–monomethod triangles are enclosed by solid lines; heterotrait–heteromethod triangles are enclosed by dashed lines; monotraits–monomethod coefficients displayed in parentheses are "reliability diagonals"; and monotraits–heteromethod coefficients shown in bold face are "validity diagonals."

Third, a validity diagonal entry should be higher than the entries in its column and row in the heterotrait–heteromethod triangles. In other words, a convergent validity entry for a trait measure should be higher than the correlations obtained between that measure and any other measure having neither trait nor method in common. For example, the correlation between Trait A measured by Method 1 and Trait A measured by Method 2, $r_{A1,A2}$ (.65), is larger than $r_{A1,B2}$ (.20), $r_{A1,C2}$ (.11), $r_{A2,B1}$ (.22), and $r_{A2,C1}$ (.12).

Fourth, different traits are supposed to be distinct, which provides discriminant validity evidence. A trait measure tends to correlate higher with an independent measure that assesses the same trait than with measures designed to assess different traits that use the same method. In the table, for a specific trait, the values in the validity diagonal are supposed to be larger than their associated entries in the heterotrait–monomethod triangles. That is, $r_{A1,A2} > r_{A1,B1}$ (.65 > .48) and $r_{A1,A2} > r_{A1,C1}$ (.65 > .40) for Trait A, $r_{B1,B2} > r_{B1,A1}$ (.60 > .48) and $r_{B1,B2} > r_{B1,C1}$ (.60 > .45) for Trait B, and $r_{C1,C2} > r_{C1,A1}$ (.55 > .40) and $r_{C1,C2} > r_{C1,B1}$ (.55 > .45) for Trait C.

Fifth, correlations between measures of different traits should not be very high. In the example, the entries in the heterotrait–monomethod triangles are quite low (ranging from .30 to .48 across methods), and those in the heterotrait–heteromethod triangles are the lowest (ranging from .11 to .22). Note that the last three observations in the example provide evidence for discriminant validity. Finally, it is desirable that the same pattern of relationship between traits is shown in all of the heterotrait triangles of both the monomethod and heteromethod blocks.

The validity requirements just described are often referred to in the literature as the Campbell-Fiske criteria. Several methods have been proposed to analyze a multitrait–multimethod matrix, using various statistical analysis techniques such as smallest space analysis, path analysis, a nonparametric approach, an analysis as smallest space analysis, path analysis, a nonparametric approach, an analysis of variance approach, and a confirmatory factor analysis approach. Some aim to quantify the variance accounted for by traits and methods while the others focus on the tests of significance of trait and method variance. A comprehensive evaluation of different methods has been presented by Neal Schmitt and Daniel M. Stults. Although the confirmatory factor analysis approach is relatively popular for analyzing the multitrait–multimethod matrix, researchers often encounter empirical problems such as model nonidentifiability and unattainable parameter estimates. Readers can refer to the work by David A. Kenny and Deborah A. Kashy for a detailed discussion regarding these potential problems.

An Example in Educational Testing

At times, the multitrait–multimethod matrix is applied in K–12 achievement testing. To validate test score interpretations of the American College Testing Program (ACT) Aspire achievement tests, for example, data from a sample of students who took both statewide achievement tests and the ACT Aspire tests were used to create a multitrait–multimethod matrix. The statewide tests and the ACT Aspire tests were built under different test specifications, but both aimed to measure student achievement in mathematics, reading, and science. In this example, the traits being measured are achievement in mathematics, reading, and science, and the two tests can be viewed as the methods. Interested readers can see further readings at the end of this entry.

Pros and Cons

The use of a multitrait–multimethod matrix is supposed to be informed by the theory of the constructs under investigation. It is effective in providing reasonable standards for collecting convergent and discriminant evidence that can demonstrate that the measure(s) of interest support the construct underlying the proposed interpretations. By studying correlations in the matrix, researchers can gather information concerning how well the measure(s) have been related to or distinguished from other measures. Also, the Campbell-Fiske criteria are easy to understand and follow since only correlations but no complex statistics are involved. However, the use of a multitrait–multimethod matrix requires a data collection design that often takes substantial time and resources in practice.

Yi-Fang Wu

See also Construct-Related Validity Evidence; Validity

Further Readings

American College Testing Program, Inc. (2016). The ACT aspire summative technical manual. Retrieved from http://actaspire.avocet.pearson.com/actaspire/home#15904

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed.), pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin, 112(1), 165–172.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.), pp. 13–103). New York, NY: American Council on Education and Macmillan. ISBN-13: 978-1573562218; ISBN-10: 1573562211

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10(1), 1–22.

David W. Stockburger David W. Stockburger Stockburger, David W.

Multivariate Analysis of Variance

Multivariate analysis of variance

1119

1127

Multivariate Analysis of Variance

Multivariate analysis of variance (MANOVA) is an extension of univariate analysis of variance (ANOVA) in which the independent variable is some combination of group membership but there is more than one dependent variable. MANOVA is often used either when the researcher has correlated dependent variables or, instead of a repeated-measures ANOVA, to avoid the sphericity assumption. While MANOVA has the advantage of providing a single, more powerful test of multiple dependent variables, it can be difficult to interpret the results.

For example, a researcher might have a large data set of information from a high school about their former students. Each student can be described using a combination of two factors: gender (male or female) and whether they graduated from high school (yes or no). The researcher wishes to analyze and make decisions about the statistical significance of the main effects and interaction of the factors using a simultaneous combination of interval predictor variables such as GPA, attendance, degree of participation in various extracurricular activities (e.g., band, athletics), weekly amount of screen time, and family income.

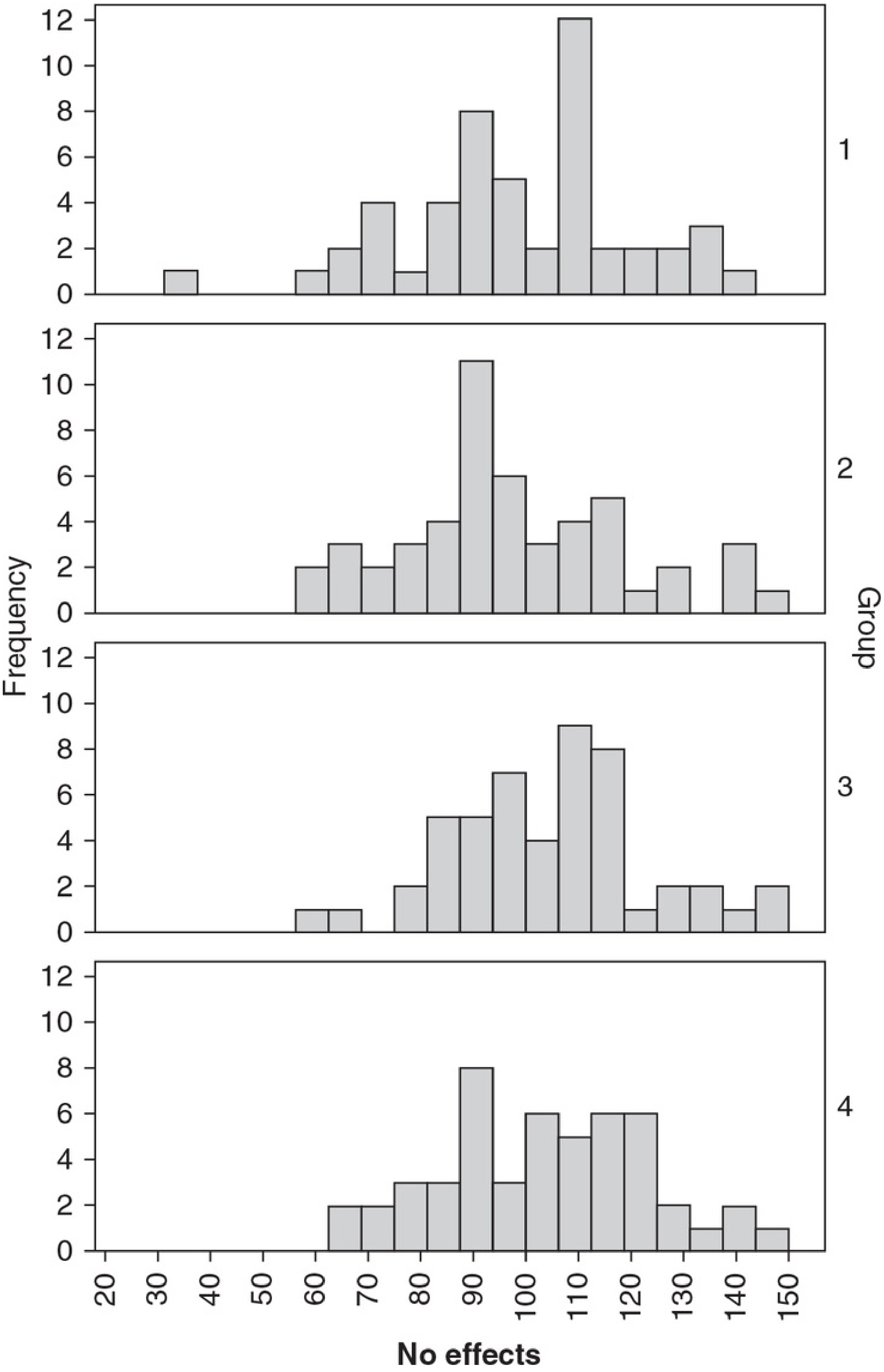
Put in a broader statistical context, MANOVA is a special case of canonical correlation and is closely related to discriminant function analysis (DFA). DFA predicts group membership based on multiple interval measures and can be used after a MANOVA to assist in the interpretation of the results.

This entry explains MANOVA by first reviewing the underlying theory of univariate ANOVA and then demonstrating how MANOVA extends ANOVA by using the simplest case of two dependent variables. After the rationale of the analysis is understood, it can be extended to more than two dependent variables but is difficult to present visually. In that case, matrix algebra provides a shorthand method of mathematically presenting the analysis.

Univariate ANOVA

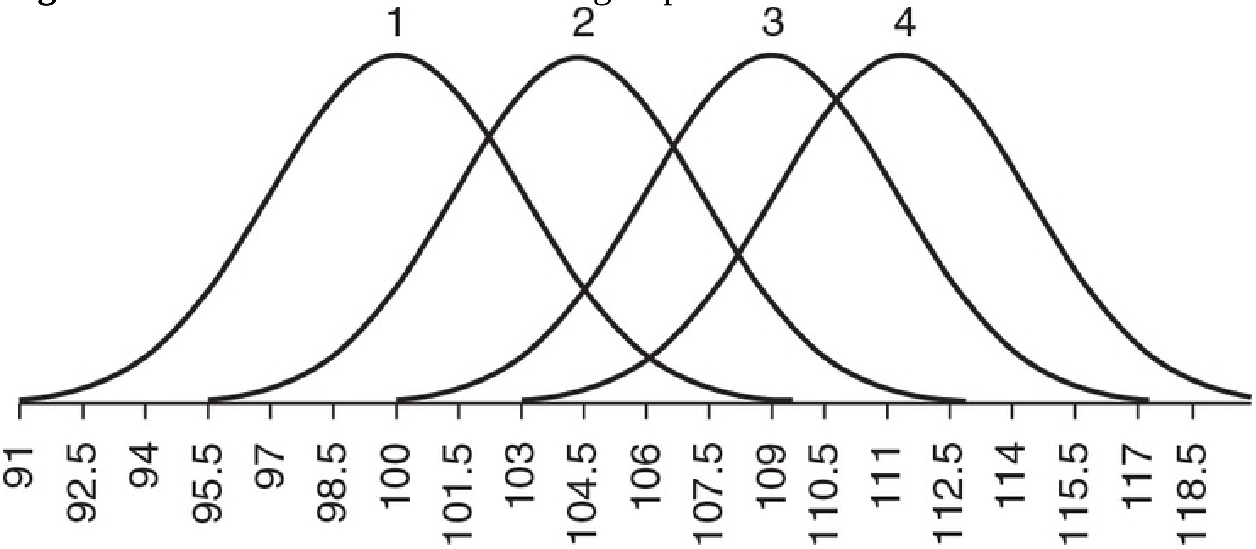
In univariate ANOVA, the independent variable is some combination of group membership and a single interval-dependent variable. The data can be visualized as separate histograms for each group, as seen in Figure 1, with four groups of 20 observations each.

Figure 1 Histogram of four groups


Frequency
Group
1
2
3
4
0
2
4
6
8
10
12
20
30
40
50
60
70
80
90
100
110
120
130
140
150
No effects

The ratio of the variability between the means of the groups relative to the variability within groups is fundamental to ANOVA. This is done by modeling the sampling distribution of each group with a normal curve model, assuming that both the separate sample means estimate μ and σ is equal in all groups and estimated by a formula using a weighted mean of the sample variances. The assumption of identical within-group variability is called the *homogeneity of variance* assumption. The model of the previous data is illustrated in Figure 2.

Figure 2 Normal curve model of four groups



From this model, two estimates of σ^2 are computed. The first, mean square between (MS_B), uses the variability of the means, and the second, mean square within (MS_W), uses the estimate of combined variability within the groups. A computed statistic, called *F*, is the ratio of the two variance estimates:

$$F = \mathrm{MS_B} \;/\; \mathrm{MS_W}\,.$$

The distribution of the *F* statistic is known, given the assumptions of the model are correct. If the computed *F* ratio is large relative to what would be expected by chance, then real effects can be inferred; that is, the means of the groups are significantly different from each other. The between variability can be partitioned using contrasts to account for the structure of group membership, with separate main effects, interactions, and nested main effects, among others, being tested using the ANOVA procedure.

MANOVA

MANOVA and ANOVA have similar independent variables, but in MANOVA there are two or more dependent variables. Although the computations involved in MANOVA are much more complicated and best understood using matrix operations, the basic concept is similar to the univariate case. This will be illustrated by first examining one of the simpler cases of MANOVA, with four groups and two dependent variables. The extension to more groups and dependent variables, while not illustrated, can be inferred from this case.

Four Groups and Two Dependent Variables

The data for four groups and two dependent variables can be illustrated using a scatterplot (see Figure 3). The paired means for each group are called *centroids*, and in matrix algebra terminology together they constitute a vector of length equal to the number of groups. Three of the four standard statistics used in hypothesis testing in MANOVA compare the variability of the centroids to the within-group variability. To do this, they model the dependent variables with a multivariate normal distribution. In a multivariate normal distribution, all univariate distributions will be normal, but having univariate normal distributions for all variables does not guarantee a multivariate normal distribution. In addition, all groups are assumed to have similar variance/covariance matrices, which corresponds to the homogeneity of variance assumption in univariate ANOVA. The bivariate normal model of the sampling distribution of data shown in Figure 3 is presented in Figure 4.

Figure 3 Scatterplot of four groups

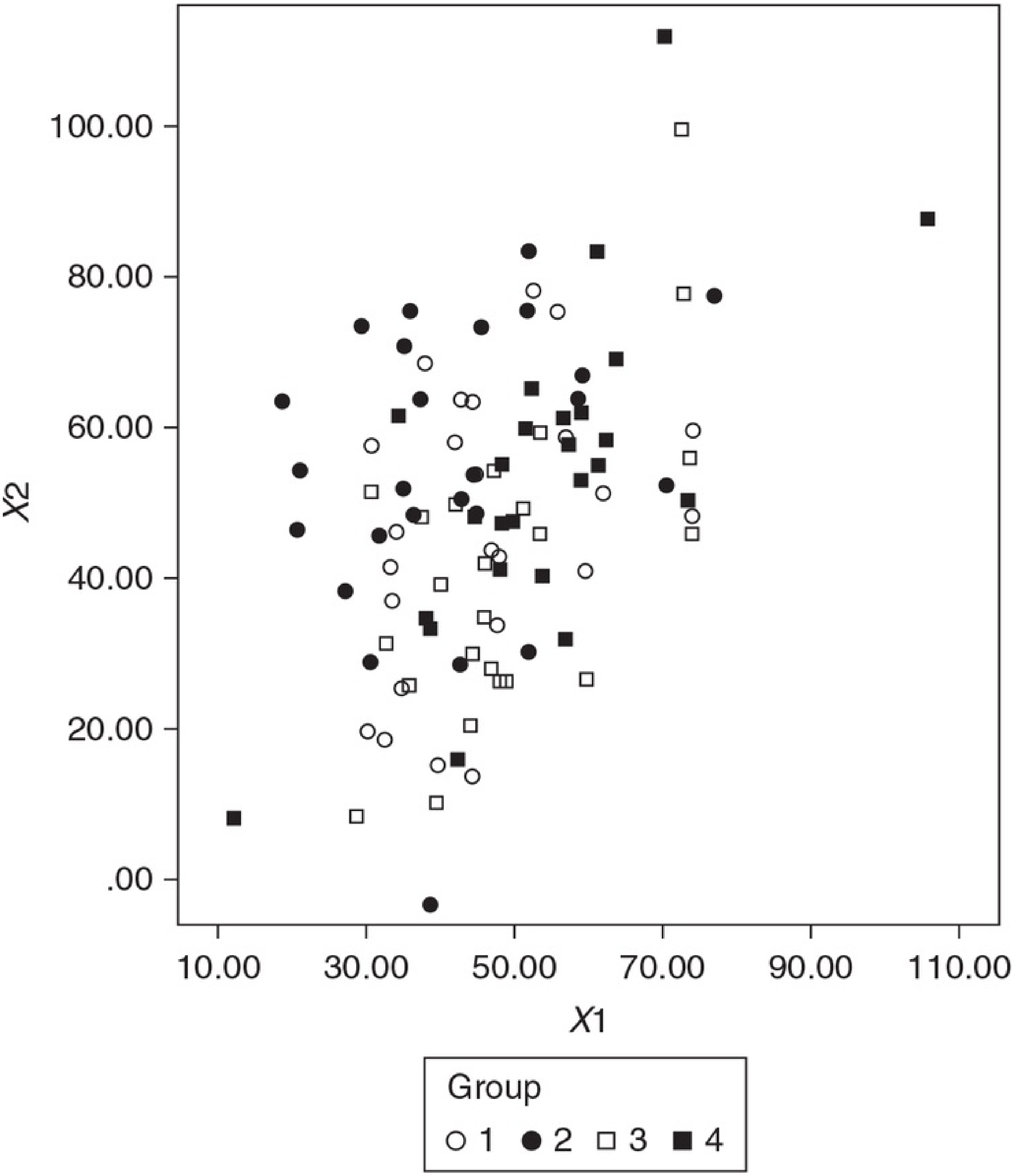
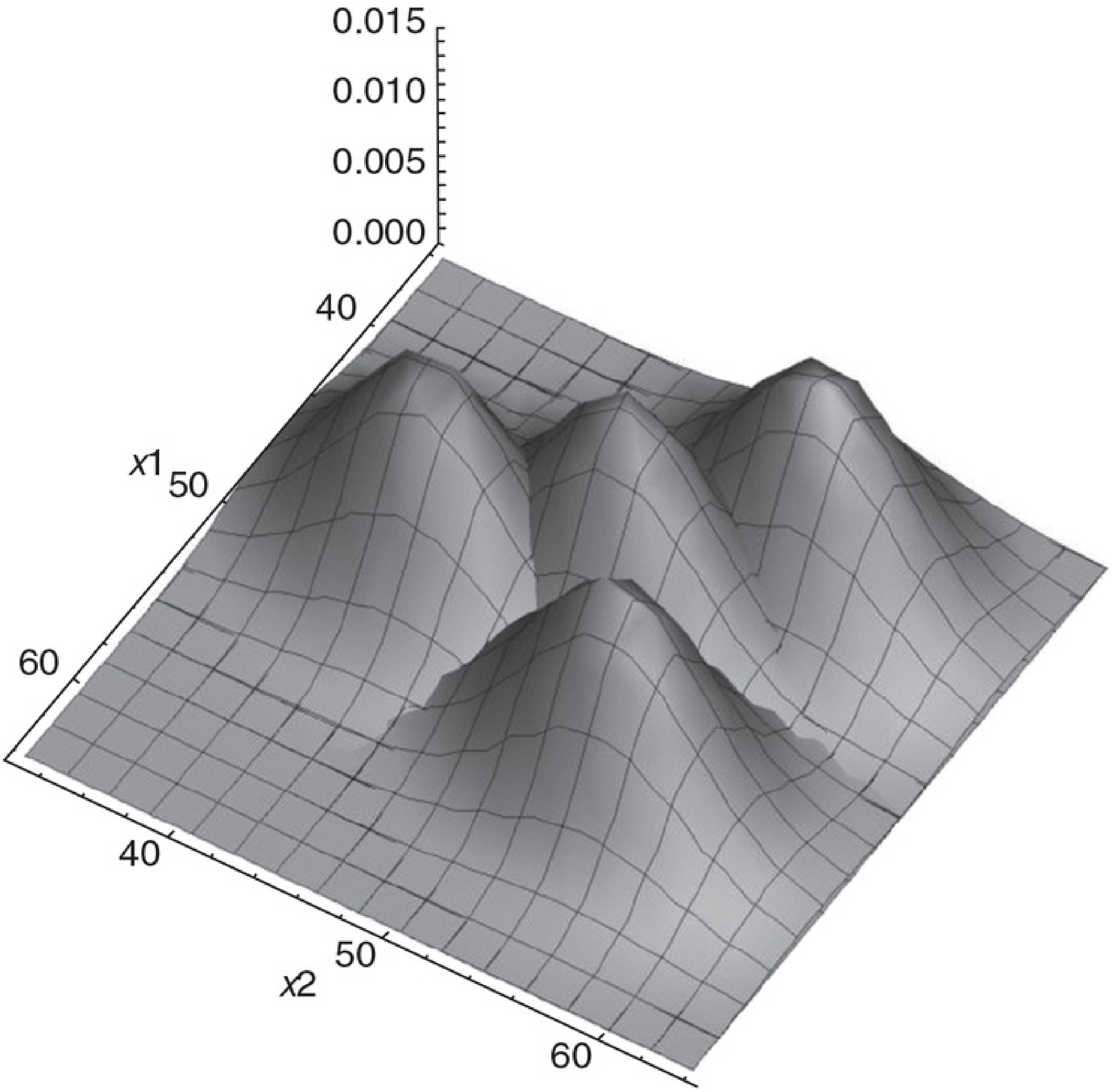


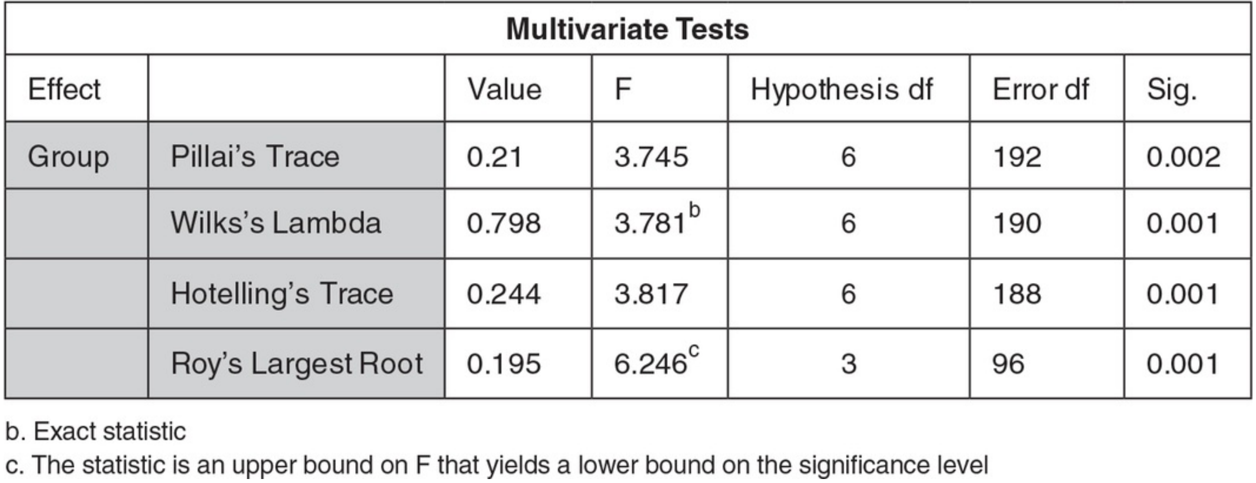
Figure 4 Multivariate normal curve model of four groups



Having data that meet the equal variance/covariance matrix assumption ensures that all individual bivariate normal distributions have the same shape and orientation.

The default SPSS MANOVA output for the example data is shown in Figure 5. The focus of the analysis is on the four “sig” levels of the group effect. Three of the four, Pillai’s trace, Wilks’s λ, and Hotelling’s trace, estimate the ratio of the variability between centroids and the within variability of the separate bivariate normal distributions. They do so in slightly different ways, but given fairly equal and large group Ns, will generate a sig level within a few thousands of each other. The interpretation of these three sig levels is that in combination, the means of dependent measures significantly differentiate between the groups. As in univariate ANOVA, the between variability can be partitioned using contrasts to account for the structure of group membership with separate main effects, interactions, and nested main effects, among others.

Figure 5 MANOVA Output Using SPSS



Multivariate Tests

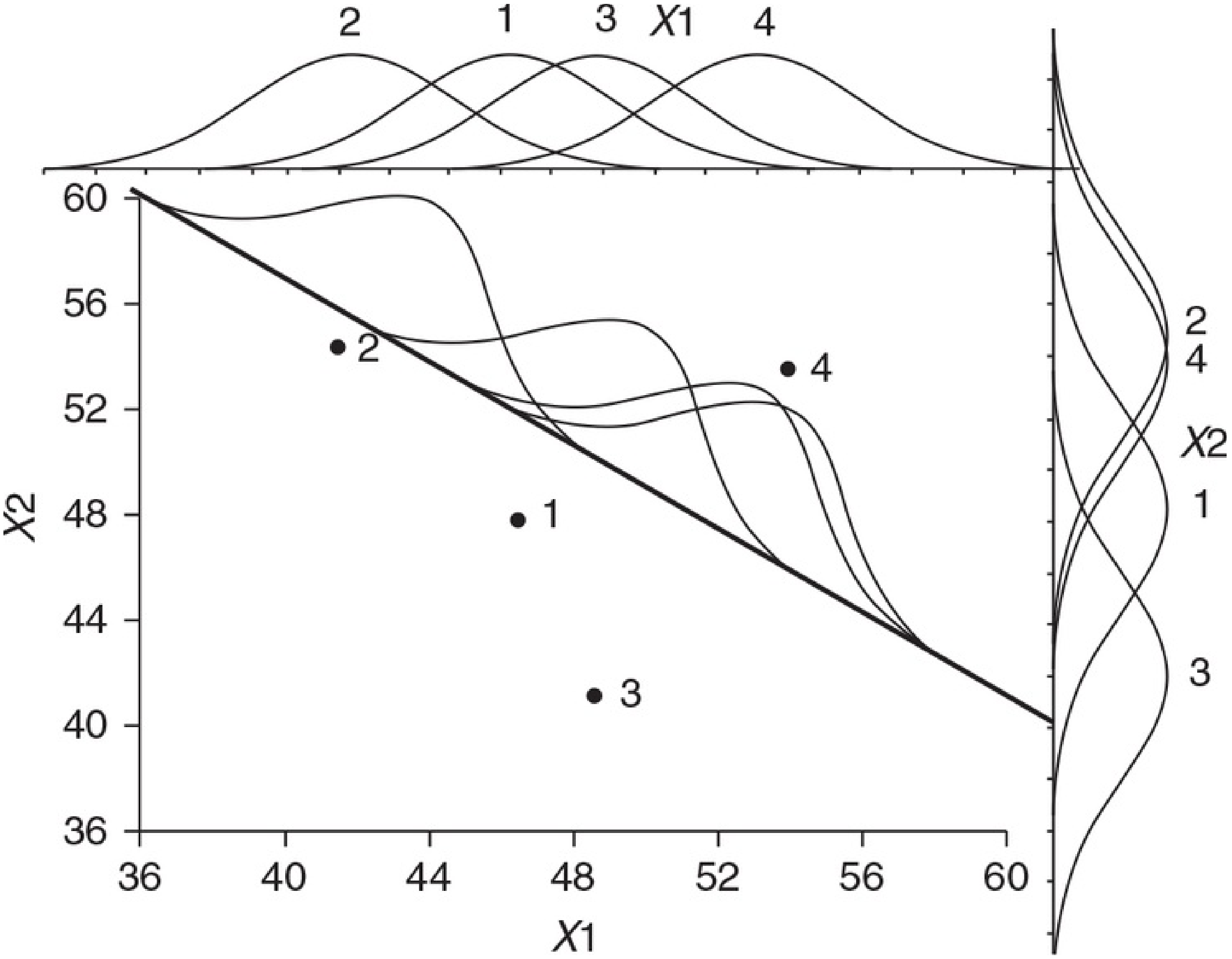
Effect		Value	F	Hypothesis df	Error df	Sig.
Group	Pillai's Trace	0.21	3.745	6	192	0.002
	Wilks's Lambda	0.798	3.781[b]	6	190	0.001
	Hotelling's Trace	0.244	3.817	6	188	0.001
	Roy's Largest Root	0.195	6.246[c]	3	96	0.001

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level

The fourth default statistic, Roy's largest root, takes a different approach to multivariate hypothesis testing. The data matrix is rotated (transformed using linear transformations) such that the variance between groups is maximized and the variance within groups is minimized. Figure 6 illustrates the rotation of the means in the example data, with the dark solid line showing the rotation. Roy's largest root is computed as a univariate ANOVA on the first extracted root and should be interpreted in light of this transformation. The *F* statistic for Roy's largest root will always be equal to or greater than the largest univariate ANOVA *F* statistic when there are only two dependent variables because if one or more of the dependent measures failed to add any discriminating ability beyond the other dependent measures, the transformation weight for those factors would be zero. Thus, the significance of Roy's largest root will always be equal to or smaller than the smallest of the significance levels. For the example data, the first root was extracted using DFA and saved as a variable to allow comparison with analyses.

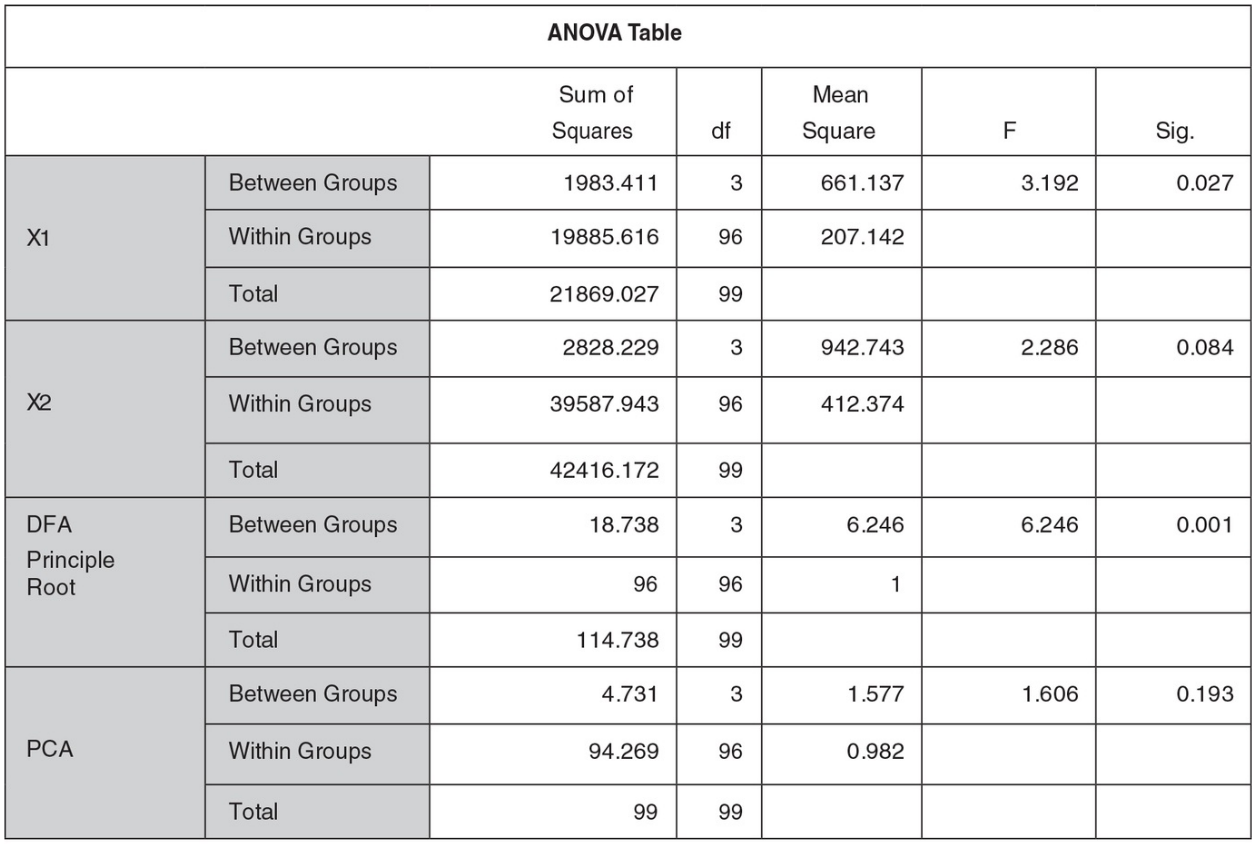
Figure 6 Extraction of largest root



With multivariate dependent measures, another option is to perform a principal component analysis (PCA) on the dependent measures and then do a univariate ANOVA on the first extracted factor, much like Roy’s largest root does on the first extracted root in DFA. In PCA, the first orthogonal factor has the greatest variance. This analysis was performed on the example data to compare its results with the others.

In order to interpret the results of MANOVA, univariate ANOVAs are often done to observe how the individual variables contribute to the variability. The results of univariate ANOVAs are presented in Figure 7 for X1, X2, DFA largest root, and the first factor in the PCA.

Figure 7 Univariate ANOVAs



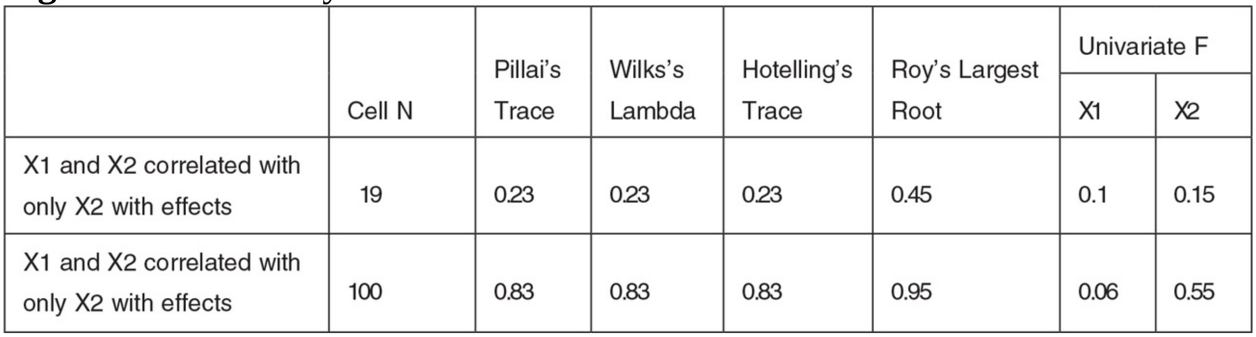
ANOVA Table						
		Sum of Squares	df	Mean Square	F	Sig.
X1	Between Groups	1983.411	3	661.137	3.192	0.027
	Within Groups	19885.616	96	207.142		
	Total	21869.027	99			
X2	Between Groups	2828.229	3	942.743	2.286	0.084
	Within Groups	39587.943	96	412.374		
	Total	42416.172	99			
DFA Principle Root	Between Groups	18.738	3	6.246	6.246	0.001
	Within Groups	96	96	1		
	Total	114.738	99			
PCA	Between Groups	4.731	3	1.577	1.606	0.193
	Within Groups	94.269	96	0.982		
	Total	99	99			

It is interesting to note that the MANOVA statistics all provided a smaller significance level than either of the two dependent measures individually. The univariate ANOVA on the DFA largest root was identical to Roy's largest root result presented in Figure 5. The PCA analysis had the largest significance level and was not statistically significant. The bottom line was that in this case MANOVA appeared to be more powerful than the individual univariate ANOVAs and that PCA did not appear to be a viable alternative.

Power Analysis of MANOVA With Three Groups and Two Dependent Measures

Power estimates for the various MANOVA statistics can be obtained by using simulated data. Figure 8 shows the estimated power of three simulations of 100 observations each and α set at .05. In the first case with a cell size of 10, X1 was generated using a random normal distribution and X2 was set equal to X1, with additional random normal error and small group effects added. That the effects were small relative to the random error can be seen in the low power (.15) observed for the univariate *F* test of the X2 variable. The power for X1 is greater than expected by chance. Pillai’s trace, Wilks’s λ, and Hotelling’s trace all showed a moderate and equal increase in power over the individual univariate power estimates. Roy’s largest root showed the greatest power at .45.

Figure 8 Power analysis



	Cell N	Pillai’s Trace	Wilks’s Lambda	Hotelling’s Trace	Roy’s Largest Root	Univariate F	
						X1	X2
X1 and X2 correlated with only X2 with effects	19	0.23	0.23	0.23	0.45	0.1	0.15
X1 and X2 correlated with only X2 with effects	100	0.83	0.83	0.83	0.95	0.06	0.55

The second analysis was similar to the first except that cell size was increased to 100. Similar results to the first analysis were found, with all power estimates except for X1 much larger than the case with the smaller cell size. Both of these simulations might be more appropriate for an analysis of covariance in which the variability of the first variable could be factored out before the second variable was analyzed.

The third analysis used a cell size of 50 and uncorrelated X1 and X2 variables, except they were each constructed with similar small effect added. Individually, the variables had power estimates of .38 and .43, respectively, but in combination, Pillai’s trace, Wilks’s λ, and Hotelling’s trace all showed a substantial increase in power. Roy’s largest root showed the greatest power at .87. Although this example is hardly a definitive power analysis, it makes a fairly strong argument that performing a MANOVA over multiple univariate ANOVAs results in a fairly significant increase in power.

MANOVA With Three or More Dependent Measures

MANOVA with three or more dependent measures provides a challenge in visualization and interpretation. Basically, the procedure is an extension of the simpler case of two variables but with a greater number of centroids for each group. MANOVA works by comparing the variability of the different centroids to the variability within cells. It requires the assumption of a multivariate normal distribution of the variables with equal variance/covariance matrices for each cell. Violation of these assumptions is likely to lead to a reduction in the power cell. Violation of these assumptions is likely to lead to a reduction in the power of the analysis.

If statistical significance is found for an effect in MANOVA using Pillai’s trace, Wilks’s λ, or Hotelling’s trace, it means that the centroids of the dependent variables are different for the different levels of the independent variable relative to the within variability. For three dependent variables, it is possible to create a three-dimensional visualization of the centroids and by rotating the vector get a reasonable understanding of the results. Beyond that, interpretation of results becomes problematic. Another caution, as in any multivariate analysis, is that when the measures are highly correlated, collinearity may generate strange results.

If statistical significance is found for an effect in MANOVA using Roy’s largest root, univariate ANOVA of the computed principal root can provide an interpretation of the results. In addition, an analysis of the linear transformation that is used to create the principal root can provide additional information, clarifying the results.

In terms of power in MANOVA, it seems reasonable to extend the limited power analysis just presented to the more complicated situation. Generally, that would mean that the power of MANOVA is greater than the individual univariate analyses. If statistical significance is found in a MANOVA, it does not necessarily mean that any of the univariate analyses will be significant. With respect to the increase in power in the case of Roy’s largest root, however, all bets are off in that if a DFA reveals more than one significant root, the power of analyzing only the principal root will be reduced.

Because of the difficulty in interpreting a MANOVA, it is recommended to use the technique to develop a deeper understanding of a data set only after a thorough understanding of the simpler, univariate data has been achieved. Rather than starting from the complicated analysis and working backward, start with the simple analysis and use the more complicated analysis to test hypotheses about multivariate relationships within the data.

Limitations

MANOVA provides an extension of univariate ANOVA to simultaneously test for effects over two or more dependent variables. In general, it delivers greater power than multiple univariate tests and its assumptions of similar power than multiple univariate tests and its assumptions of similar variance/covariance matrices for all cells are less onerous than the sphericity assumption necessary for repeated-measures ANOVA.

Although it has the advantage of generating output that is similar to ANOVA, difficulty of interpretation is MANOVA’s greatest limitation. Statistical significance in MANOVA shows that group means are different for different levels of the independent variable. With two and possibly three dependent measures, visual presentation allows the researcher some tools for analysis, but beyond that, if statistical significance is found, the researcher knows something is going on but is generally unsure of what it is.

Another limitation is the requirement that the dependent variables be a multivariate normal distribution with equal variance/covariance matrices for each cell. MANOVA is fairly robust with respect to this assumption when cell sizes are fairly large and approximately equal otherwise exploration of the reasonableness of this assumption is required.

David W. Stockburger

See also Analysis of Covariance; Analysis of Variance; Canonical Correlation; Discriminant Function Analysis; Normal Distribution; Power; Variance

Further Readings

Johnson, R. A., & Wichern, D. W. (1982). Applied multivariate statistical analysis (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Pedhazur, E. J. (1973). Multiple regression in behavioral research explanation and prediction (3rd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

van de Geer, J. P. (1971). Introduction to multivariate analysis for the social sciences. San Francisco, CA: W. H. Freeman.

N

NAEP

See National Assessment of Educational Progress

Rebecca Mazur Rebecca Mazur Mazur, Rebecca

Narrative Research

Narrative research

1127

1130

Narrative Research

The term *narrative research* (or *narrative inquiry*—the terms are often used interchangeably) has various definitions, but most broadly it refers to a research methodology that uses stories and storytelling as a source of knowledge. Interdisciplinary in nature, narrative research focuses on the structure and content of verbal communication (spoken and written) and assumes that because humans organize their memories and experiences primarily through narratives, stories contain messages about the nature of reality. Narrative research, which is an increasingly important form of inquiry in the social sciences, includes several methods such as discourse analysis, conversation analysis, sociolinguistics, and narratology. Each academic discipline, however, approaches narrative research differently and brings different sets of assumptions, concerns, and foci to the work. Narrative research is often difficult to distinguish from other forms of qualitative research, though it is best characterized by its emphasis on narrated texts that represent life stories, in whole or in part. Importantly, narrative research does not merely seek to uncover and retell stories; rather, it seeks to explore and interpret, in a disciplined way, peoples' lived experience in order to add to social science understanding. The remaining sections of this entry will investigate the philosophical origins of narrative research, various approaches to narrative research, subjectivity and influences of the researcher, the importance of providing validity and justification of claims, and ethical considerations of narrative research.

Philosophical Origins

Philosophically, John Dewey’s theory of experience is often cited as foundational to narrative research, as that theory supplies the rationale for approaching narrative through the three constructs of time, place, and social context. However, the theoretical work of numerous scholars and philosophers inform narrative researchers, including Martin Heidegger, Edmund Husserl, Ludwig Wittgenstein, Jerome Bruner, Clifford Geertz, and many others. Narrative research is situated within the reform movement that, beginning in the 1970s, posited that important aspects of personal and social experiences were inaccessible via conventional research methods. By the 1990s, the term *narrative inquiry* was being used in a variety of disciplines to describe methodology that used storytelling as the primary unit of analysis. Generally speaking, it can be said that four fundamental philosophic orientations are characteristic of narrative inquiry: the belief in the importance of the relationship between researcher and subject as a critical factor in research, a belief in the power of words (as opposed to numbers) to uncover knowledge, a belief that specific experiences have the ability to expose larger truths, and a belief that there are more ways of knowing than are generally accepted by traditional positivist research orientations.

Approaches to Narrative Research

There is not one universally accepted definition of the term *narrative*, as it tends to be interpreted differently depending on the discipline and needs of the researcher. A narrative unit can, for example, be a story told through a series of structured interviews, or it may be a story culled through hundreds of hours of observations. Written documents, transcripts of conversations, and visual artifacts may also be part of the research. Similarly, there is no single widely accepted way to go about conducting narrative research. This lack of doctrine may be both freeing and confusing to novice researchers. Usually in narrative inquiry, the research puzzle is approached by obtaining narratives from participants and then using thematic analysis, discourse analysis, or other similar analytical frameworks. Themes are explored with the understanding that people and their contexts are not fixed and that there is not one “true” representation of reality. Rather, narrative researchers assume that there are multiple and often conflicting truths at work in any story and that even the telling of a story may sincerely alter it.

Like most inquiry, narrative research is driven by a desire to fill a gap in existing knowledge about a given topic. For that reason, an extensive literature review is usually among the first steps of any narrative research project. Such a review usually among the first steps of any narrative research project. Such a review will help the researcher understand which parts of an idea or construct would benefit from being further investigated through narrative inquiry and will provide a foundation for how to approach the research puzzle.

Although narrative researchers collect data in a variety of ways, it is common to conduct interviews with subjects. Such interviews are usually unstructured or semi-structured, depending on the research question. Careful thought and planning go into narrative interviews. Closed questions (those that elicit a yes or no answer or that require a brief factual response) often do not help develop a narrative. On the other hand, questions that are too broad (e.g., “Can you tell me the story of when …?”) can intimidate or overwhelm participants. For that reason, most narrative researchers plan for a set of questions and neutral prompts that build on each other and encourage the tellers to provide details and explanations about the events or experiences involved in their story. Often the narrative interviewer does little talking, instead focusing on listening and supporting the storytelling. As participants speak, a researcher might facilitate storytelling by asking questions such as “What happened then?” or “How did you react when …?”

Whether accomplished through interviews, observations, or other means, narrative researchers keep detailed field texts that become the primary units of analysis for their work. In looking for patterns within those texts, narrative researchers will often attend to the content of a story (in other words, *what* is told) as well as to the structure of narration (*how* something is told). Although narrative researchers pay close attention to the sequence of how stories are told, they also remain aware of the common impulse to see stories as linear, with clear beginnings, middles, and ends. Although an analytic scheme emerges from thematic or discourse analysis, the stories that narrative researchers work with do not normally conform to the simpler definition of “story” that is used when discussing literature. Often, narrative researchers approach their work using the following types of questions: For what purposes, and for whom, was the story constructed? How was the story told? What social or cultural assumptions does the story make? What gaps or inconsistencies appear in the story that might indicate alternative narratives?

Researchers use an iterative process to analyze texts, often reading them multiple times to understand how themes relate to each other and to a larger whole. Often, a narrative researcher will review all texts once to get an overall picture of the story that has emerged. Then, he or she will read the text again (perhaps multiple times) to detect different “voices” that may surface, even with only one storyteller. This cycle continues until the researcher believes that he or she has a handle on the meaning and nuances of the text, including those pieces that may be confusing or contradictory. Most researchers will then seek to help situate their beliefs within a larger theoretical conversation by looking at a broad range of relevant literature.

Subjectivity and the Influence of the Researcher

Of importance to most narrative researchers is the idea of multiple subjective identities operating at once, including those of the researchers themselves. Even a single research subject has multiple facets of identity that act on the storytelling. For example, a person may clearly identify as a mother, friend, scientist, athlete, and Muslim all at once; some of these facets of identity may be stronger or more well-developed than others and thus more dominant in the narrative. This creates a tension that the researcher must be mindful of both in the field and when analyzing data. Because of the way narrative researchers are embedded in their work, and the way texts are often cocreated by the researcher and subject, the researcher, too, must often consider the researcher’s own positionality in relation to the subject and the story. Typically, narrative researchers develop a method of taking field notes that allows for continuous self-referential observations about how the researcher reacts to the telling or the teller of the story. In addition to adding richness to the data, these observations help researchers to be cognizant of and explicit about the influence that their own subjectivity may have on the telling of, or interpretation of, the story. They also help researchers navigate their work with honesty and integrity.

Some narrative researchers find that their work is improved when they are able to connect with a community of critical friends. These are usually individuals, both academic and nonacademic, whom the researchers trusts to provide accurate and responsible feedback about their ongoing work. Such critical friend communities, when carefully chosen and appropriately diverse, help researchers see the ways that they (the researchers) may be shaping the experience of subjects and the unfolding of the narrative.

Validity and Knowledge Claims

Although stories are at the heart of narrative researchers’ work, storytelling itself is not the purpose of narrative research. Rather, researchers in this tradition work is not the purpose of narrative research. Rather, researchers in this tradition work to uncover truths about the human condition, and they make knowledge claims based on the results of their inquiry. Thus, as in any other research endeavor, narrative researchers are obligated to provide readers with justification for any claims that they make. However, unlike large-sample research that, using statistics, can provide confidence intervals for researchers’ assertions, narrative research deals with complex human experiences on a small scale that cannot be statistically tested.

Instead, it is incumbent on narrative researchers to present sufficient evidence to their readers to support all assertions drawn from the research, and it is incumbent upon readers to make judgments about the extent to which narrative researchers’ claims are plausible, credible, and trustworthy. For this reason, much narrative research supplies richly detailed descriptions of human experiences, so that researchers can construct, and readers can evaluate, knowledge claims based on the story. Further, it is common (though not strictly mandatory) for narrative research to include a section that provides interpretation of the story or stories presented. These sections offer commentary about the meaning of the text and elaborate on the implications of what the research reveals. In some ways, interpretations of narrative research rely on similar tools and techniques as those used in literary criticism, such as close textual analysis. And, similar to literary criticism, narrative researchers do not need to claim that their interpretation is the only one possible; rather, they work to show readers that their knowledge claims are fair and plausible interpretations of the story and that they are well-grounded in the evidence of the narrative.

Ethical Considerations

Standard issues of privacy, confidentiality, informed consent, justice, and beneficence all apply to narrative research. However, narrative research is somewhat unique among qualitative research methods because of its emphasis on relational engagement between researchers and participants and therefore may also need to grapple with further ethical considerations. Often, narrative researchers spend a great deal of time listening to participants tell their stories, and sometimes spend time living alongside participants. This means that researchers must be responsive to ethical tensions that may arise in the course of their research, specifically related to the well-being of participants and the relationship between the researcher and participant. When working on long-term projects, researchers must attend to the continued maintenance of informed consent, especially if the scope or substance of the research changes. Also of concern is how participants consent to the final text of the research, which in some cases may reveal intimate personal thoughts, feelings, and experiences. Moreover, and especially when working with vulnerable populations, narrative researchers must grapple with a host of complicated ethical concerns about the well-being of their participants, and often researchers negotiate ways that they can be of help to participants during, and sometimes after, the research period.

Rebecca Mazur

See also Ethical Issues in Educational Research; Interviews; Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Bruner, J. (1991). The narrative construction of reality. Critical Inquiry, 18(1), 1–21.

Clandinin, D. (Ed.). (2007). Handbook of narrative inquiry: Mapping a methodology. Thousand Oaks, CA: SAGE.

Clandinin, D., & Connely, F. (2000). Narrative inquiry: Experience and story in qualitative research. San Francisco, CA: Jossey-Bass.

Polkinghorne, D. E. (2007). Validity issues in narrative research. Qualitative Inquiry, 13(4), 471–486.

Wertz, F., Charmaz, K., McMullen, L., Josselson, R., Anderson, R., & McSpadden, E. (2011). Five ways of doing qualitative analysis. New York, NY: Guilford Press.

White, H. (1980). The value of narrativity in the representation of reality. Critical Inquiry, 7(1), 5–27.

Patricia A. Jenkins Patricia A. Jenkins Jenkins, Patricia A.

National Assessment of Educational Progress

National assessment of educational progress

1130

1131

National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP) is the largest ongoing assessment of students in the United States and measures their knowledge and performance across varied subject areas. Assessments are conducted regularly in Grades 4, 8, and 12 and are essentially the same from year to year, so the results allow comparisons of, for instance, how eighth graders performed on reading in 2013 compared to eighth graders in 2015. This entry gives an overview of the history and governance of the NAEP and concludes with a discussion of the assessment design and how the results are reported.

Planning for the NAEP began in the 1960s with a grant from the Carnegie Corporation and the establishment of the Exploratory Committee for the Assessment of Progress in Education. The first national paper-and-pencil assessments were conducted in the late 1960s. Voluntary trial assessments for the states began in the 1990s; selected urban districts started offering the assessments on a trial basis in the early 2000s. District participation continues under the Trial Urban District Assessment program.

The NAEP science test has changed to include assessment via interactive computer tasks. The NAEP writing tests for Grades 8 and 12, and the NAEP technology and engineering literacy assessments, are now administered entirely on the computer. The computer-based assessments remain uniform to continue reporting comparable state-level achievement results.

NAEP is a congressionally mandated project of the National Center for Education Statistics in the U.S. Department of Education. The National Center for Education Statistics commissioner is responsible, by law, for the NAEP project. The National Assessment Governing Board is appointed by the secretary of education and is an independent, bipartisan group that oversees the NAEP. Board members include local and state school officials, educators, business representatives, governors, state legislators, and members of the general public. The board develops the NAEP framework and test stipulations, sets policies, and informs the public of results in the Nation's Report Card.

NAEP assessments cover subjects that include mathematics, reading, writing, science, U.S. history, geography, the arts, and civics, as well as technology and engineering literacy. A sampling procedure is used to represent the geographical, racial, ethnic, and socioeconomic breakdown of U.S. schools and students.

In addition to reporting on test results, the Nation's Report Card describes the school environment for populations of students (e.g., all fourth graders) and groups within those populations (e.g., Hispanic students). Test scores are not provided for individual students or schools; however, NAEP may report results of selected, large urban districts.

Patricia A. Jenkins

See also No Child Left Behind Act; Paper-and-Pencil Assessment; Partnership for Assessment of Readiness for College and Careers; Programme for International Student Assessment; Smarter Balanced Assessment Consortium; Standardized Tests

Websites

National Assessment Governing Board: www.nagb.org

National Center for Education Statistics: https://nces.ed.gov/

Laura Pevytoe Laura Pevytoe Pevytoe, Laura

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

National Council on Measurement in Education National council on measurement in education

National Council on Measurement in Education

The National Council on Measurement in Education (NCME) is an organization for professionals who conduct assessments, testing, evaluations, and other processes that make up educational measurement. The NCME seeks to advance the practice of educational measurement through the creation and implementation of standardized tests, assessment tools and materials, program design, and program evaluation.

The NCME’s goal is to act as the recognized authority in the evaluation and measurement of programs and practices in education. NCME emphasizes the importance of improving tools, procedures, and public policy in the field of educational measurement with information garnered through scholarly work. Members of the NCME include individuals involved in the areas of test development and research; program evaluation; certifying, credentialing, and licensing of professionals; and graduate students studying education and psychology.

The NCME arms its constituents, including its professional members and those who have limited or no formal assessment training such as K–12 educators, parents of students, and journalists, with an abundance of resources. Companies that create educational tests have compiled a glossary of terms related to educational testing and assessment using straightforward, nontechnical language that is accessible to the general public and is available through the NCME website. The NCME website also announces opportunities to attend training sessions or webinars for aspiring students or others interested in careers in educational measurement and assessment.

Along with formulating and publishing testing tools and materials, the NCME also publishes two quarterly journals. The *Journal of Educational Measurement* is a scholarly periodical that covers educational measurement in field settings and theories of measurement practices. Topics include original measurement research, accounts of those who have used new measurement tools, and reviews of related publications. *Educational Measurement: Issues and Practices* is a journal aimed at professional educators and practitioners, and members of the public who may be interested in learning more about educational measurement. The principal focus of the *Educational Measurement: Issues and Practices* is to encourage analytical discussion of contemporary issues in educational measurement, such as defining and measuring college readiness and evaluating growth of English-language learners. This journal also offers an outlet for NCME members to communicate with each other and with those interested individuals outside of the organization about relevant and recent topics in the field of education measurement.

In addition to publishing its own works, the NCME collaborates with several other organizations to publish testing standards. In July 2014, a new edition of *Standards for Educational and Psychological Testing* was released as a cooperative product of the NCME, the American Educational Research Association, and the American Psychological Association. These testing standards are considered the benchmark of testing guidelines and have been published jointly by these three groups since 1996.

Laura Pevytoe and Marc H. Bornstein

See also American Educational Research Association; American Psychological Association; Joint Committee on Standards for Educational Evaluation

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association. Retrieved from http://www.ncme.org/ncme/NCME/Publication/NCME/Publication/Testing_St hkey=d05e3e89-c121–401d-83af-99af87f32aed.html

National Council on Measurement in Education. (n.d.). About NCME. Retrieved from http://www.ncme.org/ncme/NCME/About/NCME/About1/About1.aspx?hkey=7548ee94–566a-4475-ba8e-926425043430.html

National Council on Measurement in Education. (n.d.). Educational measurement: Issues and practices. Retrieved from http://www.ncme.org/ncme/NCME/Publication/Educational_Measurement/NC hkey=d6608ef9–7f3b-4a57–8518-ddb968dc6a26.html

National Council on Measurement in Education. (n.d.). Glossary of important assessment and measurement terms. Retrieved from http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resour hkey=4bb87415–44dc-4088–9ed9-e8515326a061.html

National Council on Measurement in Education. (n.d.). Journal of Educational Measurement. Retrieved from http://www.ncme.org/ncme/NCME/Publication/Journal_of_Educational_Meas hkey=6380e466-a3ec-4154-b06f-96888a76ec97.html

Websites

National Council on Measurement in Education: http://www.ncme.org/NCME

John F. Dovidio John F. Dovidio Dovidio, John F.

National Science Foundation

National science foundation

1132

1134

National Science Foundation

The National Science Foundation (NSF), which was created by the U.S. Congress in 1950, is currently the second largest U.S. government research-funding agency (next to the National Institutes of Health). The president appoints the NSF director and the 24 members of the National Science Board, which establishes the overall policies of the foundation. With an annual budget of approximately $7.5 billion, NSF funds about 24% of all federally supported basic research conducted by U.S. colleges and universities. It receives 50,000 proposals per year and annually supports 200,000 scientists, engineers, educators, and students in the United States and throughout the world. NSF is the major source of federal funding in the social sciences and in science education. This entry reviews the mission and organization of NSF, types of NSF grants, and the process of submitting a proposal to NSF and its subsequent review.

Mission and Organization

According to the NSF’s website, the stated mission of NSF is “to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense.” NSF emphasizes basic research—research that addresses fundamental conceptual issues that may not necessarily have immediate practical impact but may have significant long-term theoretical, social, and pragmatic influences.

NSF includes seven directorates, with the goal of promoting research in diverse areas of science, technology, and education. These directorates include education and human resources (supporting science, technology, engineering, and mathematics education) and social, behavioral and economic sciences (supporting anthropology, economics, linguistics, management, neuroscience, psychology, and sociology). The other directorates are biological sciences, computer and information science and engineering, engineering, geosciences, and mathematical and physical sciences.

Types of NSF Grants

Researchers most commonly seek support from NSF for specific research projects. These proposals are typically for multiyear projects and may have a single principal investigator or involve collaborations within or across research institutions. Because these proposals are evaluated not only on the significance and originality of the ideas but also on the feasibility of the project and the likelihood of success, the proposals typically present a systematic program of research with some preliminary evidence of promising results.

In addition to these basic research proposals, NSF has a number of other programs designed to help achieve its general mission to promote the progress of science. The CAREER Program at NSF promotes the early development of academic faculty as both educators and researchers and is intended to foster the integration of research and education components of a faculty career. In addition, NSF has a specific grant program for research among faculty who teach at predominantly undergraduate colleges: Research in Undergraduate Institutions.

NSF also directly supports undergraduate and graduate education. The Research Experience for Undergraduates program provides funding for projects that actively and meaningfully involve undergraduate students in research in any of the areas normally funded by NSF. Research Experience for Undergraduates grants are awarded to initiate and conduct projects that engage a number of students in research from the host institutions and as well as other colleges. In addition, NSF provides Graduate Research Fellowships for outstanding graduate students, and the directorate for social, behavioral, and economic sciences offers Minority Postdoctoral Research Fellowships.

Proposal Submission and Review

Proposals to NSF, which must strictly conform to NSF guidelines, are evaluated

Proposals to NSF, which must strictly conform to NSF guidelines, are evaluated based on two primary criteria: intellectual merit and broader impacts. Intellectual merit reflects the importance of the proposed project for advancing knowledge, as well as the degree to which the project is well conceived and explores creative and potentially transformative ideas and issues. Broader impacts involve the extent to which the project promotes the teaching and training of graduate students, undergraduate students, and postdoctoral fellows; broadens the participation of members of underrepresented groups in research; enhances the infrastructure for research and education; and produces results that benefit society.

Because NSF’s budget enables it to support only a limited percentage (typically between 20% and 25%) of the proposals it receives, the merit review process represents the way the foundation evaluates and prioritizes grant proposals for funding.

In merit review, grant proposals are initially sent to outside reviewers (about 40,000 annually) who have relevant expertise on the topic. Reviewers evaluate the proposal on intellectual merit and broader impacts.

Building on these initial reviews, proposals are evaluated by standing NSF panels that are composed of scholars who, as a group, have the expertise that is necessary to evaluate the range of projects submitted to a particular program. The comments and rankings of the review panel are recommendations to the program officer, who is an NSF staff member. The panel does not make the funding decisions.

The program officer considers the recommendation and discussions of the review panel but may also include other factors, such as the transformative potential of the project or particular program objectives (e.g., balance, synergy, and diversity of ideas; encouragement for new investigators), in making funding recommendations to the NSF division director. If the division director concurs, the recommendation is submitted to the Division of Grants and Agreements for award processing.

All investigators submitting proposals to NSF receive written feedback about their proposals. In addition to a statement of the general evaluative category (e.g., “not competitive—revision encouraged”), investigators receive the initial comments from individual reviewers (anonymized) and a summary of the review panel’s deliberations. A proposal that is “declined” by NSF may be resubmitted, but only after it has been substantially revised, and it must directly address concerns made by the previous reviewers.

The support of NSF is critical for researchers in science and education. NSF is the major funding source for basic research in the United States. Grants from NSF allow scholars to pursue research that benefits society, which would not be possible by relying solely on support from the investigator's institution. NSF has provided essential support for such diverse projects as building Alvin (the deep sea exploration vessel), conducting seminal research on DNA, improving K–12 science and technology education, and supporting supercomputer facilities. Finally, NSF also shapes the direction of research by identifying particular topics of both academic and practical value to society, such as improving the quality of science education and increasing the number of members of traditionally underrepresented groups who pursue careers in science, technology, engineering, and mathematics fields.

John F. Dovidio

See also Educational Research, History of; Experimental Designs; Merit; Research Proposals; Transformative Paradigm

Further Readings

National Science Foundation. (2010). NSF sensational 60. Retrieved from https://www.nsf.gov/about/history/sensational60.pdf

National Science Foundation. (2015). Report to the National Science Board on the National Science Foundation's merit review process, fiscal year 2014. Retrieved from http://www.nsf.gov/nsb/publications/2015/nsb201514.pdf

National Science Foundation. (2016). The National Science Foundation proposal and award policies and procedures guide. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_print.pdf

Nicole Mittenfelner Carl Nicole Mittenfelner Carl Carl, Nicole Mittenfelner

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

Naturalistic Inquiry

Naturalistic inquiry

1134

1137

Naturalistic Inquiry

Naturalistic inquiry refers to research that takes place in research participants' natural settings (i.e., in a school or home rather than in a laboratory or contrived space outside of participants' daily lives and routines). Many qualitative research genres adhere to the principles of naturalistic inquiry, and ethnography, with its primary method of participant observation, is probably the genre of qualitative research that is most emblematic of naturalistic inquiry. However, many approaches to qualitative research place a primacy on conducting research in natural settings and strive to deeply understand people and phenomena. Naturalistic inquiry involves a methodological pursuit of understanding the ways individuals view, approach, and experience the world and make meaning of their experiences. Those engaged in naturalistic inquiry do not believe in or attempt to find an objective truth; instead, they are interested in individuals' subjective interpretations of experiences and events, which are embedded in multiple contexts that are temporal, societal, and personal. A central premise in both naturalistic and qualitative inquiry is that individuals are experts of their own experiences, and a primary goal of naturalistic inquiry is to collect, analyze, and present contextualized understandings of people, settings, and phenomena.

Naturalistic inquiry is an important part of educational research for a variety of reasons, including that empirical research related to education, schooling, and schools is a key aspect of the field of education. Educators and scholars can engage in inquiry-based research that generates local, embedded knowledge within practice, policy, and scholarship. Researching educational endeavors in their contextualized settings provides researchers and practitioners with more their contextualized settings provides researchers and practitioners with more holistic understandings of issues, ideas, experiences, perspectives, approaches, and events in context. In naturalistic inquiry, knowledge is constructed through individuals' subjective experiences, and because of this, conducting research in "natural" settings is vitally important to the study and practice of education. These settings are often referred to as "the field," and naturalistic research is also known as "fieldwork." This entry begins with a brief history of naturalistic inquiry and then provides an example of naturalistic research processes. It then discusses the central components of naturalistic inquiry and concludes by describing the importance of the role of the researcher in naturalistic inquiry.

History of Naturalistic Inquiry

Naturalistic inquiry developed as a formal field in the 1960s; however, the ethnographic tradition, which is a prominent example of naturalistic inquiry, began in the 19th century, and action research, which often takes place in naturalistic settings, began in the 1940s and 1950s. Ethnography and naturalistic inquiry in general have evolved significantly since their origins. Naturalistic inquiry developed, in part, as a critique of positivist research. Positivism refers to the worldview, or paradigm, that contends that there are universal laws and truths that can be objectively studied, tested, and verified. Positivist research broadly refers to experimental designs in which quantitative variables are used to measure relationships. Naturalistic inquiry, in contrast with positivism, states that social research ought to be studied in natural settings, which are the environments individuals naturally inhabit. At the heart of naturalistic inquiry is the interpretative paradigm, which asserts that knowledge is subjectively constructed and cannot be reduced to causal relationships or universal truths because actions are based on and mediated by a multitude of social and cultural beliefs, values, and experiences.

Broadly, naturalistic inquiry seeks to understand how the world is socially constructed and experienced; is based on methods that are iterative, recursive, and flexible; and focuses on developing as contextualized a picture of experiences and settings as possible. Rather than claiming there are universal "Truths," naturalistic researchers believe in multiple, situated truths and realities. Despite its association with specific disciplines such as social anthropology and theories such as naturalism and interpretivism, there are many approaches to conducting naturalistic inquiry. The concept of naturalism underscoring naturalistic inquiry contends that the social world should be examined in its “natural” state; there are other philosophical definitions of naturalism that describe it differently, and this is one reason that the term *qualitative research* is often more widely used than naturalistic inquiry. Although not every qualitative study takes place in naturalistic settings, the primary values of qualitative research are that it recognizes that there are multiple realities and truths and it attempts to study people and phenomena in natural settings. Some researchers use the terms *naturalistic inquiry* and *qualitative inquiry* interchangeably. There are many qualitative research genres, or methodological approaches, that strive to uphold the values of naturalistic research, such as action research, case study research, ethnography, evaluation research, grounded theory, narrative inquiry, participatory action research, phenomenology, and practitioner research. The specific methodological approach employed is guided by the study’s research questions and goals as well as from existing theory, empirical research, and the researcher’s beliefs and values. Researchers can also combine methods from different genres. For example, researchers conducting an evaluation may use ethnographic methods to help develop the evaluation design, or a case study may use participatory methods.

An Overview of the Processes of Naturalistic Research

Naturalistic inquiry is a form of empirical research that ideally involves the systematic collection and analysis of data through processes that are flexible, emergent, and recursive. As previously described, there are many approaches to conducting naturalistic research, and the processes for a study vary based on the genre, goals, and research questions. An overview of one process is provided to help demonstrate naturalistic research methods.

A naturalistic inquiry typically begins with a research focus, interest, problem, or question. This focus is often developed by conducting a literature review and discussing the topic with others, including mentors, advisers, colleagues, and/or other individuals with specific knowledge about a setting or phenomenon. Throughout this process, researchers tend to develop what is considered a theoretical and/or conceptual framework that informs the selection of a topic, research questions, and study design. Part of the research design process includes determining the methods that will best answer the research questions. Particulars include determining a research setting, selecting participants, and developing a plan for data collection and analysis. Data can be collected through a variety of methods, and the use of participant observation, interviews, and a review of artifacts are common in naturalistic inquiries. It is important to note review of artifacts are common in naturalistic inquiries. It is important to note that in naturalistic research, like qualitative research, the research questions can and often do evolve throughout the course of the study. The primacy is placed on the participants and phenomena, not the specific questions or research methods. The way the data are collected and analyzed can vary considerably, as there is no single way of conducting a naturalistic study, and the methods depend on the research goals and focus. Naturalistic research involves circular, back-and-forth processes. For example, researchers, depending on the specific approach used, continue to consult literature throughout their study and especially again during data analysis, not just at the beginning of a study.

Naturalistic research depends on a robust research design and rigorous data collection and data analysis methods. Furthermore, a thorough description of these processes should be included in the research report or product. Doing so allows readers to understand the processes used to collect and analyze the data and therefore to have a better sense of the validity, or trustworthiness, of a study. Related to this, naturalistic inquiry has developed a set of criteria used to determine the validity of qualitative/naturalistic research, and at the heart of these criteria are that researchers should provide as much context and detail as possible that demonstrate the quality and rigor of the study. An important strategy for achieving validity in naturalistic inquiry is thick description, which is how researchers describe the study with the goal of contextualizing the research setting, participant group(s), and participants’ experiences so that readers can create a vivid, layered picture of the setting, participants, and context in their minds and determine the quality of the research and interpretations rather than relying solely on data excerpts and written analyses.

Central Components of Naturalistic Inquiry

There is not one single way to conduct a naturalistic inquiry, and naturalistic inquiry is not limited to a specific discipline, field, theory, or approach. However, naturalistic inquiries share some common characteristics, including that they (a) involve naturalistic fieldwork, (b) evolve based on emerging learnings, (c) involve inductive data collection and analysis processes, (d) pay close and careful attention to context(s) and the ways in which they shape and inform people’s experiences and perspectives, (e) involve prolonged contact with participants, and (f) strive to develop holistic interpretations.

Naturalistic fieldwork means that researchers are physically present with individuals in the research setting, which may be a community, group, or an individuals in the research setting, which may be a community, group, or an institution. Because naturalistic inquiry tends to evolve based on emerging learnings, the research design is not fixed but rather is flexible and emerges or changes based on data and analysis. The research focuses on the phenomenon and participants rather than strictly adhering to specific methods, given that the goal is to engage with and understand the complexities of participants' experiences. Naturalistic inquiry is also characterized as being inductive because researchers develop concepts, hypotheses, and theories from the data rather than bringing preset ones derived deductively from theory. Naturalistic inquiries focus on describing individuals, situations, and experiences in context so that the research reflects how individuals' lives and experiences are complex, temporal, and influenced by a variety of mediating factors. Researchers do this by prolonged contact in the field with participants, and this prolonged contact also helps researchers to develop holistic interpretations.

It is also important to note that naturalistic research methods are iterative and recursive. This means that naturalistic research typically evolves over time and that the research is informed by, and depends on, all of its component parts. In addition to the aforementioned components, naturalistic inquiry values the interpretations, subjectivities, and the nonneutrality of the researcher, which is discussed in the next section.

The Role of the Researcher in Naturalistic Inquiry

The researcher is considered the primary instrument in naturalistic and interpretative research. This means that there is an explicit acknowledgment that researchers directly shape the data collected, and therefore, their subjectivities and choices influence the findings. Because of this, qualitative researchers should pay careful attention to issues of reflexivity, which refers to the systematic assessment of a researcher's identity, positionality, and subjectivities both broadly and then specifically in terms of the topics and setting of the specific research study. In practice, this means that researchers should consider these aspects through the processes of memo writing and in ongoing, constructively critical dialogue with peers and advisers. For example, researchers might examine how gender, social class, race, sexual identity/orientation, culture, and ethnicity may influence the research, relationships with participants, and other facets of the research process. Positionality refers to how the researcher is situated in relationship to the research context and setting, which can include social identity and/or role vis-à-vis the setting and participants, and includes the different roles and relationships that exist between the researcher and the participants. The researcher's subjectivities impact the research in myriad ways, including influencing the selection of the topic and setting as well as determining what information is focused on and what or who is included and excluded. Developing a reflexive practice as a researcher involves paying close attention to and complicating ways that the research is influenced by the identity, positionality, and subjectivities of the researcher(s). As previously stated, naturalistic research does not believe in neutral, objective research but argues that research is subjective, partial, and political (in macro and/or micro ways). Because the researcher is the primary instrument in naturalistic research, the ways that the researcher shapes the research should be acknowledged and engaged with in systematic ways through reflexive practices.

Nicole Mittenfelner Carl and Sharon M. Ravitch

See also Action Research; Conceptual Framework; Document Analysis; Ethnography; Grounded Theory; Interviews; Member Check; Narrative Research; Participatory Evaluation; Phenomenology; Positivism; Qualitative Data Analysis; Qualitative Research Methods; Trustworthiness; Validity

Further Readings

Cochran-Smith, M., & Lytle, S. L. (2009). Inquiry as stance: Practitioner research for the next generation. New York, NY: Teachers College Press.

Creswell, J. W. (2013). Qualitative inquiry and research design: Choosing among five approaches (3rd ed.). Thousand Oaks, CA: SAGE.

Denzin, N. K., & Lincoln, Y. S. (2011). Introduction. In N. K. Denzin & Y. S. Lincoln (Eds.), The SAGE handbook of qualitative research (4th ed.), pp. 1–19). Thousand Oaks, CA: SAGE.

Hammersley, M., & Atkinson, P. (2007). Ethnography: Principles in practice (3rd ed.). Hoboken, NJ: Taylor & Francis.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: SAGE.

Maxwell, J. A. (2013). Qualitative research design: An interactive approach (3rd ed.). Thousand Oaks, CA: SAGE.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). Qualitative data analysis: A methods sourcebook. Thousand Oaks, CA: SAGE.

Ravitch, S. M., & Carl, N. M. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: SAGE.

Ravitch, S. M., & Riggan, M. (2016). Reason and rigor: How conceptual frameworks guide research (2nd ed.). Thousand Oaks, CA: SAGE.

NCES

NCES

1137

1137

NCES

See Normal Curve Equivalent Score

NCLB

See No Child Left Behind Act

NCME

See National Council on Measurement in Education

Valeisha M. Ellis Valeisha M. Ellis Ellis, Valeisha M.

Needs Assessment

Needs assessment

1137

1139

Needs Assessment

The term *needs assessment* is a systematic approach that gathers data by means of established procedures and methods through a defined series of phases. Needs assessment sets priorities and determines criteria for solutions, so that stakeholders can make informed decisions. It also sets criteria to determine the most effective method to use human capital, revenue, and other resources. This entry describes how needs assessment leads to action that will improve services, programs, operations, organizational structure, and the significance of needs assessment in education research, measurement, and evaluation.

A “need” is the gap between the current state (what is) and a desired state (what should be). Needs assessments are conducted to determine needs, study their nature and causes, and prioritize future action. Needs assessments are focused on specific targeted populations in an organization. In education, a targeted population may be students, parents, teachers, administration, or community in general. Although initially a needs assessment is conducted to determine the needs of the population for whom the organization or system exists, a comprehensive needs assessment often takes into account needs identified in other parts of the system. It is critical to understand that a needs assessment is not complete unless plans are made to use the information in a practical and meaningful way.

There are several basic approaches to identifying needs for an organization. The discrepancy views a need as a discrepancy between a desired performance and an observed or predicted performance. A democratic view identifies a need as a desired change by the majority of a particular group. An analytic view perceives a need to be the direction in which improvement can be predicted to occur given a need to be the direction in which improvement can be predicted to occur given information about the current status. A diagnostic view perceives a need as something that causes harm by its absence or deficiency. Although there is not a common accepted definition of need, the need assessor decides which definition will be an appropriate guide for a study.

A needs assessment is a systematic approach with three distinct phases. The three phases have distinct outcomes. The outcome in Phase 1 is the preliminary plan for data collection in Phase 2. The outcome of Phase 2 is the criteria for action based on high-priority needs. The outcome for Phase 3 is the action plan(s), written and oral briefings, and the final report.

The essence of Phase 1 is to identify what is already known about the needs of the target population by establishing a commitment to the needs assessment and obtaining reassurance that the administration will use the findings with appropriate action in a suitable manner. To completely explore the needs in Phase 1, the team must prepare a management plan, identify major concerns, determine need indicators, and consider data sources.

The fundamental purpose of Phase 2 is to gather and analyze data. The first step is to establish the scope of the needs assessment and determine the target groups. The next step in Phase 2 is to gather data to define needs. Prioritizing needs is the third step. The fourth step is to identify and analyze causes. The final step in Phase 2 is to summarize findings and share the results with the needs assessment committee, managers, or other key stakeholders. The major accomplishment of Phase 2 is a set of needs statements prioritized of importance.

Phase 3 is the decision-making phase that moves the needs analysis to action. First, one should group priority of needs into two stages. The first stage of Step 1 is to identify broad areas and then critical areas within each area. The second step is to identify possible solutions. During Step 3, solutions that move toward the contemplated change are selected. An action plan that includes descriptions of the solutions, rationale, proposed timelines, and resource requirements is created in Step 4. The final step in Phase 3 is a final report. The final report should include a description of the needs assessment process, major outcomes (identified needs), priority needs (and criteria used to determine such priorities), an action plan (with the data and criteria used to arrive at the solution strategies), and recommendations for future needs assessments.

The needs assessment is an integral part of continuous program improvement. Planning, implementing, evaluating, and improving the needs assessment is a

Planning, implementing, evaluating, and improving the needs assessment is a cyclical process that should move from conducting a needs assessment to updating the needs assessment continuously for systematic organizational change. There is a variety of data that can be collected from needs assessments in education (i.e., test scores, school involvement, academic progress, health indicators, personal characteristics). A thorough needs assessment provides a foundation for other assessment measures. Therefore, the needs assessment measure is not meant to be used in isolation. The data collected from needs assessments add value and provide actionable steps for improving education programs, operations, and organizational structure.

Valeisha M. Ellis

See also Process Evaluation; Program Evaluation

Further Readings

Gupta, K. (2011). A practical guide to needs assessment. New York, NY: Wiley.

Igarashi, M., Suveges, L., & Moss, G. (2013). A comparison of two methods of needs assessment: Implications for Continuing Professional Education. Canadian Journal of University Continuing Education, 28(1).

McClelland, S. B. (1995). Organizational needs assessments: Design, facilitation, and analysis. Westport, CT: Greenwood.

Stufflebeam, D. L., McCormick, C. H., Brinkerhoff, R. O., & Nelson, C. O. (2012). Conducting educational needs assessments (Vol. 10). Dordrecht, the Netherlands: Springer Science & Business Media.

Tobey, D. D. (2005). Needs assessment basics: A complete, how-to guide to help you: Design effective, on-target training solutions, get support, ensure bottom-line impact. Alexandria, VA: ASTD Press.

Bo Hu Bo Hu Hu, Bo

Network Centrality Network centrality

1139

1140

Network Centrality

The term *network centrality* refers to a measure of the prominence or importance of an individual actor within a social network. A social network can be defined as a network formed by a set of interacting social entities (actors) and the linkages (relations) among them. The definition of centrality was first developed in the late 1940s and early 1950s by Alex Bavelas and Harold Leavitt when studying communication structure. From the perspective of social network analysis, the prominent actors are those that are extensively involved in the relationships with other actors. There are several different ways to define the involvement of an actor with regard to the relationships with others. These methods focus on some meaningful centrality indicators, such as degree, closeness, betweenness, information, and the rank of actors. Moreover, an individual actor's centrality indices can be aggregated across a group of actors to represent the centralization of a network. This entry introduces the definition and calculation of network centrality based on three mainly used centrality indicators: degree, closeness, and betweenness.

Degree Centrality

Degree centrality represents the simplest way to define network centrality. *Degree* can be simply interpreted as how active an actor is in a social network. Accordingly, the idea underlying degree centrality is that the most central and prominent actor within a network must be the most active one (i.e., having the most ties to other actors).

Let g be the size of an undirected network with a single, dichotomous relation. Let $C_D(n_i)$ denote the degree centrality of the ith actor, $i = 1, \dots , g$, in the network. Mathematically, the degree centrality of an individual actor can be expressed as

$$C_D(n_i) = \sum_j x_{ij},$$

where x_{ij} is the value representing whether a relation exists between the ith and jth actor for all $i \neq j$.

It can be noted that $C_D(n_i)$ is a function of network size with the maximum value of $g - 1$. In order to compare across networks with different size, $C_D(n_i)$ needs to be standardized by dividing by its maximum value $g - 1$. The standardized degree centrality $C_D(n_i)$ can be expressed as

$$C'_D(n_i) = \frac{C_D(n_i)}{g-1}.$$

In a directed network, where each tie has a direction, degree can be further differentiated between in-degree and out-degree. The former refers to the number of ties directed to an individual actor, and the latter refers to the number of ties that an individual actor directs to the other actors. In-degree is a measure of the popularity and out-degree is a measure of gregariousness. For directed networks, the degree centrality corresponds to out-degree centrality. Let x_{ij+} denote the out-degree of actor i, similar to Equations 1 and 2, and the degree centrality can be expressed as

$$C_D(n_i) = \sum_j x_{ij+},$$

and the standardized degree centrality can be written as

$$C'_D(n_i) = \frac{\sum_j x_{ij+}}{g-1}.$$

Closeness Centrality

Closeness centrality focuses on how close an actor is to all other actors within a network. It can be measured as a function of geodesic distances (i.e., the number of linkages between two actors in a shortest path). Let $d\ (x, y)$ denote a distance function. Then, $d\ (n_i, n_j)$ represents the number of lines in the geodesics linking actors i and j. The total distance from i to all other actors is for all $j \neq i$. Because closeness decreases with the increase of distance, the index of closeness centrality can be simply expressed as the inverse of the total distance:

$$c_c\left(n_i\right) = \left[\sum_{j=1}^{g} d\ \left(n_i, n_j\right)\right]^{-1}.$$

When an actor is adjacently tied to all other actors, the index reaches its maximum value, $(g - 1)^{-1}$, and when one or more actors are isolated and unreachable from other actors, the index attains its minimum value, 0. The standardized closeness centrality, , can be calculated by multiplying $c_c(n_i)$ by $(g - 1)$:

$$C_c^{'}\ \left(n_i\right) = (g-1)\, c_c\left(n_i\right).$$

For a directed network, the index of closeness centrality and its standardized version can be obtained using exactly the same equations (Equations 5 and 6) as for an undirected network. However, it should be noted that in a directed network, $d\ (n_i, n_j)$ may not always equal $d\ (n_j, n_i)$.

Betweenness Centrality

The idea underlying betweenness centrality is that the actor is central and prominent if it serves as an intermediate for extensive indirect linkages between nonadjacent actors. In other words, the prominent actors are those that have control over the interactions among nonadjacent actors. Specifically, an actor is central if it lies between other actors on their geodesics, and therefore, the index of betweenness centrality should capture the involvement of an actor in the geodesics among other actors. Usually, betweenness centrality is presented as a probability.

Let g_{jk} be the number of geodesics linking two actors, j and k. If all these geodesics are equally likely to be chosen as the route, the probability of choosing any one of them is $1/g_{jk}$. Suppose that actor *i* serves as a "messenger" in the communication between the two actors. Let $g_{jk}(n_i)$ denote the number of geodesics that involve actor *i*. Then, the probability of taking the paths containing actor *i* can be obtained by $g_{jk}(n_i)/g_{jk}$. The index of betweenness centrality for the *i*th actor is simply the sum of these probabilities over all pairs of other actors:

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk}.$$

The values of $C_B(n_i)$ range from 0 to $(g - 1)(g - 2)/2$. Obviously, the index reaches 0 when n_i fails to be part of any geodesics. It should be noted that the maximum value of betweenness centrality equals to the number of pairs of actors not including n_i. To obtain the standardized betweenness centrality, needs to be divided by its maximum value $(g - 1)(g - 2)/2$:

$$C'_B(n_i) = C_B(n_i)/\left[\frac{(g-1)(g-2)}{2}\right].$$

In the case of a directed network, the betweenness centrality can be directly calculated using Equation 7. However, to obtain the standardized index, $C_B(n_i)$ needs to be divided by $(g - 1)(g - 2)$; as in a directed network, geodesics linking actor *j*–*k* and geodesics linking *k*–*j* are not identical and need to be treated differently.

Bo Hu

See also Network Cohesion; Network Density

Further Readings

Bavelas, A. (1948). A mathematical model for group structure. Human Organizations, 7, 16–30.

Bavelas, A. (1950). Communication patterns in task-oriented groups. Journal of the Acoustical Society of America, 22, 366–371.

Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31, 581–603.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications: Vol. 8. New York, NY: Cambridge University Press.

Bo Hu Bo Hu Hu, Bo

Network Cohesion Network cohesion

1141

1143

Network Cohesion

In social network analysis, the term *network cohesion* refers to a measure of the connectedness and togetherness among actors within a network. A social network can be defined as a network formed by a set of interacting social entities (actors) and the linkages (relations or edges) among them. The index of network cohesion is a single value that captures the togetherness of the group. Network cohesion can be measured in a variety of different ways, most of which are based on the dyadic cohesion. Dyadic cohesion refers to the closeness between a pair of actors. It should be differentiated from the *closeness centrality*, which measures how close an actor is to all other actors within a network. However, from the angle of cohesion, closeness centrality can be seen as a measure for the actor-level cohesion. This entry introduces several different measures for network cohesion and demonstrates how to calculate each index through examples of both undirected and directed networks.

Measures for Network Cohesion

The simplest way to measure network cohesion is to examine how many ties that a network contains. In this sense, network cohesion can be simply expressed as the sum of all observed edges from a network. This index has a disadvantage in that the sum of edges is dependent on the size of a network. A poorly connected network with more actors may have the same total number of edges as a small cohesive network. For this reason, in order to compare across networks of different size, a standardized index is needed. The standardized index can be obtained by dividing the sum of edges by the maximum possible edges of a network. In social network analysis, this standardized index is also known as *network density*, denoted as *D*. Let *N* be the size of a network, and *E* denote the number of observed edges. Then, the network density for an undirected network can be expressed as

$$D = \frac{2E}{N\ (N-1)},$$

and for a directed network:

$$D = \frac{E}{N\ (N-1)}.$$

Network cohesion can also be measured by the average degree of the network (d). In social network analysis, *degree* can be simply interpreted as how active an actor is in a network and can be measured by the number of ties the actor has to other actors. Thus, the average degree is simply the average number of ties each actor has. For an undirected network,

$$\bar{d} = \frac{2E}{N} = D \times (N-1),$$

and for a directed network,

$$\bar{d} = \frac{E}{N} = D \times (N-1).$$

Actually, both network density and average degree are measures that are directly built upon the dyadic cohesion. Measures for network cohesion can also be built upon the subgroup-level cohesion, namely, structural cohesion. These measures not only consider the number of ties but also the structure and clustering among ties. Sometimes the subgroup is also termed as a component, which refers to the substructure of networks connected internally but disconnected between each other. One of the measures under this umbrella is component ratio (*CR*). Let *C* denote the number of components and *N* denote the number of actors in a network, then:

$$CR = \frac{\mathrm{C}-1}{N-1}.$$

Another component-based measure for network cohesion is fragmentation (*F*). Fragmentation is defined as the proportion of pairs of actors that are not located in the same component. Let r_{ij} be any pair of actors *i* and *j* in a network. If *i* and *j* are observed in the same component, $r_{ij} = 1$; otherwise, $r_{ij} = 0$. Let *N* be the number of actors, and the fragmentation can then be calculated by

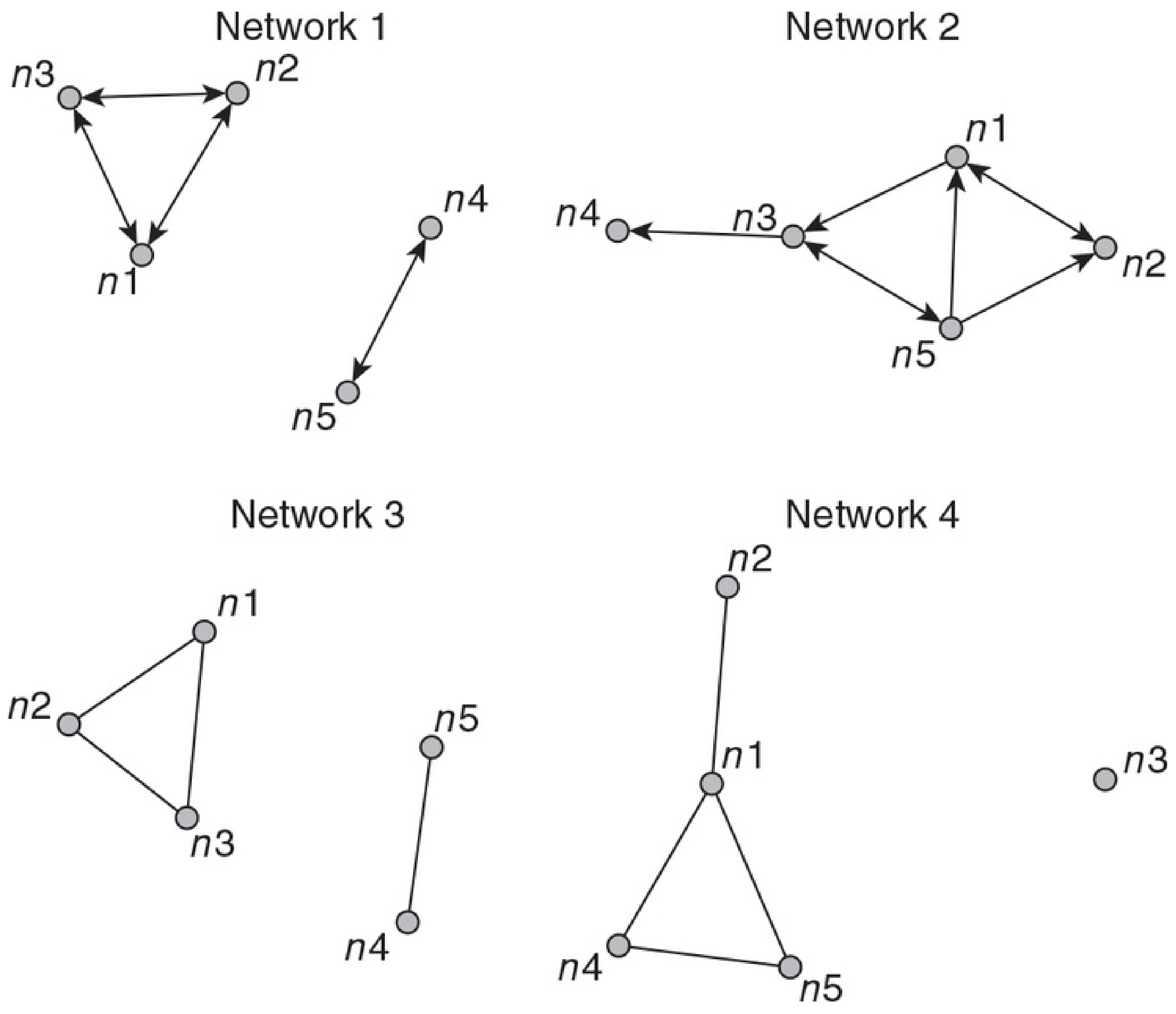
$$F = 1 - \frac{\sum_{i,j} r_{ij}}{N\ (N-1)}.$$

From Equations 5 and 6, it can be noted that both CR and fragmentation are inverse measures of network cohesion. When all actors are located in one component, both indices reach their minimum value of 0, indicating a perfect cohesion. In contrast, when all actors are isolated from each other, making each single actor itself a component, both indices reach their maximum value of 1, suggesting the network is completely disconnected. Also, because both indices are independent of network size, *N*, they are standardized measures and can be compared across networks with different size.

Illustrative Examples

As illustrative examples, Table 1 presents four adjacency matrices (a squared (0,1) matrix with rows and columns labeled by the name of actors) containing the artificial relational data for two directed (Networks 1 and 2) and two undirected networks (Networks 3 and 4). Data in the first two matrices represent the perceived collaboration among five agencies involved in a state-funded community healthy project. Data in the last two matrices represent the working alliance data among five members in a self-directed therapeutic group. Note that the adjacency matrices for undirected networks are symmetric. The network graphs for four networks are shown in Figure 1.

Figure 1 Network graph


Network 1
n3
n2
n1
n4
n5
Network 2
n1
n4
n3
n2
n5
Network 3
n1
n2
n5
n3
n4
Network 4
n2
n1
n3
n4
n5

Network 1

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	1	0	0
$\mathbf{n}_2$	1		1	0	0
$\mathbf{n}_3$	1	1		0	0
$\mathbf{n}_4$	0	0	0		1
$\mathbf{n}_5$	0	0	0	1	

Network 2

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	1	0	0
$\mathbf{n}_2$	1		0	0	0
$\mathbf{n}_3$	0	0		1	1
$\mathbf{n}_4$	0	0	0		0
$\mathbf{n}_5$	1	1	1	0	

Network 3

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	1	0	0
$\mathbf{n}_2$	1		1	0	0
$\mathbf{n}_3$	1	1		0	0
$\mathbf{n}_4$	0	0	0		1
$\mathbf{n}_5$	0	0	0	1	

Network 4

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	0	1	1
$\mathbf{n}_2$	1		0	0	0
$\mathbf{n}_3$	0	0		0	0
$\mathbf{n}_4$	1	0	0		1
$\mathbf{n}_5$	1	0	0	1	

In Network 1, the matrix contains 8 “ones,” representing 8 directional relationships. Therefore, for the first network, N equals 5 and E equals 8. Based on Equations 2 and 4, $D = 0.4$ and $d = 1.6$. Following the same procedure, D and d are 0.4 and 1.6, respectively, for Network 2. According to network density and average degree, there seems to be no difference between the first two networks with respect to cohesion. However, as shown in Figure 1, intuitively, Network 2 seems more cohesive than Network 1, which has a divided structure.

To further examine the difference in cohesion between two networks, one may need to consider the component-based measures. In Network 1, two components can be identified. Component 1 contains n_1, n_2, and n_3 and Component 2 contains n_4 and n_5, whereas all actors in Network 2 are located in only one component. Based on Equation 5, CR is 0.25 for Network 1 and 0 for Network 2. These values indicate that compared to Network 1, every actor in Network 2 is reachable. Also, using Equation 6, the fragmentation can also be calculated. Because there is only one component in Network 2, it is easy to obtain $F = 0$ for Network 2. In Network 1, there are four intracomponent pairs (n_1 vs. n_2, n_1 vs. n_3, n_2 vs. n_3, and n_4 vs. n_5), indicating eight unique intracomponent ties. Thus, and $F = 0.6$.

To calculate the density and average degree of an undirected network, one only needs to consider either the upper triangular or the lower triangular of the adjacency matrix. For example, considering the upper triangular of the third matrix, there are 4 “ones.” Thus, E equals 4 for Network 3. Through Equations 1 and 3, $D = 0.4$ and $d = 1.6$. One can also calculate that the density and average degree are 0.4 and 1.6, respectively, for Network 4. Again, there seems to be no difference between two networks in terms of cohesion. According to the network graphs, it is easy to see both networks have two components. Therefore, they have the same CR, which is 0.25. Also, both networks contain four intracomponent pairs. Based on Equation 6, the fragmentation is 0.8 for both networks.

Bo Hu

See also Network Centrality; Network Density

Further Readings

Borgatt, S. P., Everett, M. G., & Shirey, P. R. (1990). LS sets, lambda sets, and other cohesive subsets. Social Networks, 12, 337–358.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., & Morris, M. (2008). A statnet tutorial. Journal of Statistical Software, 24(9), 1.

Krackhardt, D. (1987). Cognitive social structures. Social Networks, 9, 109–134.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications: Vol. 8. New York, NY: Cambridge University Press.

Bo Hu Bo Hu Hu, Bo

Network Density Network density

1143

1145

Network Density

In social network analysis, the term *network density* refers to a measure of the prevalence of dyadic linkage or direct tie within a social network. A social network can be defined as a network formed by a set of interacting social entities (actors) and the linkages (relations or edges) among them. A dyad, referring to a pair of actors, is the smallest structure of a social network. Through examining the prevalence of dyadic connections, researchers can gain insight into the interaction between actors in a network. Network density is an important attribute and property used to describe a network. Usually, a dyadic relation is numerically coded as a binary variable with 1 and 0 representing the presence and absence of a tie, respectively. The index of network density is expressed as the ratio of observed ties (edges) to all possible pairwise ties in a network. It can be interpreted as the proportion of potential ties that are actually present. This entry traces the development of network analysis and demonstrates the calculation of network density through examples using both directed and undirected networks.

Social Network Analysis

Social network analysis is a quantitative method widely used to investigate a social environment with a focus on the relationship among social entities and the pattern derived from the relationships. Researchers’ interest in modeling the property of pairwise relation in a network can be traced back to Leonhard Euler’s work, *Seven Bridges of Königsberg*, in 1736, which laid the foundation for the graph theory in mathematics. In the 1930s, Gestalt psychologist Jacob Moreno developed sociograms to visualize the social structure of a group of elementary school students. Motivated by the use of sociograms, in the 1940s and 1950s, plenty of analytic techniques and mathematical models were developed to measure network properties, such as reciprocity, mutuality, balance, and transitivity. At the same time, network analysis was intensively used by anthropologists and social psychologists in studying complex society and human communication. During the 1980s, the application of log linear models and Paul Holland and Samuel Leinhardt's p_1 model in network analysis expedited the use of the method. Ever since, social network analysis was extended to modeling nominal, ordinal data, as well as multivariate relational and longitudinal data.

The Calculation of Network Density

In social network analysis, the index of network density is simply defined as the ratio of observed edges to the number of possible edges for a given network. Before calculating network density, it is necessary to differentiate between two types of networks: undirected networks and directed networks. In undirected networks, ties are nondirectional. That is, for each dyadic relation, there is no way to distinguish between the "initiator" and "receiver." In a therapeutic group, the working alliance between the therapist and each individual client is an example of a nondirectional relationship, and the whole therapeutic group can be seen as an undirected network. By contrast, in a directed network, each tie has a direction, orienting from "initiator" to "receiver." For example, a sociogram based on the aggression of a group of high school students is a directed network.

Let N denote the size of a network, which refers to the number of nodes (actors). The number of all possible edges for an undirected network of N size is , and for a directed network, the number is $N\ (N - 1)$. Let E be the observed edges in a network. Then, the network density D for an undirected network can be expressed as

$$D = \frac{2E}{N(N-1)},$$

and for a directed network,

$$D = \frac{E}{N(N-1)}.$$

As illustrative examples, Table 1 presents four adjacency matrices (a squared (0,1) matrix with rows and columns labeled by the name of actors) containing the artificial relational data for two directed (Networks 1 and 2) and two undirected networks (Networks 3 and 4). Data in the first two matrices represent the perceived collaboration among five agencies involved in a state-funded community healthy project. Data in the last two matrices represent the working alliance data among five members in a self-directed therapeutic group. Note that the adjacency matrices for undirected networks are symmetric.

Network 1

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	1	1	1
$\mathbf{n}_2$	1		1	1	1
$\mathbf{n}_3$	1	0		0	0
$\mathbf{n}_4$	0	0	1		1
$\mathbf{n}_5$	1	1	1	0	

Network 2

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	1	0	0
$\mathbf{n}_2$	1		0	0	0
$\mathbf{n}_3$	0	0		0	1
$\mathbf{n}_4$	0	0	0		0
$\mathbf{n}_5$	1	0	0	0	

Network 3

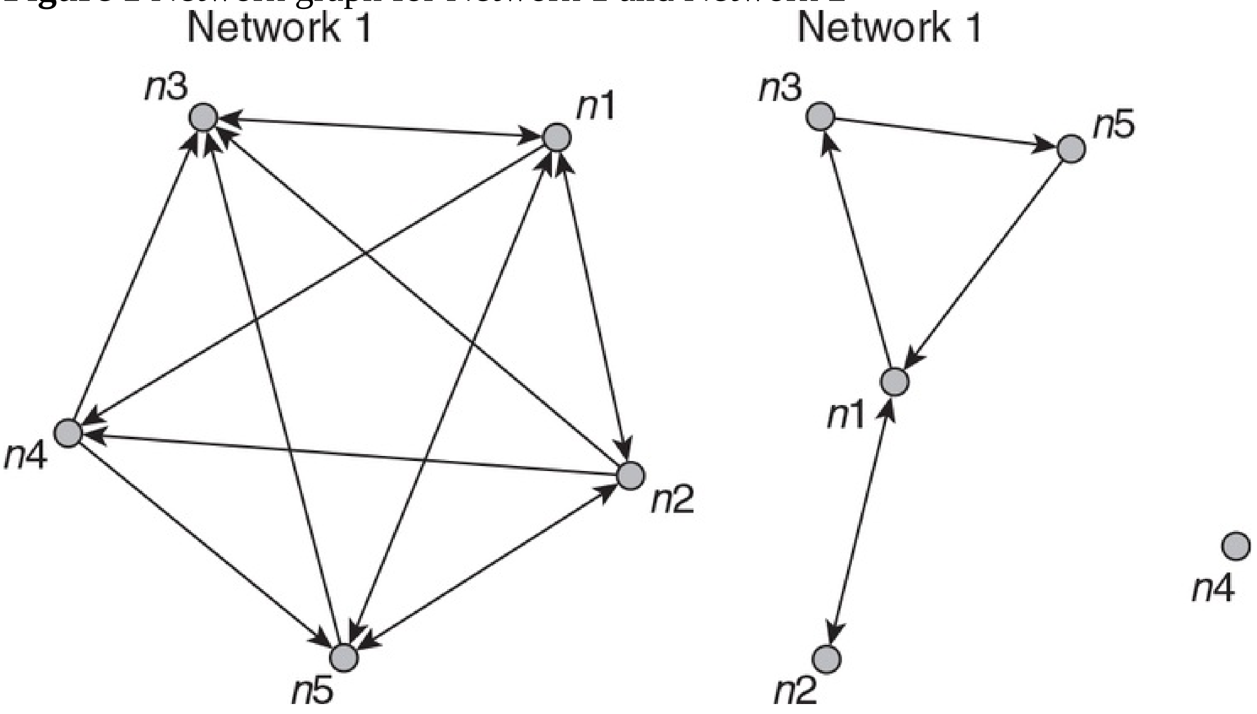
	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	1	1	1
$\mathbf{n}_2$	1		1	0	1
$\mathbf{n}_3$	1	1		1	1
$\mathbf{n}_4$	1	0	1		0
$\mathbf{n}_5$	1	1	1	0	

Network 4

	n_1	n_2	n_3	n_4	n_5
$\mathbf{n}_1$		1	0	0	1
$\mathbf{n}_2$	1		0	0	0
$\mathbf{n}_3$	0	0		0	0
$\mathbf{n}_4$	0	0	0		0
$\mathbf{n}_5$	1	0	0	1	

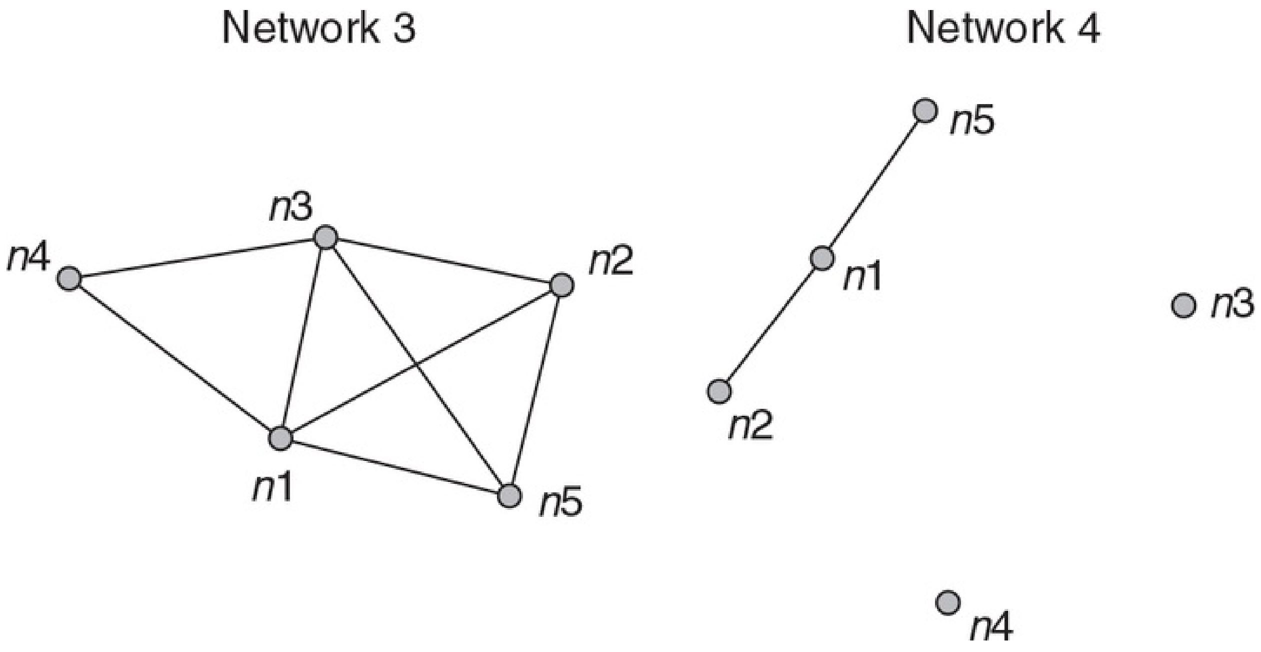
The network graphs for the first two matrices are shown in Figure 1. In Network 1, the matrix contains 14 “ones,” representing 14 directional relationships. Therefore, for the first network, *N* equals 5 and *E* equals 14. Based on Equation 2, *D* = 0.7. Similarly, with five observed ties, the network density for Network 2 is 0.25. These values suggest in Network 1, 70% of potential ties are present, while in Network 2, only 25% of potential ties are present. As shown in Figure 1, with a higher density, Network 1 looks more cohesive and every actor perceives collaborations with most of the other actors. In contrast, Network 2 has a divided structure with one actor being isolated from the others.

Figure 1 Network graph for Network 1 and Network 2



The adjacency matrix for an undirected network is symmetric and redundant. Therefore, to calculate *E* in an undirected network, one only needs to consider either the upper triangular or the lower triangular of the matrix. For example, considering the upper triangular of the third matrix, there are 8 “ones.” Thus, *E* equals 8 for Network 3. Through Equation 1, *D* = 0.8. One can also calculate that the density is 0.2 for Network 4. The network graphs for Networks 3 and 4 are shown in Figure 2. Clearly, Network 4 is poorly connected with only few ties and paths having been built among actors.

Figure 2 Network graph for Network 3 and Network 4



Bo Hu

See also Network Centrality; Network Cohesion

Further Readings

Carrington, P. J., Scott, J., & Wasserman, S. (Eds.). (2005). Models and methods in social network analysis (Vol. 28). New York, NY: Cambridge University Press.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., & Morris, M. (2008). A statnet tutorial. Journal of Statistical Software, 24(9), 1.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association, 76(373), 33–50.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). New York, NY: Cambridge University Press.

Randall Reback Randall Reback Reback, Randall

No Child Left Behind Act No child left behind act

1145

1148

No Child Left Behind Act

The federal government dramatically expanded its role in K–12 education policy with the No Child Left Behind Act of 2001 (NCLB), a reauthorization of the Elementary and Secondary Education Act of 1965. This entry explains the provisions of NCLB, including its requirements for annual student testing and for schools to meet targets for student proficiency in English and math. It then discusses the consequences the law outlined for schools that failed to meet the targets, the debate over the effects of NCLB, the flexibility granted to certain states in meeting the law’s requirements, and the replacement of NCLB with the Every Student Succeeds Act.

In January 2002, NCLB was signed into law by President George W. Bush, after receiving bipartisan support in Congress. The original Elementary and Secondary Education Act established what is now referred to as Title I funding, federal revenues designed to provide additional resources for schools serving students from low-income households. NCLB required states to adopt testing and accountability systems or else states would lose their Title I funding.

State compliance with NCLB required annual testing in mathematics and English-language arts for public school students in Grades 3 through 8 and in at least one high school grade. States were also required to calculate the percentages of students achieving grade-level proficiency on those exams. They were required to calculate both schoolwide proficiency rates and proficiency rates for various subgroups at the school—students from various ethnic subgroups, students from low-income families, students with disabilities, and students with limited English proficiency. Based on these proficiency rates and a few other indicators, states were required to classify schools as failing to meet adequate yearly progress (AYP) if proficiency rates were less than targeted levels. NCLB required states to administer baseline standardized exams to students during the 2001–2002 school year and to begin annual AYP determinations starting in the 2002–2003 school year.

Specific Determinants of Meeting AYP

Proficiency rate targets required for schools to make AYP were set for 2002–2003 based on each state's own baseline exam proficiency rates. These targets escalated steadily each year, culminating in 100% proficiency rate targets for 2014. A school's actual targets each year were often lower than the stated targets because states could use confidence interval formulas to adjust targets downward based on tested subgroup size—the smaller the tested number of students, the more generous the downward adjustment of proficiency rate target.

Along with student proficiency rates, AYP determination was based on student test-taking participation rates and one other indicator—typically student attendance rates for elementary and middle schools and student graduation rates for high schools. Making AYP required adequate performance for every single indicator; a school could fail to make AYP if it failed only one performance indicator in either mathematics or English-language arts.

The lowest performance indicator ultimately determined a school's AYP status, with two exceptions. First, states were allowed to use a "safe harbor" formula in the case that only one student subgroup's performance would cause the school to fail to make AYP; if that subgroup showed sufficient improvement from its performance the prior year, then the school could make AYP through safe harbor. Second, schools were not held accountable for a subgroup's performance if the number of students in this subgroup taking the test fell below a prespecified level.

States had discretion along several dimensions of determining accountability rules under NCLB. States could choose the standardized exams, the dates on which the exams were administered, and the definition of grade-level proficiency on the exams. States could also choose formulas for confidence intervals, safe harbor rules, and minimum subgroup size rules. States could determine whether proficiency rate targets were uniform across subjects and grade levels and whether to disaggregate student proficiency rates by grade level. This wide discretion, especially the use of different formulas, created large cross-state variation in the fraction of schools making AYP—variation that did not correspond with external measures of student performance such as statewide proficiency rates on the National Assessment of Educational Progress.

Consequences for Schools Failing to Meet AYP

For public schools receiving Title I funds (Title I schools), failing to make AYP had formal consequences under NCLB. First, the state would place a Title I school failing to make AYP in two consecutive years on a list of schools in need of improvement. The school district had to allow students attending these schools to transfer to another public school inside the same district that was not on the list of schools in need of improvement. School districts could limit open spots due to schools’ capacity constraints, so intradistrict public school choice as a result of this provision expanded unevenly across the country.

If a school failed to make AYP for three consecutive years, the school district had to pay for “supplemental educational services” for any of that school’s students from Title I-eligible (low-income) households who took part in these services. These supplemental educational services were primarily tutoring programs, often operated by private, for-profit organizations. School districts could also run their own tutoring programs to retain some of the revenues and directly serve students.

If a school failed to make AYP for 4 years, it had to take “corrective action” such as replacing certain staff members or introducing a new curriculum, while schools that failed to meet AYP for 5 years had to draft a plan for restructuring that could involve removal of the principal, closing of the school, or conversion from a traditional public school to a charter school.

For all public schools, informal consequences of failing to meet AYP may have included the stigma involved for principals, teachers, and the community. Research studies have found evidence that school accountability grades can influence households’ residential choices and affect property values. The use of the word “failing” as in “failing to make AYP” may have heightened parents’ concerns about these schools; NCLB-era research studies have found effects of accountability pressure on teacher turnover rates, the types of teachers placed in the high-stakes tested grade levels, and teachers’ concerns about their job security.

Debating the Positive and Negative Effects of NCLB

Debating the Positive and Negative Effects of NCLB

Critics of NCLB derided the emphasis on student proficiency rates rather than broader measures of student learning or measures that track the performance of the same students over time. Critics also noted that NCLB held more diverse schools to a wider range of performance standards than schools with more homogeneous student populations. Some critics questioned the use of the word *failing* and raised concerns that it may undermine educators' sense of professionalism, especially if educators felt they had little control over whether the school would receive a failing designation.

Teachers have voiced concerns about their principals managing them in a way that is overly focused on test preparation, high-stakes subjects, and the students who are on the margin for reaching proficiency that year. Studies have confirmed that the greater focus on mathematics and English-language arts came at the cost of reduced instructional time for subjects such as science and social studies.

Supporters of NCLB, or of school accountability policies more generally, point toward several positive effects. Academic achievement gaps across ethnic groups generally narrowed during the NCLB era. Research studies also reveal that student performance on external measures of mathematics and English-language arts skills either stayed the same or improved for schools facing the largest additional accountability pressure from NCLB. The increased focus on mathematics and English-language arts and on historically low-performing student groups yielded some desirable outcomes, and the greater time spent on test preparation may have helped to align curriculum with state standards.

NCLB Waivers

Most schools were unable to make a steady climb toward 100% student proficiency rates by 2014. As states became concerned with the growing number of schools failing to make AYP, the U.S. Department of Education offered the possibility of deviating from some of the rules of NCLB.

Beginning in 2012, under the administration of President Barack Obama, the U.S. Department of Education began granting states formal waivers from some provisions of NCLB. States had to apply for these temporary waivers, and the U.S. Department of Education reviewed the applications based on the alternative forms of evaluation metrics proposed and on the presence of other educational forms of evaluation metrics proposed and on the presence of other educational policies championed by the Obama administration. By 2014, the vast majority of states had received NCLB waivers from the U.S. Department of Education.

Every Student Succeeds Act

In December 2015, President Obama signed the Every Student Succeeds Act, replacing the NCLB. The Every Student Succeeds Act preserved annual testing requirements from NCLB but expanded the types of performance measures that states would have to use in their accountability systems. Rather than identifying the schools that fail to make AYP, Every Student Succeeds requires states to identify a minimum of 5% of schools as being in need of improvement.

States have wide discretion in how to weight the various performance measures, which must include both student proficiency rates and measures of student performance other than test scores. Unlike NCLB, states may design accountability programs that incorporate measures showing changes in performance over many years of time. States may also choose to design school evaluation systems that highlight areas of relative strengths and relative weaknesses for all schools, rather than simply identifying which schools are most in need of improvement.

Randall Reback

See also Adequate Yearly Progress; Every Student Succeeds Act; Family Educational Rights and Privacy Act; Federally Sponsored Research and Programs; Individuals with Disabilities Education Act; National Assessment of Educational Progress; Policy Evaluation; Standardized Tests; Standards-Based Assessment

Further Readings

Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. Educational Researcher, 44(6), 347–358.

Dee, T., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. Education Evaluation and Policy Analysis, 35(20), 252–279.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., & Barney, H. (2007). Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/content/dam/rand/pubs/monographs/2007/RAND_MG58

Kane, T. J., & Staiger, D. (2003). Unintended consequences of racial subgroup rules. In P. E. Peterson & M. R. West (Eds.), No Child Left Behind? The politics and practice of accountability (pp. 152–176). Washington, DC: Brookings Institution Press.

Manna, P. (2010). Collision course: Federal education policy meets state and local realities. Washington, DC: CQ Press.

Polikoff, M., McEachin, A., Wrabel, S., & Duque, M. (2014). The waive of the future: School accountability in the waiver era. Educational Researcher, 43, 45–54.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. American Economic Journal: Economic Policy, 6(3), 207–241.

Susan Prion Susan Prion Prion, Susan

Nominal-Level Measurement Nominal-level measurement

Nominal-Level Measurement

Nominal data are one of the four levels of measurement described in 1946 by S. S. Stevens, a Harvard psychologist. The four levels are nominal, ordinal, ratio, and interval data, and all have specific definitions of their characteristics. It is important to identify the type of data being collected in a research study, so that the correct type of statistical analysis is performed. This entry describes the unique characteristics of nominal data and outlines the data analysis techniques permissible to use with these results.

Nominal data are considered the most crude or simplest of the four levels of measurement. Nominal data are also called categorical, labeled, or nonranked information because the value given functions only to delineate each individual result and to allow the researcher to place similar values into categories. Nominal data refers to a discrete type of information, in which the results are neither measured nor ordered, but simply allocated into distinct categories according to some sort of arbitrary organizing scheme. One category is not considered to be higher or lower than the others. Nominal scales are considered qualitative classifications and are not treated as continuous.

Nominal data are classified as discrete and are analyzed using the binomial class of statistical tests. Nominal data have three characteristics that differentiate them from ordinal, ratio, and interval data: (1) There is no ordering of the different categories, (2) there is no measure of distance between values, and (3) the categories can be listed in any order without influencing the relationship between and among the categories.

It is not possible to conduct arithmetic, statistical, or logical operations on nominal data because the numerical value has meaning only as an identifier rather than an integer. A person’s home address is an example of a number that rather than an integer. A person's home address is an example of a number that functions only as a nominal value. A street number of "100 Grove" carries no significance except to identify the building that has been designated "100 Grove Street." The number "100" does not declare that this home is 100 feet, or 100 miles, from a clear landmark or that the home is 100 times more comfortable than other homes in the area. The number "100" is simply a way to identify the specific home, much as we might use "green house with big tree near the park" to identify a specific home.

It is possible to measure the number of occurrences in each nominal category and calculate a frequency count for that category. Some nominal variables are dichotomous or binary, defined as only two categories or levels. Examples of dichotomous nominal results include such things as a surgical outcome (dead or alive), a smoker (yes or no), an epidemiological status (healthy or ill), or a functioning status (on or off).

Nominal data are important for researchers, as they often provide key descriptive information about the subjects and their features. Common measure nominal variables include gender, race, ethnicity, marital status, nationality, language, biological species, and religious preference.

Clustering is a data mining technique that has been used successfully with nominal variables. Clustering is the grouping of a set of values in such a way that those in the same cluster are homogeneous or similar to each other. By the same reasoning, objects that belong to different clusters or categories are dissimilar to each other—a phenomenon known as separation—and the distance or dissimilarity measure between clusters can be calculated. Distance calculations such as simple matching—Russell–Rao, Jaccard, Dice, Rogers–Tanimoto, and Kulczynski—can be completed. Distance measures such as Yule, Sokal-Sneath-c, and Hamann can be calculated for binary data.

A discreet dependent variable or scale using nominal-level measurement should be analyzed with a binomial or parametric class of a statistical test. Examples would include χ^2 and logistic regression. When the independent and dependent variables are both discrete, χ^2, logistic regression, π, and Cramer's V can be used to analyze the results. Other possibilities for data analysis include cross tabulations, frequencies, proportions, correspondence analysis, multiple classification analysis, Wilcoxin's two-sample test, and the Kolmogorov–Smirnov test. The mode is the only measure of central tendency that is applicable to nominal data because it is simply a set of frequency counts.

There is some controversy about the original levels of measurement as initially defined by Stevens. Critics contend that the four types of results neither correspond to the characteristics of real data that lead to a robust statistical analysis nor give researchers a mutually exclusive classification system that works for current methods of data analysis. Ultimately, the type of measurement scale is not so much an attribute of the data collected but is dependent on the research questions being asked. For researchers and educators, selecting the correct statistical test depends not only on the measurement scale of data but on the type of variables and the ultimate purpose of the analysis.

Susan Prion

See also Interval-Level Measurement; Levels of Measurement; Ordinal-Level Measurement

Further Readings

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., & Cheney, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. Statistics in Medicine, 9(9), 1103–1115.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677–680.

Velleman, P., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading for classifying statistical methodology. The American Statistician, 47, 65–72.

Grant B. Morgan Grant B. Morgan Morgan, Grant B.

Rachel L. Renbarger Rachel L. Renbarger Renbarger, Rachel L.

Nonexperimental Designs Nonexperimental designs

1149

1151

Nonexperimental Designs

Nonexperimental designs are those research designs that lack manipulation of an independent variable and/or control of nuisance variables through random assignment into control or treatment groups. As a result, cause-and-effect relationships cannot be inferred from nonexperimental designs. This entry describes nonexperimental designs, discusses the reasons they are used, and gives recommendations for dealing with confounding variables.

Despite experimental designs typically being considered the gold standard within the context of science, researchers may opt for a nonexperimental design when manipulation of an independent variable and/or random assignment of study participants into groups is not possible, feasible, ethical, or of interest to the researcher. Nonexperimental designs do not allow for causal inferences to be made from relationships observed between variables. This should not, by itself, relegate nonexperimental designs to being weaker, less rigorous, or incapable of making important contributions. They frequently serve a different purpose than experimental designs or are necessary given ethical, feasibility, or practicality constraints.

Common nonexperimental designs that use qualitative methodologies are case study, ethnography, and grounded theory. Common nonexperimental designs that use quantitative methodologies are comparative, survey research, and retrospective. It should be noted that there are so-called quasi-experimental designs that are not considered “true” experiments due to the lack of random assignment (with or without matching) but are still often considered strong designs within the scientific community. These designs include time series, regression discontinuity, and comparative design in which propensity score regression discontinuity, and comparative design in which propensity score matching/analysis is incorporated.

Within the scientific framework, the importance of manipulation and random assignment to experimentation is clearly evident given that their absence is the distinguishing feature between designs classified as experimental and nonexperimental. As noted, these may be beyond the researcher’s ability, interest, or ethics. The remainder of this section discusses scenarios in which a nonexperimental design would be chosen.

First, a researcher may wish to make meaning and organize concepts from an extended series of observations to produce a theory, grounded in observation and experience, about how some social phenomenon might exist or occur. The design for such a study would likely involve purposeful sample selection and extended and extensive data collection via group or individual interviews, naturalistic observation, and/or review of artifacts (e.g., documents and archives) in order to develop a framework (i.e., theory) about the nature of a social phenomenon. The theory might later be subjected to additional investigation. This design involves neither manipulation of any variable(s) nor does it involve random assignment of participants into a control group or one or more treatment groups. As a result, cause-and-effect relationships cannot be inferred; however, potential explanations can be proposed and investigated further in future research.

Second, a researcher may wish to study the effect of being a victim of assault during college on degree completion. To carry out such a study, it would be necessary to collect, at minimum, degree completion data for students who had experienced assault. This approach would not allow the researcher to compare the degree completion rates of those students who had not been assaulted. An alternative approach would be to collect the degree completion rates from students who had experienced assault versus those who had not. An experimental design would require the researcher to randomly assign students to one of the assault conditions. Such a scenario is neither possible nor ethical; thus, a nonexperimental design would be necessary to examine the relationship between these variables.

Recommendations for Dealing With Confounding Variables

Readers will likely

Readers will likely encounter recommendations in research design texts on ways of “strengthening” studies using nonexperimental designs. Such recommendations should be recognized as being aligned with the perspective that experimental designs are stronger than nonexperimental designs. It is true that when causal inference is the goal of a research study, the experimental design is the strongest design toward that end. The recommendations for strengthening a design should be situated within the causal inference framework. This is not meant to imply that nonexperimental designs are “weak”; they simply do not allow causal inferences.

Perhaps a better description of the recommendations for strengthening a nonexperimental design would be that they are recommendations for addressing confounding variables. These recommendations include matching (also called blocking), holding confounding variables constant, or including extraneous variables in the statistical model. These methods are attempts to mimic the randomization mechanism by limiting or partitioning a portion of the variability in the outcome so the relationship of interest is made more precise, in a sense.

Each of the recommendations described in this section is based on the following scenario: Suppose previous reading achievement is theorized or hypothesized to influence the effectiveness of a reading intervention on an end-of-grade reading test, and the intervention is administered within select intact classrooms at a school. Each method attempts to control for the effect of the confounding variable, previous reading achievement, without using random assignment of students into intervention groups. If random assignment were used, any differences in previous reading achievement between the groups would be due to randomness and would thus most likely be roughly equivalent across groups.

Matching

Matching on variables that are (1) of theoretical importance and (2) available to the researcher can improve the understanding of relationships between variables, even in experimental designs. In the reading intervention scenario just described, the researcher may choose to match each student who received the intervention with a student who did not receive the intervention on the previous year’s end-of-grade test score before making a comparison between the sets of students.

Holding Confounding Variable Constant

Holding one or more confounding variables constant removes the variability due to the confounding variable completely. In the same reading intervention scenario, the researcher could compare reading scores of students who received the reading intervention and the scores of those who did not among only the students who did not meet the reading proficiency standard on previous year's end-of-grade test.

Including Confounding Variable in Statistical Model

The confounding variable(s) can also be added in a statistical model, such as a multiple regression model, so that the variability in the confounding variable(s) can be accounted for. In the reading achievement scenario, the researcher could add the previous year's end-of-grade test and an indicator of whether each student received the intervention into a regression model that predicts the end-of-grade test scores in the current year.

Grant B. Morgan and Rachel L. Renbarger

See also Causal Inference; Correlation; Experimental Designs; Propensity Scores; Qualitative Research Methods; Regression Discontinuity Analysis; Time Series Analysis

Further Readings

Kirk, R. E. (2013). Experimental design: Procedures for the behavioral sciences (4th ed.). Thousand Oaks, CA: SAGE.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Harcourt.

Trochim, W. M. K., & Donnelly, J. P. (2008). The research methods knowledge base (3rd ed.). Mason, OH: Atomic Dog.

Parvati Krishnamurty Parvati Krishnamurty Krishnamurty, Parvati

Nonresponse Bias

Nonresponse bias

1151

1153

Nonresponse Bias

Nonresponse occurs when all the sampling units selected for a sample in a survey are not interviewed. A sampling unit could be an individual, household, business, or other entity being interviewed. The term *nonresponse bias* refers to the potential bias that can occur in surveys due to nonresponse. Particularly since the start of the 21st century, surveys based on probability samples have been experiencing declining response rates, and therefore nonresponse bias has become a growing concern for surveys in all fields. Nonresponse bias can impact education surveys as well as educational measurement and research based on surveys. If survey participants are systematically different from nonparticipants on measures related to the study, then the accuracy of the estimates, analysis, and inferences from the survey results will be affected. This entry describes the definition, identification, and measurement of nonresponse bias and describes the techniques used to adjust for it. It concludes with a list of resources for further reading on nonresponse bias.

A famous example of nonresponse bias is from the 1936 presidential election in which Democrat Franklin D. Roosevelt and Republican Alfred Landon were the two candidates. *The Literary Digest* voter survey predicted that Landon would beat Roosevelt. The prediction was based on only 2.4 million responses from a total of 10 million mail-in questionnaires (a 24% response rate). Poll results indicated that Landon would win a majority and Roosevelt was expected to get only 43% of the vote. Actually, Roosevelt won the election with 62% of the vote. Nonresponse bias was one of the reasons for this error because respondents tended to be Landon supporters and nonrespondents tended to support Roosevelt. An additional reason for the error was sampling bias due to the undercoverage of low-income voters who tended to be Democrats.

Defining Nonresponse Bias

Bias is the difference between a survey estimate and the actual value in the target population. There are two components of nonresponse bias associated with an estimate—the amount of nonresponse and the difference in the estimate between the respondents and nonrespondents. The nonresponse bias of the mean can be described by the following expression in which *Y* is the measure of interest:

$$\text{Bias} = mn\left(\overline{Y}_r - \overline{Y}_m\right),$$

where Bias = the nonresponse bias of the respondent mean; = the mean of the respondents in a sample of the target population; = the mean of the nonrespondents in the target population; *m* = the number of nonrespondents in the target population; and *n* = the total number in the target population.

An alternative approach to measuring nonresponse bias considers every potential respondent to have a propensity (or probability) of participation in the survey. Response propensity can be affected by various factors including demographic characteristics, employment status, sponsorship of the survey, or the length (or burden) of the questionnaire.

This propensity of participation is denoted by ρ, and nonresponse bias is expressed as

$$\text{Bias} = \sigma_{y\rho}/\rho,$$

where $\sigma_{y\rho}$ is the covariance between the measure (*y*) and response propensity (ρ), and ρ is the mean response propensity of the sample.

The covariance ($\sigma_{y\rho}$) is the product of the correlation of *y* and ρ, the standard deviation of *y* and the standard deviation of ρ. Both definitions of nonresponse bias assume that there is no other source of bias in the measure such as measurement error.

Response Rates and Nonresponse Bias

Nonresponse bias is a potential consequence of low response rates. However, Nonresponse bias is a potential consequence of low response rates. However, using the response rate as a measure of data quality can be misleading, as low response rates do not always cause nonresponse bias. Response rates are measured at the level of the entire survey, but nonresponse bias can affect different measures or statistics differently. This is because nonresponse bias is a function of the response rate and the difference between respondents and nonrespondents on the specific measure. So if the difference between respondents and nonrespondents is small, high nonresponse could lead to low nonresponse bias.

Although the response rate is a useful indicator of data quality, it is not directly related to nonresponse bias, and it is survey specific rather than estimate specific. Several indicators have been developed as alternatives to the response rate to take into account the complexity of measuring nonresponse bias and its specificity to the estimate or model. These include representativity indicators (*R* indicators), balance indicators (*B* indicators), and various estimate-level indicators.

Identifying and Adjusting for Nonresponse Bias

There are several approaches to identifying and studying nonresponse bias. These include the following:

- *Comparing survey estimates with external data*. Survey estimates can be benchmarked against estimates from large national surveys like the Current Population Survey or administrative data to assess nonresponse bias. Bias is indicated by differences between the survey estimates and external data.
- *Studying variation within the survey*. Data from earlier rounds of a survey, screener interviews, or level of effort data (such as number of calls made, early or late response) may be available and can be used to study nonresponse bias. Nonresponse follow-up, which usually involves interviewing a sample of nonrespondents to gather data on selected variables, is another method of assessing nonresponse bias. Such surveys can be expensive because nonrespondents can be hostile or difficult to reach, and high response rates are needed in the follow-up survey.
- *Comparing alternative postsurvey adjustments for nonresponse bias*. Several sophisticated statistical techniques are used to adjust estimates for nonresponse bias, including weighting class adjustments, raking, calibration methods, propensity models, and post stratification. The results from these

techniques can be compared to assess how well they correct for nonresponse bias.

- *Modeling of response based on variables available in the sample*. Respondent and nonrespondent characteristics from the sampling frame (the list from which the sample is drawn) can be combined into a model to assess the extent of nonresponse bias. In some surveys, interviewer-observed data on housing type or neighborhood is recorded during data collection. This data can later be used to judge whether nonrespondents and respondents differ on the basis of interviewer-observed characteristics.

In practice, considerable thought needs to be given to potential response rates and nonresponse bias at the design stage of a survey. Different modes, frames, and survey designs allow for different kinds of nonresponse studies and postsurvey adjustments. Additional features of a survey such as advance letters, incentives, or number of callbacks allowed are often used as strategies to increase response rates and may impact nonresponse bias.

Parvati Krishnamurty

See also Correlation; Response Rate; Sample Size; Standard Deviation

Further Readings

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly, 70(5), 646–675.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). Survey nonresponse. New York, NY: Wiley.

Tourangeau, R., & Plewes, T. (Eds.). (2013). Nonresponse in social science surveys: A research agenda. Washington, DC: The National Academies Press. Retrieved from www.nap.edu

Craig A. Mertler Craig A. Mertler Mertler, Craig A.

Normal Curve Equivalent Score Normal curve equivalent score

1153

1154

Normal Curve Equivalent Score

Normal curve equivalent scores (NCE scores) are one type of normalized standard, norm-referenced scores. Norm-referenced scores report the results of standardized assessments and other instruments in a way that permits the comparison of an individual's performance with a very well-defined norm group of similar individuals who have completed the same assessment. Publishers of norm-referenced assessments will typically transform scores so that they can be placed along a common distribution. This common distribution is called a normal distribution, normal curve, or bell-shaped curve.

There are numerous types of scores that are transformed for ease of interpretation in relation to a norm group. These include linear standardized scores (e.g., *Z* scores and *T* scores), which report how far a raw score is located from the mean score of a norm group, reported in standard deviation units. A distribution of linear standard scores will typically retain the same shape as the distribution of raw scores (i.e., *not* a normal distribution). A disadvantage of these types of scores is that they are often misinterpreted due to the nature of their respective scales (i.e., the majority of *Z* scores range from −3.00 to +3.00 and the majority of *T* scores range from 20 to 80). This problem can be overcome through a transformation of raw scores to normalized standard scores. Transformation to any normalized standard score type will, in essence, make the distribution conform to a normal distribution. Examples of these types of transformed scores are stanines, SAT scores, deviation IQ scores, and NCE scores.

NCE scores have a mean of 50 and a standard deviation of 21.06. NCE scores range from 1 to 99. The somewhat "odd" value for the standard deviation has been established so that NCE scores will precisely align with percentile ranks at three specific points: at scores of 1, 50, and 99. The basic advantage of NCE scores is that they represent equal units across the entire continuum (i.e., 1–99), unlike percentile ranks. NCE scores are calculated in the following manner:

1. A *Z* score is calculated (from the obtained raw score) and multiplied by 21.06 (the “new” value for a standard deviation).
2. The value of 50 (the “new” value for the mean) is added to the resulting value in order to obtain the NCE score.

This can be expressed with the following formula:

$$\text{NCE} = 21.06z + 50$$

NCE scores, along with other normalized standard scores, allow for ease of interpretation of scores that have been transformed to the same normal distribution scale. Regardless of the specific score being used, all normal standardized score scales provide the same information about a particular individual’s performance.

Craig A. Mertler

See also Ability Tests; Achievement Tests, ACT; Aptitude Tests; Areas Under the Normal Curve; High-Stakes Tests; Iowa Test of Basic Skills; Normal Distribution; Norming; Percentile Rank; Reading Comprehension Assessment; SAT; Standardized Scores; Standardized Tests; Stanines; *T* Scores; *Z* Scores

Further Readings

McMillan, J. H. (2014). Classroom assessment: Principles and practice for effective standards-based instruction (6th ed.). Boston, MA: Pearson.

Mertler, C. A. (2007). Interpreting standardized test scores: Strategies for data-driven instructional decision making. Thousand Oaks, CA: SAGE.

Popham, W. J. (2011). Classroom assessment: What teachers need to know (6th ed.). Boston, MA: Pearson.

Waugh, C. K., & Gronlund, N. E. (2013). Assessment of student achievement

(10th ed.). Boston, MA: Pearson.

Dorothy J. Musselwhite Dorothy J. Musselwhite Musselwhite, Dorothy J.

Brian C. Wesolowski Brian C. Wesolowski Wesolowski, Brian C.

Normal Distribution

Normal distribution

1154

1157

Normal Distribution

The normal distribution is a hypothetical symmetrical distribution used to make comparisons among scores or to make other kinds of statistical decisions. The shape of this distribution is often referred to as “bell shaped” or colloquially called the “bell curve.” This shape implies that the majority of scores lie close to the center of the distribution, and as scores drift from the center, their frequency decreases.

Normal distributions belong to the family of continuous probability distributions or probability density functions. A probability density function is a function meant to communicate the likelihood of a random variable to assume a given value. This function is graphed by plotting the variable, *x*, by the probability of that variable occurring, *y*. These normal probability distributions are characterized by the aforementioned symmetric bell shape but can have any real mean, labeled μ, and any positive real standard deviation, labeled σ. Specifically, the normal distribution is characterized by continuous data, meaning the data can occupy any range of values. In special cases, the normal distribution can be standardized in which the mean becomes 0 and the standard deviation becomes 1. All normal distributions can be transformed or standardized to the standard normal distribution.

The normal distribution is commonly named the “Gaussian distribution,” after Carl Friedrich Gauss, a German mathematician who made significant advancements of statistical concepts. Less frequently, the normal distribution may be called the “Laplace distribution,” after Pierre-Simon Laplace. The remainder of this entry reviews the history of normal distribution, explains the defining function of normal distribution, explores its properties, highlights the differences between normal distribution and standard normal distribution, and reviews assumptions and tests of normality.

History

The first affiliation with normal distribution stemmed from errors of measurement. Galileo Galilei looked specifically within astronomy to notice that the errors in observations were not random. Small errors far outweighed the larger errors, and these errors had a tendency to be symmetrically distributed around a peak value.

In 1895, Karl Pearson is credited with the first appearance of the term *normal distribution* from his seminal paper. However, the term also appeared in work by Charles Peirce in 1783, Francis Galton in 1889, and Henri Poincaré in 1893. The first mathematical derivation of the normal distribution is attributed to Abraham DeMoivre in his *Approximatio ad summam terminorum binomii* $(a + b)^n$ *in seriem expansi*. DeMoivre used integral calculus to estimate a continuous distribution, resulting in a bell-shaped distribution.

In 1808, Robert Adrain, an American mathematician, debated the validity of the normal distribution, expounding on distributions of measurement errors. His discoveries led to further work in proving Adrien-Marie Legendre’s method of least squares. In 1809, without knowledge of Adrain’s work, Gauss published his *Theory of Celestial Movement*. This work presented substantial contributions to the statistics field, including the method of least squares, the maximum likelihood parameter estimation, and the normal distribution. The significance of these contributions is possibly why Gauss is given credit over Adrain in regard to the normal distribution. In use from 1991 to 2001, the German 10 DM banknote displayed a portrait of Gauss and a graphical display of the normal density function.

In the early 1800s, Adolphe Quetelet, Walter Weldon, and Pearson worked to apply the concept of the normal distribution to the biological and social sciences, eventually cofounding the journal *Biometrika*. In 1994, American psychologist Richard Herrnstein and political scientist Charles Murray published *The Bell Curve: Intelligence and Class Structure in American Life*. This publication led to the term *bell curve* becoming a more widely known concept. Herrnstein and Murray looked at the relation between intelligence scores and social outcomes, resulting in implications of an ever-increasing social stratification based on intelligence.

Definition

The normal distribution is constructed using the normal density function:

$$px = e - x - \mu^2 / 2\sigma^2\sigma^2\pi.$$

This exponential function is comprised of a constant (e), the mean (μ), the standard deviation (σ), and the variance (σ^2). The formula is often shortened to $N(\mu, \sigma^2)$. If $N(0, 1)$, so that $\mu = 0$ and $\sigma^2 = 1$, the resulting distribution is the standard normal distribution.

The shape of the normal distribution is based on two parameters: the mean (μ) and the standard deviation (σ). The mean controls the x-axis, and the standard deviation controls the y-axis. The mean influences the location of the apex of the distribution and is therefore called the location parameter. The variance (σ^2) influences how wide the distribution appears and is therefore called the scale parameter. A larger variance will result in a wider bell curve.

Properties

In a normal distribution, the curve is entirely symmetrical around the mean, such that $x = \mu$. This symmetrical distribution shows that with the mean, the median and mode must also coincide. There are more data observations closer to these central tendency values than to the extremes of the bell shape. Along the y-axis, the graph stretches from −∞ to +∞. The normal distribution, however, is a purely hypothetical model. Rarely do observations exist in the world that fit the model perfectly. Rather, scores and distributions come close to the normal distribution. As the sample size increases, the distribution becomes closer to the hypothetical model.

The function approximates the number of observations that should fall within specific areas of the curve. For example, within 1 standard deviation of the mean (+1 and −1), approximately 68.3% of observations should appear here. Looking specifically at 1 standard deviation above the mean, only 34.1% of observations fall between the mean and the value of the mean plus 1 standard deviation above that mean. Roughly, 95.4% of observations should fall within 2 standard deviations from the mean (+2 and −2). Finally, about 99.7% of observations fall within 3 standard deviations from the mean (+3 and −3).

The normal curve can be used to determine the percentage of scores above or below a certain score. Specifically, between the mean and 3 standard deviations above the mean, approximately 50% of observations should occur in this interval (34.13 + 13.59 + 2.14).

Skewness and kurtosis are two ways a distribution can deviate from the normal, idealized shape. In a normal distribution, skew and kurtosis have values of 0. As the distribution deviates further from normal, these values move above and below 0. Skewness refers to a lack of symmetry, where most of the scores are gathered at one end of the scale. Positively skewed refers to a distribution with a cluster of scores at the lower end of the scale with the tail at the higher, more positive end of the scale. Negatively skewed refers to a distribution with the tail at the lower end of the scale, and the cluster of scores positioned at the higher, positive end.

Kurtosis refers to the pointiness of the distribution. More specifically, kurtosis describes the degree to which scores cluster at the ends of the distribution or the tails. Positive kurtosis is a pointier version of the normal distribution, with many scores in the tails. It is often called a heavy-tailed distribution or leptokurtic. Negative kurtosis, often called platykurtic, is flatter than the normal distribution and is relatively thin in the tails.

Standard Normal Distribution

The standard normal distribution is a version of the normal distribution in which the normal random variable has a mean of 0 and a standard deviation of 1. In the standard distributions, the random variables are transformed into *z* scores using the following formula for use with a population:

$$z = (X - \mu)\sigma,$$

where *X* is the normal random variable, μ is the mean of the data, and σ is the standard deviation. The following formula for *z* scores is used with a sample:

$$z=(X-X)s,$$

where X is the normal random variable (or score), X is the mean of the data, and s is the standard deviation. Most frequently in use with a standard distribution is the standard normal distribution table, which dictates cumulative probability based on the z score calculated. This table gives values of the area under each part of the curve at the value of z. The areas are related to probability. Only in a standardized normal distribution does the total area under the curve equal one (1.0). The area above the z score indicates the likelihood of those values occurring, and the area below the z score indicates the likelihood of those values occurring.

Assumptions of Normality

In the interpretation of data, it is important that all evidence is evaluated objectively or free of bias. Outliers can directly affect this interpretation as well as any violation of the assumption of normality. This assumption is one of a varied list of assumptions of statistical tests but relates directly to the normal distribution.

Extreme scores may bias estimates of parameters. Specifically, the mean may be influenced more by outliers than the median. Confidence intervals are based on parameter estimates and therefore are also influenced by the bias of outliers. To achieve accurate confidence intervals, estimates must come from a normal distribution. Null hypothesis significance testing assumes that parameter estimates are normally distributed because other test statistics, such as those from the t test, F test, and χ^2 distribution, are normally distributed. Because populations are often unavailable for testing, significance tests will be accurate when the sampling distribution is normally distributed.

If the sample size is large enough, however, the assumption of normality becomes less of a concern. A larger sample size increases the normality of the distribution and therefore will result in more accurate confidence intervals, significance tests, and estimates of parameters. The definition of “large” will vary from distribution to distribution. The most generally accepted value for sample size is 30, but skew and kurtosis can also impact how large this value should be. Sample sizes upward of 100 may be necessary to achieve a more accurate sampling distribution.

The misunderstanding that occurs most frequently with the assumption of normality is that the data alone need to be normally distributed, which is not the case. The errors, or residuals, of the data should be normally distributed as well as the sampling distribution. However, the raw data are likely to have a varying shape.

Tests of Normality

Many parametric tests are based on the assumption of normality, which assumes the sampling distribution of the population parameter is normally distributed. This assumption does not mean that the sample data being analyzed should be normally distributed.

Two tests of normality exist to compare scores in a sample to a normally distributed set of scores with identical mean and standard deviation. A significant *p* value ($p < .05$) from these tests indicates that the distribution is significantly different from a normal distribution. The Kolmogorov–Smirnov test and the Shapiro–Wilk test both test for this significance. However, the Shapiro–Wilk test has more power to detect differences from normality. Therefore, the Shapiro–Wilk test may have significant values when the Kolmogorov–Smirnov test does not. These tests should be used carefully, as false significance may occur when testing larger samples. Both tests should be used simultaneously with histograms or plots and the aforementioned values of skew and kurtosis.

Dorothy J. Musselwhite and Brian C. Wesolowski

See also Kurtosis; Skewness; Standard Deviation; Standard Error of Measurement; Variance

Further Readings

Boyle, J. D., & Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York, NY: Schirmer Books.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. London, UK: SAGE.

Gravetter, F. J., & Wallnau, L. B. (2011). Essentials of statistics for the behavioral sciences. Belmont, CA: Wadsworth.

Gross, J. (2004). A normal distribution course. Frankfurt, Germany: Peter Lang.

Kubiszyn, T., & Borich, G. (2003). Educational testing and measurement: Classroom application and practice. New York, NY: Wiley.

Payne, D. A. (2003). Applied educational assessment. Toronto, Canada: Wadsworth.

Reid, H. M. (2014). Introduction to statistics: Fundamental concepts and procedures of data analysis. Thousand Oaks, CA: SAGE.

Rachel L. Renbarger Rachel L. Renbarger Renbarger, Rachel L.

Grant B. Morgan Grant B. Morgan Morgan, Grant B.

Norming

Norming

1157

1160

Norming

Norming refers to the process of constructing norms or the typical performance of a group of individuals on a psychological or achievement assessment. Tests that compare an individual's score against the scores of groups are termed *norm-referenced assessments*. These norm-referenced assessments help educational stakeholders such as administrators, teachers, and parents make informed educational decisions about an individual student and the student's progress.

In the field of education, the challenges include accurately representing test populations and interpreting the scores. These challenges involve addressing issues of understanding the test takers, calculating many types of scores based on specific needs, and sampling for representative scores, among others. For this reason, norming requires multiple considerations throughout the steps of the process. This entry first discusses the process of norming, including the selection of norm groups, the procedures used, and sampling of the target population. It then looks at the types of scores obtained from norming and some issues with norming.

Norms and Norm Groups

To make an appropriate comparison, the background characteristics of the norm group and the individual test taker should be similar. The performance of an 18-year-old cannot be compared with the norm group comprising students of 12–15 years. The student's math results might look good compared to that norm, but the score could look poor when compared to a group of engineering professionals. The norm group information (e.g., age, gender, socioeconomic status, location) should appear in the testing manual, so that administrators can make informed decisions about the reliability of the student's performance. To avoid grouping issues, the score of an individual should be compared with similar test takers taking the assessment at the same time.

Often, several groups of individuals are considered the reference groups. These norms can be based upon developmental time points, such as age or grade in school. Geography can influence norms as well. For national norms, test developers often use nationally representative samples. Less frequently, developers or test administrators report local norms of a smaller population to determine the performance of students in other state districts, for example. A controversial norm called race norming involves the comparison based on race or ethnicity. Smaller, defined groups within the larger sample are broadly referred to as subgroup norms. User norms include the performances of the test takers during one time period. Most tests provide only the user norms for the age, grade, geographic, and/or race or ethnicity reference groups.

Procedures

To construct the norms, test developers must define and identify the specific testing population (e.g., students applying to postsecondary institutions) and decide the statistics to be calculated. These decisions will impact the test developers in drawing a sample from the target population. Once the sample takes the psychological assessment, the group statistics are calculated along with standard errors. These statistics describe the performance of the norm group, the individuals used for comparison, representative of the target population. Based on the desired types of normative scores to be reported, test developers then create conversion charts.

To do this, the test sample is divided into appropriate subgroups, such as age subgroups, and percentiles are identified along the range of year or month time points. The norm group scores are commonly assumed to fall within a normal distribution, allowing the conversion based on the mean and standard deviation. The new score distribution must then be smoothed to the curve to avoid irregularities between the percentile points. This conversion table helps exam administrators transform raw scores, the initial score achieved from the assessment, into interpretable data, frequently called derived or scaled scores. If the data are not normally distributed, test developers may transform the distribution into a normal distribution or include additional computations for interpretation.

Sampling

To construct the norms, test developers must select the target population in a systematic way. The two types of sampling are probabilistic and nonprobabilistic. With probability sampling, the participants are selected with some degree of randomness, increasing the likelihood that the sample is representative of the target population. The target population and resources available affect the type of sampling used, and the variability within the target population may be a contributing factor in determining the appropriate sample size.

Four types of probabilistic sampling techniques are simple random, systematic, stratified, and cluster sampling. Simple random sampling involves randomly choosing people within the target population where every member has an equal chance of being selected. Stratified sampling, in which a researcher selects proportions of individuals within certain groups to reflect the larger population, is beneficial to guarantee that certain subgroups are represented in the sample. If the norm group does not contain the correct proportion of certain groups, sampling weights are used when analyzing the data.

In systematic random sampling, developers have a list of all people within the target population and choose every *n*th person. The sampling error depends on how the list is arranged. If the list is randomly ordered, the sampling error is smaller than when it is ordered by a variable, such as name or race. Cluster sampling involves randomly selecting intact groups within the population (e.g., schools, neighborhoods) for testing. These sampling techniques are often used in conjunction to decrease error and remain practical.

Although probabilistic sampling is preferred, this can be difficult for a myriad of reasons. Often, test developers do not have a list of the entire population to randomly select from and therefore cannot choose accordingly. It may also be difficult to accurately identify the composition of the target population ensuring that all groups are represented at the naturally occurring rates. In creating nationally representative samples, countries with a large number of participants or a large geographic area may have additional difficulties in testing the appropriate sample.

With nonprobabilistic samples, norming occurs on samples of convenience. This sampling is typically used when there are few resources available to reach all subgroups within the heterogeneous population. However, because the test was not given to a randomly sampled population, the test developers cannot generalize to the larger population. Bias may also affect the test performance. If norming occurs in a university setting, background variables such as educational level may skew the scores. Due to this concern about representativeness, norming with nonprobabilistic sampling is not common.

Derived Score Types

The most prominent type of derived score is percentiles. Percentiles are the scores below which a certain percentage of the scores fall. A similar idea, percentile rank, refers to the rank of that score compared to the population. For example, a student whose score was at the 75th percentile had a score higher than 74% of the scores. The median, the middle score within the group of scores, indicates the 50th percentile because half of the scores fall below this point and half of the scores fall above. Other common percentiles include quartiles, in which the range of scores is divided into 4ths (at the 25th, 50th, and 75th percentiles), or deciles, in which the range of scores is divided into 10th (10th, 20th, … , 90th).

One disadvantage to using percentiles includes not being able to calculate the average scores for subgroup samples who have taken the test. Another disadvantage deals with the fact that the majority of the scores fall in the middle range of the distribution. When test takers score closer to the ends of the distribution (very high or very low), the percentile ranks begin to increase or decrease dramatically. To overcome possibilities for error, many standardized tests construct percentile bands. These bands include a range of percentiles within which the test taker’s true score should fall.

Standard scores can also be calculated from raw scores or from percentiles to provide for greater differentiation between scores. Standard scores include *Z* and *T* scores, and both of these have a constant mean and standard deviation. These scores are often calculated to be used across measures, measures with different means and distributions. For example, if a student achieved a math test score of 85/100 and a science score of 30/50 questions, standard scores allow researchers to compare the achievement of each measure. *Z* scores range from +3 to −3, with negative values indicating scores below the mean. These numbers align with the standard deviation units.

Z scores present difficulties in that because the mean is 0, half of the scores will be negative. Furthermore, *Z* scores differentiate between scores rather well because of the decimal use. Negative values and decimals can be inconvenient for mathematical reasons. For this reason, *T* scores can be converted from *Z* scores. *T* scores range from *20 to* 80, have a mean of 50, and align with standard deviation units. To calculate either *Z* or *T* scores, the interpreter needs three pieces of information: the raw score, the mean of the norm group, and the standard deviation of the norm group.

Stanines are another standard score. These standard scores range from 1 to 9, with a mean of 5 and a standard deviation of 2. This results in a less specific level of performance for the individual. Stanines are assigned based on the raw score falling on its percentile rank.

Another common reported score is the mean, the average score of the norm group. The mean is often calculated for the entire norm group as well as for each subgroup.

Issues

There are many arguments against using norms as the sole criterion for decision making. One prominent issue deals with the performance levels. No test is completely culture free, meaning that members of different groups respond differently to the test and their scores vary. This has been seen with minority groups and women wherein their scores were lower than those of the norm group, affecting important decisions such as college entry and employment.

Performance differences result when those sampled for testing do not complete the procedures. An example of this is a stratified sample of occupations. Perhaps those who are unemployed have more time to complete the assessment or those from professions with more to gain complete the assessment at higher rates. Additionally, for those who miss the test, makeup testing may not be available. This would result in a nonrepresentative norm group, especially when there are no updated norms. Performance levels stay relatively stable from year to year, but occasional renorming should be completed to provide a current reference group.

There are also questions as to whom to be tested. Certain individuals, such as those with disabilities or who do not speak the same language as the measure, may or may not be included. It is difficult further to define these individuals. Should they include those who speak a basic level of English? How do you know who meets these criteria without any additional measures? Federal regulations protect certain groups of individuals, such as those in prison, and testing them may require extensive procedures. Prior to norming inception, developers must make these important decisions. Ultimately, the decision about who to include in the norming sample will depend on the intended use of and target population for the psychological assessment. In order to make the assessment fairer for all groups, some additional sampling procedures and statistical analyses can be completed.

Rachel L. Renbarger and Grant B. Morgan

See also Achievement Tests; Aptitude Tests; Intelligence Tests; Norm-Referenced Interpretation; Percentile Rank; Standardized Scores

Further Readings

Coaley, K. (2014). An introduction to psychological assessment and psychometrics. Thousand Oaks, CA: SAGE.

Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. Fort Worth, TX: Holt, Rinehart, and Winston.

Furr, R. M., & Bacharach, V. R. (2013). Psychometrics: An introduction. Thousand Oaks, CA: SAGE.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), Educational measurement (pp. 155–186). Westport, CT: Praeger.

Craig A. Mertler Craig A. Mertler Mertler, Craig A.

Norm-Referenced Interpretation Norm-referenced interpretation

1160

1163

Norm-Referenced Interpretation

When norm-referenced standardized tests are administered to students, the results are reported in a way that permits comparisons with a well-defined group (i.e., a *norm group*) of other students who have taken the same assessment. The primary difference between criterion-and norm-referenced scores is that with norm-referenced test scores, interpretations of individual student scores are entirely dependent upon the performance of other students. Norm-referenced tests and their resulting scores provide data that assist educators in answering the following questions:

- What is the relative standing of this student across this broad domain of content?
- How does the student compare to other similar students?

This entry will explore the various types of norm-referenced scores, which include raw scores, percentile ranks, developmental/growth scores, grade equivalent scores, age equivalent scores, and standardized scores, and concludes with a note about interpretation of norm-referenced scores.

Types of Norm-Referenced Scores

There are numerous types of norm-referenced scores. Many of them are based on mathematical transformations. In other words, the raw scores are converted to some other scale. These new scales, then, conform to the characteristics of the normal distribution. It is important to bear in mind that norm-referenced test scores are all based on the notion of how an individual student performs as compared to a large group of similar students. Most of these students will be “average,” with their performance being located near the middle of the distribution.

Raw Scores

Raw scores (i.e., the number of items answered correctly) are the main method of reporting results of criterion-referenced tests and are also provided on norm-referenced test reports. However, these scores are not very useful when interpreting test results for purposes of norm-referenced comparisons. Often, educators need to know how a particular student’s raw score compares to the specific norm group. In order to make these comparisons, raw scores must be converted to another scale. These new scales are referred to as *transformed* or *derived scores* and include percentile ranks, *Z* scores, *T* scores, stanines, normal curve equivalent (NCE) scores, and deviation IQ scores.

Percentile Ranks

A percentile rank is a single number that indicates the percentage of the norm group that scored below a given raw score. Percentile ranks range from 1 to 99. However, because percentile ranks indicate percentages of individuals above and below scores that are normally distributed, they do not represent equal units. Percentile ranks are more compactly arranged in the middle of the normal distribution because that is where the majority of individuals fall.

Consider a hypothetical test report for a student, Emma. Let us assume that Emma correctly answered 34 out of a possible 45 items on a reading subtest. When converted, this raw score converts to a percentile rank of 86. This means that Emma scored higher than 86% of the other students (in the norm group) who took the test. In other words, 86% of those students answered fewer than 34 items correctly.

Developmental/Growth Scores

Developmental scales indicate a student’s development across various levels (e.g., grade or age) of growth. Their purpose is to compare a student’s performance to a series of reference groups that vary with respect to developmental growth.

Grade Equivalent S

Grade Equivalent Scores

A common type of developmental score is the grade equivalent score. A grade equivalent score indicates the grade in the norm group for which a certain raw score was the median performance and is intended to estimate a student's developmental level. Grade equivalent scores are expressed in years and 10ths of years; they consist of two numerical components, separated by a period. The first number indicates the year or grade level, and the second indicates the month during that particular school year, ranging from “0” (i.e., September) to “9” (i.e., June). For example, if a student receives a raw score of 67 on the mathematics portion of an achievement test, this score might be transformed to a grade equivalent score of 4.2. This means that this student's performance corresponds to the performance of a typical student taking the same test in November (i.e., the second month) of fourth grade.

Age Equivalent Scores

Similar to grade equivalent scores, age equivalent scores are based on the average test performances of students at various age levels. Their units are also unequal, meaning that equal age units (e.g., 6 months or 1 year) do not correspond to equal age equivalent score units. Age equivalent scores are also useful for measuring growth in mental ability, reading ability, and other types of characteristics that exhibit fairly consistent patterns within an instructional program. Age equivalent scores are expressed in similar fashion to their grade equivalent counterparts.

Standardized Scores

Percentile ranks and grade equivalent/age equivalent scores exist on scales with unequal units. This characteristic seriously limits the interpretability of each type of score. Standardized scores (or standard scores) are obtained when raw scores are transformed to “fit” a distribution whose characteristics are known and fixed —such as a normal distribution, where the scores are reported in equivalent standard deviation units. As a result of these transformations, scores can be interpreted in a way that is unaffected by the characteristics of a particular test. Regardless of the test, standardized scores efficiently indicate whether a particular score is typical, above average, or below average as compared to others who took the test and also clearly indicate the magnitude of the variation away from the mean score.

Moreover, standardized scores allow for comparisons of test performance across two different measures. For example, suppose you want to compare performances on a reading test (containing 65 items) and a mathematics test (containing 34 items). The mean score on the reading test is 45 and on the math test is 24. Simply comparing raw scores would not tell you very much about a student’s relative standing as compared to the norm group. If Katie received a raw score of 40 (out of 65) on the reading test and a raw score of 30 (out of 34) on the math test, it would be incorrect from a norm-referenced perspective to say that she performed better on the reading test, even though she answered more items correctly. One might notice on the score report that her score of 40 on the reading test is below the average, whereas her score of 30 on the mathematics test is above average. This type of norm-referenced comparison is possible only through the use of standardized scores due to the fact that scores from different subtests are put on the same score scale. Two categories of standard scores are linear standard scores and normalized standard scores.

Linear Standard Scores

Linear standard scores tell how far raw scores are located from the mean of the norm group, expressed in standard deviation units. A distribution of linear standard scores will have the same general shape as the distribution of raw scores from which the standard scores were derived. This is not the case for percentile ranks, grade or age equivalent scores, or nonlinear standard scores. These types of scores are often used to make two distributions (e.g., scores from a science test and a mathematics test) more comparable by placing them on the same scale. Types of linear standard scores include *Z* scores and *T* scores.

Z scores exist on a continuum, where more than 99% of the scores range from −3.00 to +3.00. The sign indicates whether the raw score is above or below the mean; the numerical value indicates how many standard deviations it is located away from the mean. An individual’s *z* score is calculated as follows:

$$z = X - \mu\sigma.$$

The mean of the set of scores is subtracted from the student’s raw score. The resulting value is divided by the standard deviation for the set of scores.

Assume that the administration of a standardized test results in a mean score of

75 and a standard deviation of 8. A student whose raw score is 75 would receive a *Z* score equal to 0 (i.e., zero standard deviation units from the mean). Another student whose raw score is 91 receives a *z* score of +2.00 (i.e., two standard deviations above the mean). Finally, a student who earns a raw score of 63 receives a *z* score of −1.50 (i.e., 1½ standard deviations below the mean). By definition, half the students receive scores below the mean; in other words, they will receive negative *z* scores.

One way to avoid half of the students receiving negative scores is through the use of *T* scores. A *T* score provides the location of a raw score in a distribution that has a mean of 50 and a standard deviation of 10. More than 99% of the *T* scores on a standardized test will range from 20 (three standard deviations below the mean) to 80 (three standard deviations above the mean). A student's *T* score is calculated as follows:

$$T = 10z + 50.$$

A *z* score is calculated, then multiplied by 10 (the value for a standard deviation on the "new" scale). The resulting value is added to 50 (the new value for the mean) to obtain the *T* score.

If we return to the earlier hypothetical example, the first student's *z* score of 0 would equate to a *T* score of 50; the second student (*z* score = +2.00) would have a *T* score of 70; and the third student (*z* score = −1.50) would have a *T* score of 35, thus eliminating the negative scores.

Normalized Standard Scores

Raw scores can also be transformed to scores that are distributed normally, regardless of the shape of the original distribution. This type of transformation actually changes the shape of the distribution by making it conform to a normal distribution. Once the shape has been altered, various types of standard scores can be derived. These derived scores are collectively known as normalized standard scores.

Stanines comprise a common type of normalized score scale used to report norm-referenced performance, but do so by representing a band of scores, as opposed to precise score values. A stanine (short for "standard nine") provides the location of a raw score in a specific segment of the normal distribution. Stanines range in value from 1 (i.e., the extreme low end) to 9 (i.e., the extreme high end), where the mean is equal to 5 and the standard deviation is equal to 2.

The main disadvantage of stanines is that they represent coarse groupings of scores, especially when compared to percentile ranks. However, a stanine is likely a more accurate estimate of the student's achievement because it represents a range within which the student's test performance truly belongs, as opposed to a precise estimate of the student's performance. An individual's stanine score is calculated as follows:

$$\text{Stanine} = 2z + 5.$$

A *z* score is calculated, then multiplied by 2 (the value for a standard deviation on the new scale). The resulting value is then added to 5 (the new value for the mean) to obtain the stanine score.

NCE scores have a mean of 50 and a standard deviation of 21.06. Similar to percentile ranks, NCE scores range from 1 to 99. The "odd" value for the standard deviation has been established so that NCE scores precisely match percentile ranks at three specific scores: 1, 50, and 99. The advantage of NCE scores is that they represent equal units across the entire continuum (i.e., 1–99), unlike percentile ranks. NCE scores are calculated in similar fashion to the scores previously discussed:

$$\text{NCE} = 21.06z + 50.$$

A *z* score is calculated and multiplied by 21.06 (the new value for a standard deviation). The value of 50 (the new value for the mean) is added to obtain the NCE score.

A final type of standardized score, used primarily with assessments of mental ability, is a deviation IQ score. Deviation IQ scores provide the location of a raw score in a normal distribution having a mean of 100 and a standard deviation equal to 15 (or 16, depending on the specific test). For a test with a standard deviation of 15, an individual's deviation IQ score is calculated in the following manner:

$$\text{Deviation IQ} = 15z + 100.$$

A *z* score is first calculated and then multiplied by 15. The value of 100 is added in order to obtain the deviation IQ score.

A Final Note About Norm-Referenced Interpretation

A Final Note About Norm-Referenced Interpretation

All norm-referenced scores provide essentially identical information concerning the location of an individual raw score within a distribution; they simply do so on different scales. It is also important to remember the unequal nature of percentile ranks, as well as the first and ninth stanines, which represent much larger bands than the other stanines. It really does not matter which specific norm-referenced score educators choose to interpret, as they all provide the same information about a particular student’s test performance.

Craig A. Mertler

See also Age Equivalent Scores; Areas Under the Normal Curve; Grade Equivalent Scores; Normal Distribution; Norming; Percentile Rank; Standardized Scores; Standardized Tests; Stanines; *T* Scores; *Z* Scores

Further Readings

Kubiszyn, T., & Borich, G. D. (2016). Educational testing and measurement: Classroom application and practice (11th ed.). Hoboken, NJ: Wiley.

McMillan, J. H. (2014). Classroom assessment: Principles and practice for effective standards-based instruction (6th ed.). Boston, MA: Pearson.

Mertler, C. A. (2007). Interpreting standardized test scores: Strategies for data-driven instructional decision making. Thousand Oaks, CA: SAGE.

Popham, W. J. (2011). Classroom assessment: What teachers need to know (6th ed.). Boston, MA: Pearson.

Waugh, C. K., & Gronlund, N. E. (2013). Assessment of student achievement (10th ed.). Boston, MA: Pearson.

NSF

See National Science Foundation

Dianne Nutwell Irving Dianne Nutwell Irving Irving, Dianne Nutwell

Nuremberg Code

Nuremberg code

1164

1164

Nuremberg Code

The Nuremberg Code is a set of 10 principles intended to satisfy moral, ethical, and legal concerns involving the use of any human subjects in research. An American military tribunal issued the Nuremberg Code in 1947 as part of the judgment in the so-called Doctors' Trial, part of the Nuremberg Trials at the end of World War II. Some of the Nazi doctors and administrators prosecuted in the Doctors' Trial were involved in medical research on concentration camp prisoners.

The code, a set of voluntary guidelines, was written to apply to medical experimentation involving human subjects and focuses on the physical and mental safety of the human subjects. Many of the following principles, which are paraphrased from the code, can apply to other forms of research:

1. Voluntary-informed consent of the human subject is essential;
2. Research should be intended to lead to results for the good of society that cannot be attained through other means;
3. Research should be based on prior animal research and knowledge of the disease or problem being studied;
4. Physical and mental suffering and injury must be avoided;
5. An experiment should not be conducted if there is reason to anticipate that death or a disabling injury will occur;
6. The degree of risk should be no greater than the humanitarian importance of the problem to be solved;
7. Precautions should be taken and facilities provided to protect research subjects against the possibility of injury, disability, or death;

8. An experiment should only be conducted by those who are scientifically qualified;
9. The subject should be able to end his or her participation in the study; and,
10. The scientist should be able to end the study at any stage after determining that continuing is likely to cause injury, disability, or death of the subject.

Under the code, informed consent must be based on legal capacity and cannot involve coercion, so research on children and others not capable of deciding for themselves (e.g., the mentally ill) is prohibited.

The Nuremberg Code informed subsequent international ethics statements. In 1964, the World Medical Association issued the voluntary international Declaration of Helsinki guidelines, which have been revised multiple times since then. Unlike the Nuremberg Code, the Declaration of Helsinki does allow for research on children and others who cannot decide for themselves as long as consent has been obtained from parents or other legal proxies.

Dianne Nutwell Irving

See also Belmont Report; Conflict of Interest; Declaration of Helsinki; Ethical Issues in Educational Research; 45 CFR Part 46; Human Subjects Protections; Informed Consent; Institutional Review Boards

Further Readings

Angell, M. (2015). Medical research: The dangers to the human subjects. The New York Review of Books. Retrieved from http://www.nybooks.com/articles/2015/11/19/medical-research-dangers-human-subjects/

Annas, G., & Grodin, M. (Eds.). (1992). The Nazi doctors and the Nuremberg Code: Human rights in human experimentation. New York, NY: Oxford University Press.

Beecher, H. K. (1966). Ethics and clinical research. The New England Journal of Medicine, 274(24), 1354–1360.

Shuster, E. (1997). Fifty years later: The significance of the Nuremberg code. The New England Journal of Medicine, 337, 1436–1440. Retrieved from https://doi.org/10.1056/NEJM199711133372006

U.S. Department of Health & Human Services. (2016, February 19). Ethical Codes & Research Standards. Retrieved from https://www.hhs.gov/ohrp/international/ethical-codes-and-research-standards/index.html

U.S. Holocaust Memorial Museum. (n.d.). Nuremberg Code. Retrieved from https://www.ushmm.org/information/exhibitions/online-exhibitions/special-focus/doctors-trial/nuremberg-code

U.S. Holocaust Memorial Museum. (n.d.). The Nuremberg Race Laws. Retrieved from https://www.ushmm.org/outreach/en/article.php?ModuleId=10007695

World Medical Association. (n.d.). Declaration of Helsinki. Retrieved from https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Judith Davidson Judith Davidson Davidson, Judith

NVivo

NVivo

1165

1168

NVivo

NVivo is a member of a class of computer software commonly referred to as qualitative data analysis software (QDAS). It was developed to serve the needs of researchers working with nonnumerical, unstructured data, otherwise known as qualitative data. NVivo also integrates with quantitative software tools, making it a valuable tool for mixed methods research. Qualitative research is employed by many educational scholars, program evaluators, and practitioners. Increasingly, QDAS tools like NVivo bring coherence and clarity to studies in this area. This entry provides a detailed explanation of the ways NVivo is structured; information on applications to educational research, measurement, and evaluation; understanding of the historical context; and consideration of the challenges and future for NVivo.

NVivo: Structure

NVivo, like other software tools of this sort, provides the means to organize and conduct analysis of visual, textual, audio, and social media data. Working within NVivo, researchers can exercise flexibility and individuality in research design, as theoretical and methodological decision making remains clearly in the hands of the researcher, not the technology. It is designed to serve the needs of individual researchers or team-based work.

The primary organizational unit in this process is the digital container NVivo provides for each individual project, often referred to as the e-project. The easiest way to imagine how this all works is to think of NVivo as having three major components such as (1) data (sources), (2) indexical features (cases, nodes, and classifications), and (3) interrogational features (links, query or search tools, visualization tools, and memos). Simply stated, NVivo helps researchers to organize their data sources using a range of indexical features. Then, sources and indexes are mixed, matched, juxtaposed, and repositioned to develop new meaning using the range of interrogational features provided by NVivo.

Data: Sources

Within the e-project, a researcher can store all data, digital artifacts, and the digitalized literature relevant to the discussion of the work. These materials are referred to as sources. Digital sources stored within NVivo can be as varied as word documents, pdfs, photographs, audio recordings, or video downloads. They can also include web pages, Twitter feeds, and spreadsheets of quantitative data. Nondigitalized materials are logged in the sources section of the e-project as externals with information on where they are physically located if needed. Source attributes provide researchers with the ability to connect and track important characteristics of documents (such as reference information for digital literature).

Indexical Features: Cases, Classifications, and Nodes

Within each e-project, researchers have access to several forms of indexes or tag systems that allow them to develop well-structured and efficiently organized databases. The indexical systems provide ways to identify and follow cases (a special kind of code that serves as folders for the units of analysis on a project, such as people, places, organization, etc.). Cases, in turn, have the capacity to have a classification type with fixed attribute (or variable) values attached to them, such as gender, age, language, and so on. NVivo also allows for deductive and inductive codes (called nodes in NVivo) that can be linked to a whole source or any part of a source. Researchers can apply codes from within the source or from within a node category where that source material is found.

Interrogation Tools

NVivo provides multiple tools for the analysis of data. The number and kind of tools available increase annually as new possibilities come on the market and new partnerships are forged with software companies and services that can new partnerships are forged with software companies and services that can expand NVivo’s capacity to support qualitative research. Links, query or search tools, visualization tools, and memos are considered the skeleton supports of this area.

Links

As researchers develop hypotheses and uncover relationships among sources and source materials stored at nodes, they can hyperlink their ideas across the parts of the e-project. This allows researchers to virtually jump into different parts of the project in pursuit of emerging understanding.

Query or Search Tools

These tools allow researchers to ask questions of source materials or the components of source materials stored at nodes. Searches range, at the simple end, from word frequency counts (producing word clouds and other visualizations) and word retrieval to, on the difficult end, complex queries employing Boolean logic that allow researchers to look for presence, absence, and context of the coding within their materials.

Visualization Tools

A variety of chart forms and modeling tools provide researchers with opportunities to manipulate raw or summarized materials within an e-project and to create new understanding from that assisted vision. As with other items created in NVivo, the visualizations can be saved and exported for discussions with others.

Memos

Memos provide a place within the project for ongoing writing about the project. Memos can be linked in whole or part to sources or nodes (as opposed to sources or data, which do not link, but are linked to).

Although the three large components of NVivo—sources, indexical features, and interrogation tools—may remain relatively similar in relationship to each other, the contents—their organization and capacities—are not static, and updates are rapidly introduced. For instance, in recent years, NVivo has added options for collection of new forms of data (NCapture add-on for web pages and social media data) and integration with bibliographic management software (Endnote, Mendeley, RefWorks, and Zotero), note-taking software (Evernote and OneNote), and survey tools such as Survey Monkey and Qualtrics. In addition, there is an option within the software to link to TranscribeMe for transcription of audio and video files. These additions have greatly expanded NVivo's capacity and possibilities.

Applications to Educational Research, Measurement, and Evaluation

NVivo is used in any disciplinary field that conducts qualitative and mixed methods studies, from education and medicine to criminal justice, sociology, and anthropology, to name just a few. It is used in governmental agencies and nonprofits and commercial ventures where there is a need to analyze qualitative research data.

Given the range of studies in education that use qualitative research data, NVivo has been employed in everything from studies of individual children or teachers, to classrooms, schools, districts, initiatives, and governmental policy. As a tool to organize nonnumerical data, NVivo has been used in conducting portfolio analysis, calls for public response, and similar text-heavy tasks. It has application in K–12, higher education, and nonformal education. The capacity to integrate with survey tools makes it very valuable to mixed methodologists. Increasingly, those in the education field will also want to make use of its capacity to work with web pages and social media data.

History and Context

NVivo came into being in 1981 when an Australian scientist named Tom Richards decided to assist his sociologist wife (Lyn Richards) by developing a computer program that would help her to better organize the qualitative research data that were spreading over the house. They called the first version of their software program NUD*IST, which was short form for NonNumerical Unstructured Data Indexing Searching and Theorizing. It was one of the first software programs of its kind.

In 1995, the commercial firm QSR International was formed, which since that time has been responsible for the development and dissemination of the time has been responsible for the development and dissemination of the software. In 1999, a completely new and revised form of the software was launched, now named NVivo. The most recent version of the software—NVivo11—was released in 2015, shortly after a new Mac version of the software was made available.

In the most recent iteration of the software, the Windows platform comes in three editions—starter, pro, and plus. Starter is the basic software that supports just textual data analysis with basic search tools. Pro offers the full range of support for a variety of data types as well as more complex queries and visualizations. Plus incorporates some artificial intelligence allowing automatic coding for noun phrases and sentiment as well as social network analysis. There is also a server-based version—NVivo for Teams—that enables the real-time collaboration on an NVivo project.

QSR now advertises itself as the largest privately owned qualitative research software developer in the world. It has offices in North America, the United Kingdom, Japan, and Australia and works with academic, government, and commercial organizations in almost every country in the world. The company has received numerous software development and business awards.

Future Challenges

NVivo faces challenges similar to those faced by all QDAS tools. Many qualitative researchers have been reluctant to leave their traditional practices and embrace digital methods, particularly when they find that the new tools come with a steep learning curve. Another ongoing concern that veteran and new users face is that there is no official standard for reporting QDAS-supported qualitative research, which is why many qualitative research accounts say little about the ways QDAS was involved in the production of the results.

Like other software developers that got into the race before the development of the Internet, QSR must contend with challenges to move its products to the digital cloud. As qualitative research becomes increasingly team-based, QDAS users have also called for more collaborative capacities within and across QDAS packages. At the same time that QDAS developers must consider how to prevent digital attacks and data breeches, they must also grapple with issues related to the archiving of qualitative research data, an arena that is only beginning to receive attention.

Although there are many digital tools that perform some of the tasks qualitative researchers need, such as word processors, wikis, blogs, and a variety of apps, QDAS packages like NVivo are unique in the number and kind of tools integrated within one package. Because QDAS are expressly created for the conduct of qualitative research, they offer researchers many advantages that cannot be obtained from using multiple tools, each with only a portion of the needed components.

In an era of big data, in which the necessity of quantitative tools to make sense of massive amounts of data cannot be denied, there is also a place for small data, that is, the raw human material of narratives, symbols, and representations, which have not yet crossed over the boundaries into massed numerical structures. QDAS tools, like NVivo, have the capacity to help researchers make sense of qualitative data and to develop meaningful explanations that stand on their own or can be integrated with explanations derived from big data.

Judith Davidson

See also Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Bazeley, P., & Jackson, K. (2013). Qualitative data analysis with NVivo (2nd ed.). London, UK: SAGE.

Davidson, J., & di Gregorio, S. (2011). Qualitative research and technology: In the midst of a revolution. In N. Denzin & Y. Lincoln (Eds.), Handbook of qualitative inquiry (4th ed.), pp. 627–643). Thousand Oaks, CA: SAGE.

Davidson, J., & di Gregorio, S. (2011). Qualitative research, technology, and global change. In M. Giardina & N. Denzin (Eds.), Qualitative inquiry and the global crisis (pp. 79–96). Walnut Creek, CA: Left Coast Press.

Paulus, T., Lester, J., & Dempster, P. (2014). Digital tools for qualitative research. London, UK: SAGE.

Silver, C., & Lewins, A. (2014). Using software in qualitative research: A step by step guide (2nd ed.). London, UK: SAGE.

Websites

QSR International Website: http://www.qsrinternational.com/CAQDAS

Web Page: University of Surrey
http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/

O

Andrew Maul Andrew Maul Maul, Andrew

Objectivity

Objectivity

Objectivity

Objectivity is a key concept in scientific and philosophical reasoning. In general, objectivity refers to the idea that the results of scientific inquiry do not or should not depend on the idiosyncratic features of any given individual or group of individuals, such as opinions, values, biases, interpretations, or feelings. Objectivity is often associated with being “bias free,” “value neutral,” or “fair.” Many controversies have surrounded and continue to surround the concept of objectivity, particularly as it is understood in the social sciences. This entry reviews foundational concepts related to objectivity and briefly describes some associated challenges.

Objectivity and Subjectivity

Objectivity is often regarded as an ideal for scientific inquiry, expressing the idea that the knowledge gained as a result of inquiry should depend on the *object* of investigation but not on the *subject* doing the investigation. In general, this requires that the claims made on the basis of inquiry should be testable independently of the individual or individuals making the claims. The truth of the claim that Mount Everest is higher than Mount Kilimanjaro can be verified independently of any individual, whereas the truth of the claim that Gauguin is a better painter than Renoir cannot be verified. The concept of objectivity thus expresses a value that underlies or is closely related to other scientific values including replicability, reproducibility, generalizability, validity, and invariance. Historically, the perception that scientific inquiry leads to objective knowledge has played a significant role in explaining the prestige and authority commonly afforded to science on the part of the general public.

A classic perspective holds that objectivity implies *faithfulness to reality*, that is, claims are objective if they accurately describe facts about the world, implying that such facts are true independently of the perspective of any given individual. This view capitalizes on the intuition that, although each individual experiences the world from a particular perspective, there are features of the world itself that seem to be constant and thus (at least possibly) to exist and have properties independently of the minds of those who perceive them. A common perspective on science maintains that at least one major aim of scientific inquiry is to understand such observer-independent aspects of reality and that the best explanation for the demonstrated value of science (e.g., in terms of historically progressive success in prediction, explanation, and control) is that many fields of inquiry have, at least to some extent, succeeded in this aim. A somewhat softer version of this perspective is captured by the term *intersubjectivity*, which refers to the idea that all members of a community of observers perceive something in common and agree on what is perceived. In practice, intersubjective agreement may have many of the same consequences as objective knowledge but does not preclude the possibility that what is agreed upon does not actually refer to objective reality, nor the possibility that some other subject might disagree with the original community of observers.

Especially in the social sciences, it may be of value to distinguish between claims regarding *epistemic* and *ontological* forms of objectivity and subjectivity. This entry has so far focused on epistemic objectivity, which is a matter of whether knowledge and the methods for acquiring knowledge are independent of individual perspectives. Ontological objectivity and subjectivity, on the other hand, refer to whether the existence of a thing or property of a thing depends on individual perspectives. While most of the subject matter of the physical sciences is commonly regarded as ontologically objective—that is, the physical world is assumed to exist independently of the way it is perceived by any individual—much of the subject matter of the social sciences is ontologically subjective. That is, phenomena such as feelings, goals, beliefs, hopes, fears, and desires exist insofar as they are experienced as existing by an individual; in other words, unlike with ontologically objective phenomena, there is no distinction between appearance and reality. Further, some phenomena, such as money, countries, and marriages, appear to exist and have properties independently of any given individual but not independently of groups of individuals taken together (e.g., money has value insofar as it is perceived by multiple individuals to have value). This distinction is relevant for the psychological and social sciences because ontological subjectivity is not in itself a barrier to epistemic objectivity. It could even be said that, to a large extent, the purpose of the psychological sciences is to provide objective knowledge about subjective phenomena.

Challenges to Objectivity

The notion that knowledge is objective if it directly corresponds to reality seems to imply that objective knowledge is universal and timeless. This idea has been criticized from a number of directions. One of the most famous criticisms comes from Thomas Kuhn, who helped popularize the idea that knowledge is to a great extent contingent upon particular social and historical circumstances. Kuhn proposed that scientists view the world through the lens of a paradigm—a set of shared values, beliefs, vocabulary, norms of practice, and so forth, all at least potentially influenced by broader social, cultural, and historical issues—and that scientists working under different paradigms could wind up coming to conclusions or making claims that are each justified within their original paradigms but are mutually unintelligible (or “incommensurable”). Kuhn (and others) argued that all observation is *theory laden*, and thus that there can never be a completely objective, aperspectival “view from nowhere.” Scholars working in traditions such as feminist philosophy of science argue that perspectivality is not only unavoidable but can actually be good, especially when we acknowledge and make explicit our perspective and positioning within the world. Having a diversity of opinions and perspectives on an issue can be beneficial both scientifically, insofar as intellectual diversity has well-demonstrated epistemic benefits, and morally, insofar as the acknowledgment of the situated nature of perspective can help give voice to historically underrepresented viewpoints and reduce the likelihood that the perspectives of those currently in positions of power will be unquestioningly accepted as representing timeless and universal truths.

Andrew Maul

See also Epistemologies, Teacher and Student; Hypothesis Testing

Further Readings

Harding, S. (1991). Whose science? Whose knowledge? Thinking from women's lives. Ithaca, NY: Cornell University Press.

Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago, IL: University of Chicago Press.

Longino, H. (1990). Science as social knowledge: Values and objectivity in scientific Inquiry. Princeton, NJ: Princeton University Press.

Maul, A. (2013). The ontology of psychological attributes. Theory & Psychology, 23, 752–769.

Reiss, J., & Sprenger, J. (2014). Scientific objectivity. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Summer 2016 ed.). Retrieved from http://plato.stanford.edu/archives/sum2016/entries/scientific-objectivity/

Massimiliano Sassoli de Bianchi Massimiliano Sassoli de Bianchi de Bianchi, Massimiliano Sassoli

Observer Effect

Observer effect

1171

1174

Observer Effect

The term *observer effect* generally refers to the possibility that an act of observation may affect the properties of what is observed. However, depending on the context and the mechanisms involved, it may indicate effects of a very different nature. Observer effects are a threat to validity in much of educational research. After providing examples of observer effects to illustrate its meaning, this entry examines how to limit observer effect. The entry also considers common misconceptions about observer effect.

Imagine measuring the pressure of an automobile tire. When connecting the manometer, it is easy to let out some air, so that the measured pressure may not correspond to the pressure of the tire before the act of the measurement was initiated. Similarly, imagine that we want to know the temperature of a liquid and that to do so we use a thermometer. Because the latter has its own temperature, when it is immersed in the liquid, it can change the liquid’s temperature, so again the observed value may not correspond to the temperature of the liquid before the measurement.

Effects of this kind occur in many domains of physics and are usually named probe effects. They also occur in other domains, like for the so-called *heisenbug* in computer programming (the term is a pun on the name of physicist Werner Heisenberg), denoting a software bug that can alter its behavior, or even disappear, when one attempts to probe it. In social science, the term *observer’s paradox* (coined by the American linguist William Labov) is used to refer to situations in which the presence of the observer can alter the results of the observation. For instance, in sociolinguistics, when a researcher attempts to gather data on natural speech, the researcher may alter the way of speaking of the interviewed (which may become more formal) by the researcher's mere presence. Other designations denote variants of the observer's paradox, like the Hawthorne effect or the experimental demand effect, always describing situations in which the behavior of persons may be altered in ways that are usually not intended by the experimenters, just because they are monitored or placed in a specific experimental context.

Limiting Observer Effects

Observer effects can, in principle, be eliminated, or considerably reduced, by using more sophisticated instruments, improved observational techniques, and other precautions. However, this cannot be done with all observations, as some of them may have a built-in invasiveness, impossible to reduce or eliminate. But the fact that an observation is intrinsically invasive does not necessarily mean that it will alter the result. As an example, consider a wooden cube and the observation of its burnability. For this, we have to put the cube in contact with a flame to see if it transforms into ashes, and this means that the observation of the burnability property destroys it by destroying the entity possessing it (ashes are not burnable). On the other hand, because we know in advance that a wooden cube has the disposition to burn (i.e., we can predict with certainty the result of our observation), although the latter has destructive effects, it nevertheless provides the correct answer.

The previous example shows that an observation, even if irreducibly invasive, can still describe a process of discovery of properties that were actuals before the observation was carried out. However, there are also situations in which the observation can literally create the properties that are observed. In this case, the process not only changes the state of the observed entity, but it does so in a way that cannot be predicted in advance. As a consequence, one lacks a criterion for asserting that the observed property was possessed by the entity, prior to its observation. This is the typical situation in quantum mechanics, engendering the so-called measurement problem. Note that in physics one usually speaks of measurements, rather than observations. However, these two notions are intimately related, as is clear that measuring a physical quantity, like the position of a particle, is about observing (i.e., determining) its value.

Misconceptions

A common misconception is the belief that the quantum observer effect would have something to do with a psychophysical effect, with the consciousness of the observer causing the so-called collapse of the wave function (i.e., the transition from possibilities to actualities). In this view, the selection of the outcome would occur at the level of the conscious observer who performs the measurement and observes the measurement apparatus, as for instance described in Wigner's friend, a famous thought experiment proposed by physicist Eugene Wigner.

Another misconception is that the effect would be limited to the domain of physics. On this last point, it is important to note that the quantum formalism has been applied with success to also model human decision making, in a new research domain called quantum cognition. The reasons for this success are numerous, but one of them is specifically related to the quantum observer effect, which has a natural counterpart in psychology. Indeed, in many interrogative contexts the answers that are obtained are not only discovered, but also created, in an unpredictable way. In short, judgments create rather than record.

Consider the example of a survey where 100 people are asked if they are smokers or nonsmokers. If 20 answer *yes* (and consequently 80 answer *no*), we can say that the probability of finding a smoker in the group of participants is 20%. This probability describes a condition of a lack of knowledge about actual properties. Indeed, participants usually know in advance if they are smokers or nonsmokers, and when asked the question, their responses simply reflect the actual state of affair of their behavior in relation to smoking. This means that the 20% probability reflects the actual presence, prior to the measurement, of 20 smokers in the sample of participants.

Suppose now that the same group of persons is asked whether they are for or against the use of nuclear energy and that 30 answer *for*. Can we still interpret the probability of 30% by saying that, before the question was addressed, 30 persons in the group were in favor of nuclear energy and 70 were against it? This interpretation would be clearly incorrect, as now many participants can be expected not to have a well-formed opinion about the nuclear issue prior to the survey, which means that they will be forced to create one when providing their answers. Thus, no longer is the researcher in a situation in which individuals already have a stored answer, which they can simply retrieve in a deterministic way. Instead, it is a situation in which a question that is new for the participants is posed, so that most of their answers are actualized at that moment, in a highly contextual way (i.e., in a way that depends not only on the state of the participants and the way the question is formulated but also on the unpredictable fluctuations occurring within the participants' mind, when confronted with the new cognitive situation). This means that the 30% probability does not correspond to a situation of a lack of knowledge about actual properties but to a situation of a lack of knowledge about the actual breaks in the symmetry of the potential properties. Also, it can be shown that situations of this kind cannot be properly described by classical (i.e., Kolmogorovian) probabilities but require nonclassical probability models like those used in quantum mechanics.

If a survey like the one just described is interpreted as a measurement, we can say that we are in the presence of an observer effect for two distinct reasons: (1) the process is invasive (as the measurement context forces the respondents to produce an answer) and (2) it literally creates the properties that are observed as outcomes. *Mutatis mutandis,* the same happens in physics, when a measurement is performed on a physical entity in the so-called superposition state. The superposition describes a condition such that the entity, for example, an electron, does not possess in actual terms the properties that are observed, like being localized in a specific region of space. An electron in a superposition state with respect to the position observable is a nonspatial entity (i.e., an entity that is only potentially present in space). However, when interacting with the measuring apparatus, a specific spatial localization can be created, in an unpredictable way, producing the collapse of the wave function describing the electron's state. The observer effect then corresponds to the fact that by observing the position of the nonspatial electron (by means of the measuring apparatus), we force it to acquire one. In other terms, our observation actualizes properties that were only potential prior to its execution.

The previous explanation should clarify one of the common misconceptions about the quantum observer effect, which is about believing that our reality would arise, in the last analysis, as a result of our acts of observation. It is in fact the opposite that appears to be true: One of the consequences of quantum theory is precisely that there must be a reality independent from the observers. It is precisely that reality that lends itself to be observed, which may prove to be quite different than expected. For example, it can be the reality of an electron in a nonspatial state or of a conceptual situation with respect to which we may not have yet taken a final position.

Another important consideration of observer effect is that resulting from the fact

Another important consideration of observer effect is that resulting from the fact that observations can disturb each other and therefore can be experimentally incompatible (as exemplified in Heisenberg's famous uncertainty principle). This means that if we perform a sequence of different measurements, which are not mutually compatible, the order of the sequence can have an influence with regard to the statistics of outcomes. The same happens in psychological measurements, as is clear that when we ask a sequence of questions, their order can influence the answers that are given. For instance, asking first "Is Clinton honest?" and then "Is Gore honest?" does not produce the same statistics of outcomes than asking first "Is Gore honest?" and then "Is Clinton honest?" Question-order effects are clearly a concern for psychologists and sociologists when they perform measurements on beliefs, attitudes, intentions, and behaviors, and one of the ways to attenuate them is to randomize the questions, so that not all respondents will answer the questions in the same order.

It is also important to consider the quantum Zeno effect (the name comes from the famous arrow's paradox devised by the Greek philosopher Zeno of Elea), a situation in which the continuous observation of a physical system can "freeze" its evolution. For example, if an unstable atom is observed with increased frequency, it can be prevented from decaying. A Zeno-like effect has also been described in neuroscience research, by noticing that continuous focus attention can generally stabilize the neural circuits of the brain. In a quite different ambit, the effect of multiple observations is also described in the psychological phenomenon known as the bystander effect, according to which the more spectators are present in an emergency situation, the less likely it is that one of them will help.

To conclude, it is worth mentioning a last occurrence, often also described as an observer effect and able to affect data collection, research design, and data analysis. It occurs when the desire to observe something is so strong that it can lead individuals to believe what they want to believe or to "observe" something that is not really there (see, for instance, the infamous N-Ray affair, where an entire community of scientists deceived themselves). The scientific method was precisely designed in the attempt to neutralize self-deception, but of course human beings cannot be fully immunized from individual and collective prejudices.

The examples and explanations provided cannot exhaust the vast subject of the effects of our observations, particularly in the different fields of inquiry, like physics, psychology, and social science. These effects, however, should not always be considered as a limitation in our ability to access reality. Many times they only tell us that our observations are processes that are more complex than initially expected, which can be deterministic but also indeterministic, noninvasive but also invasive, comprising discovery but also creation aspects, and which of course can also be more or less subjective or contaminated by our personal beliefs.

Massimiliano Sassoli de Bianchi

See also Attention; Double-Blind Design; Experimental Designs; Hawthorne Effect; John Henry Effect; Nonexperimental Designs; Order Effects; Tests

Further Readings

Aerts, D., & Sozzo, S. (2017). Quantum structure in cognition: Origins, developments, successes and expectations. In E. Haven & A. Khrennikov (Eds.), The Palgrave handbook of quantum models in social science: Applications and grand challenges. London, UK: Palgrave & Macmillan.

Busemeyer, J. R., & Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge, UK: Cambridge University Press.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. Journal of Applied Psychology, 73, 421–435.

Krips, H. (2013). Measurement in quantum theory. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Retrieved from http://plato.stanford.edu/archives/fall2013/entries/qt-measurement/

Resnik, D. B. (2005). The ethics of science: An introduction. London, UK: Routledge.

Sassoli de Bianchi, M. (2013). The observer effect. Foundations of Science, 18, 213–243.

Mary L. McHugh Mary L. McHugh McHugh, Mary L.

Odds Ratio

Odds ratio

1174

1175

Odds Ratio

The odds ratio (*OR*) is a nonparametric, inferential test of association. It evaluates whether the odds of a certain event (or outcome) happening is the same for two groups. Specifically, the *OR* measures the ratio of the odds that an event or result will occur to the odds of the event not happening, given one group's exposure to some factor to which the other group has not been exposed. This entry discusses the use of *OR*, its assumptions, and its calculation, significance and strength testing for the *OR*, and interpretation of the *OR*, giving an example of *OR* use.

The *OR* answers the following questions: "What are the odds that a child who has not been vaccinated will contract chicken pox as compared with the odds of a vaccinated child contracting chicken pox?"

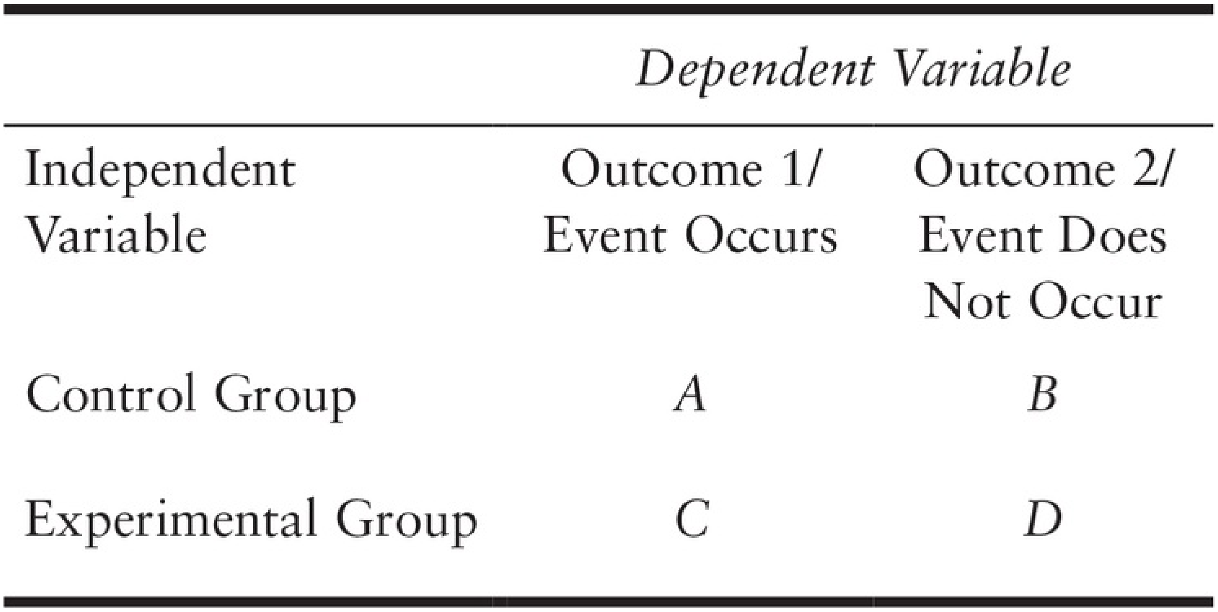
The *OR* is the ratio of the following two:

1. the ratio between the number in the control group with outcome 1 and the number in the control group with outcome 2 and
2. the ratio between the number in the experimental group with outcome 1 and the number in the experimental group with outcome 2.

The *OR* is a robust and fairly versatile statistic used in many clinical studies and educational research designs with two study groups. The *OR* provides an effect size like other correlational statistics, but in a form very different from other effect size statistics. The *OR* provides information on how high the odds are for one condition versus the other. Therefore, the interpretation of the *OR* result is very different from the interpretation of other effect size statistics.

The *OR* assumes subjects were randomly and independently sampled from the population of interest. That is, selection of one subject is unrelated to the selection of any other subject. For example, in drug treatment studies, patients are assigned randomly to either the experimental or control group such that there is no bias in group assignment. The variables are the counts of subjects in each condition, not ratios or proportions.

To calculate the *OR*, the data must be in a 2 × 2 table, as shown in Table 1.



	Dependent Variable	
Independent Variable	Outcome 1/ Event Occurs	Outcome 2/ Event Does Not Occur
Control Group	*A*	*B*
Experimental Group	*C*	*D*

The formula to calculate the *OR* is as follows:

$$OR = \frac{A/B}{C/D} \text{ or } \frac{A \times D}{B \times C}.$$

The formula AD ÷ BC is called the *cross product* and is mathematically equivalent to the original formula (A ÷ B) ÷ (C ÷ D). That is, the results for the two formulas are the same.

For interpretation, it is very important that the table is set correctly and the numbers correctly entered into the formula’s numerator and denominator for the independent and dependent variables. The odds of the experimental group experiencing the outcome must be placed in the numerator, while the odds of the control (untreated) group experiencing the outcome must be placed in the denominator. Reversing the placement will lead to an uninterpretable result.

denominator. Reversing the placement will lead to an uninterpretable result.

Several significance tests may be used for the *OR*, but the Fisher exact test is typically used for a 2 × 2 table. The researcher may also use a chi-square test or a maximum likelihood ratio chi-square test. Strength testing can also be done using the φ coefficient.

Interpretation

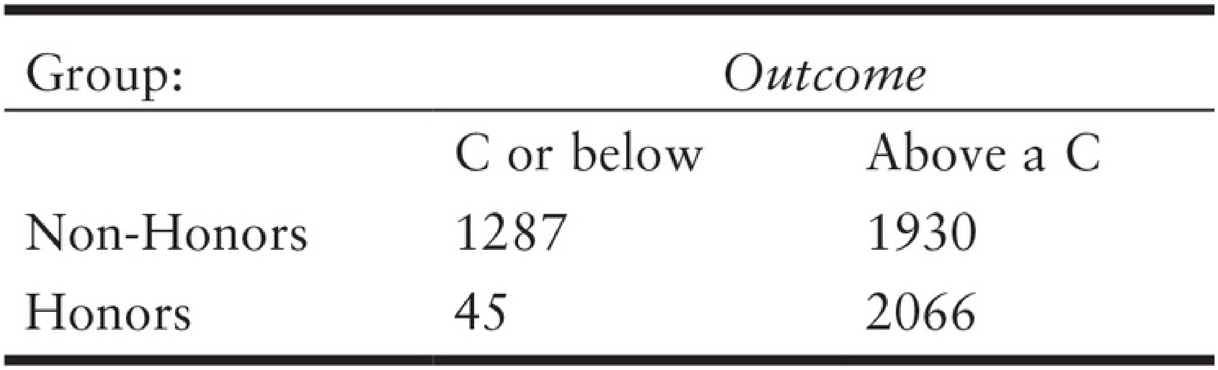
The *OR* is interpreted very differently from other effect size statistical results. Other effect size statistics (such as the Pearson *r* or the Cramér’s *V* coefficient) are interpreted as the amount of covariation of two variables. They can be squared to obtain a coefficient of variation that is a direct measure of the amount of variation in the dependent variable attributed to the independent variable. The *OR* is directly interpreted as the odds of one outcome over the odds of another outcome.

The *OR* value that represents “no effect” is 1.0 as compared with the correlation value zero (0). An *OR* value of 4.37 is read as follows: The *OR* of a member of the control group having the condition is 4.37 higher than a member of the experimental group, meaning the unexposed group will experience the outcome 4.37 times more often than the exposed group. An *OR* greater than 1.0 means there is a higher chance of the outcome of interest for the unexposed (control) group than for the exposed (experimental) group. A negative *OR* indicates that a lower chance of the outcome interest for the control group than for the experimental group.

Example of *OR* Use

Consider the relationship between students’ math scores and being in an honors class. The *OR* can be used to examine the math score by comparing the ratio of those who are in the honors class and those who are not.

Assume that two groups are being studied. In one group, 2,111 students are enrolled in the honors course. The other group of 3,217 students is not enrolled in the honors course. The results after the final math exam were that 45 students in the honors course received a C or below, while 1,287 people in the nonhonors course received a C or below. What is being examined is the ratio of the odds of scoring above a C in the honors group to the odds of scoring above a C in the nonhonors group. The table is set up as shown in Table 2.



Group:	*Outcome*	
	C or below	Above a C
Non-Honors	1287	1930
Honors	45	2066

The *OR* was calculated as follows:

$$(1287 \div 1930) \div (45 \div 2066) = .6668 \div .02178 = 30.6.$$

Alternately the ratio can be calculated using the second equivalent formula:

$$(1287 \times 2066) \div (45 \times 1930) = 1{,}658{,}942 \div 86{,}850 = 30.6.$$

The result of this hypothetical example is interpreted as follows: The odds of a nonhonors student receiving a C or below is 30.6 higher than the odds of an honors student receiving a C or below.

For a significance test, this example used Fisher exact test, for which the p value was <.0001. Alternatively, the chi-square (df = 1) = 975.14, p < .0001.

Mary L. McHugh

See also Logistic Regression; Multiple Linear Regression

Further Readings

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? BMJ, 316(7136), 989–991.

Evidence-Based Medicine Notebook. (1996). Down with odds ratios! Evidence-Based Medicine, 1(6). Retrieved from https://pdfs.semanticscholar.org/96b5/8b85d1678b87902b36662264cf0c04a0c

Scotia, N. (2010). Explaining odds ratios. Journal of Canadian Academy of Child Adolescent Psychiatry, 19, 227.

Laura Pevytoe Laura Pevytoe Pevytoe, Laura

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Office of Elementary and Secondary Education Office of elementary and secondary education

1175

1176

Office of Elementary and Secondary Education

The Office of Elementary and Secondary Education (OESE) is a division of the U.S. Department of Education. The OESE is charged with overseeing and improving the quality of teaching and learning that takes place in elementary, middle, and high school education settings and ensuring that children, youth, and families have equal opportunity to access education services.

History

In 1867, the Department of Education was established under President Andrew Johnson, yet formal organization of the OESE came nearly 100 years later. The 1960s brought a climate of political and social revolution, which led to an increase in federal funding for education and the Elementary and Secondary Education Act (ESEA) of 1965. ESEA provided financial assistance for the education of children of low-income families and funding for school library resources, instructional materials, supplementary educational centers and services, and educational research and training. These efforts continued as Congress passed the Department of Education Organization Act in 1979, and offices from several agencies were combined to allow the department to begin full operations in 1980. The OESE was created to oversee and enforce all aspects of the ESEA.

Organization

Organization

As of 2016, the OESE was composed of eight programs: Office of Academic Improvement, Office of Early Learning, Office of Impact Aid Programs, Office of Indian Education, Office of Migrant Education, Office of Safe and Healthy Students, Office of School Support and Rural Programs, and Office of State Support. The primary adviser for the OESE who oversees these eight programs is the assistant secretary for elementary and secondary education. The assistant secretary serves as the principal consultant to the secretary of education on all issues related to elementary and secondary education. Additionally, the assistant secretary leads the OESE in duties of coordination and recommendation of program policies.

These eight programs are in place to improve achievement of preschool, elementary, and secondary school students at state and local levels, ensure that students from all socioeconomic levels and backgrounds have equal access to the eight programs, and offer financial assistance to local educational organizations whose income is affected by the federal government’s activities. For example, the Office of Academic Improvement has an initiative to increase the participation of low-income students in pre-advanced placement and advanced placement courses and exams by providing grants to qualified entities and supporting the development of enhanced advanced placement courses. The Office of Safe and Healthy Students sponsors a program that assists in the education of homeless children and youth by supporting offices in all 50 states, the District of Columbia, and Puerto Rico that coordinate those education programs and collect data on barriers that may impede students’ regular school attendance.

Laura Pevytoe and Marc H. Bornstein

See also Great Society Programs; Institute of Education Sciences; National Science Foundation; U.S. Department of Education

Further Readings

Stallings, D. T. (2002). A brief history of the United States Department of Education 1979–2002. Durham, NC: Duke University.

U.S. Department of Education. (2014, March 27). Advanced placement incentive program grants. Retrieved from http://www2.ed.gov/programs/apincent/index.html

U.S. Department of Education. (2015, August 5). The Elementary and Secondary Education Act of 1965 as Amended Reports to Congress. Retrieved from http://www2.ed.gov/about/reports/annual/nclbrpts.html

U.S. Department of Education. (2016, December 12). Office of Elementary and Secondary Education, programs/initiatives. Retrieved from https://www2.ed.gov/about/offices/list/oese/programs.html

U.S. Department of Education. (2016, December 22). Education for homeless children and youths: Grants for state and local activities. Retrieved from http://www2.ed.gov/programs/homeless/index.html

U.S. Department of Education. (2016, January 19). Office of Elementary and Secondary Education, overview. Retrieved from http://www2.ed.gov/about/offices/list/oese/index.html

Yanyun Yang Yanyun Yang Yang, Yanyun

Omega

Omega

1176

1180

Omega

Reliability is concerned with the consistency of scores on a measure when the measure is administered to the same group of individuals under comparable circumstances. Scores on a measure are often computed as the sum of scores across items in the measure and are referred to as scale scores. Reliability of scale scores is defined in classical test theory as the ratio of scale true score variance to observed scale score variance. It can also be defined as the correlation between scale scores and the scores from its parallel form, or the square of correlation between scale true scores and observed scores. That is, . The population $\rho_{XX'}$ is unknown in practice and needs to be estimated from data.

Omega ω is ground on factor analysis. It is one of the popular methods for estimating reliability for scale scores. It was initially termed by David R. Heise and George W. Bohrnstedt in 1970 and has been discussed extensively by Roderick P. McDonald since 1978. The classic notion of ω has four assumptions: (1) a set of items measures a single-latent construct (or factor) of interest; (2) the latent factor is the only common cause of the inter-item correlation, and consequently, residual scores are independent across items; (3) the relationship between the latent factor and item scores is linear; and (4) a one-factor model adequately represents the data. After further defining and describing how to calculate ω, this entry details the difference between ω and the coefficient α, which is another common way to estimate reliability of scale scores. Next, the entry describes how to obtain an ω estimate from sample data. Finally, consideration is given to some situations that do not conform to the common assumptions of the classic notion of ω.

Within the factor analysis framework, an individual’s scores on an item x_{ij} is expressed as

$$x_{ij} = \tau_j + \lambda_j F_i + e_{ij},$$

where τ_j is the intercept for item *j*, F_i is the factor score for individual *i*, λ_j is the factor loading of the *j*th item on the factor, and e_{ij} is the individual’s residual score on item *j*. The terms $\tau_j+\lambda_j F_i$ and e_{ij} can be conceptualized as true and error scores, respectively, as defined in classical test theory. But note that e_{ij} includes both specific and random error components. The metric of latent factor is arbitrary. By fixing the variance of factor at one and following the assumption of independency between factor scores and residuals, the variance of observed scores on the *j*th item and the variance of scale scores on the measure are, respectively,

$$\sigma^2_{x_j} = \lambda^2_j + \psi^2_j \text{ and } \sigma^2_X = \left(\sum_{j=1}^{J} \lambda_j\right)^2 + \sum_{j=1}^{J} \psi^2_j,$$

where denotes the residual score variance for item *j*. ω is computed as

$$\omega = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\sigma^2_X},$$

or

$$\omega = 1 - \frac{\sum_{j=1}^{J} \psi^2_j}{\sigma^2_X}.$$

It is interpreted as the proportion of scale score variance that is attributable to the common latent factor or that is not attributable to the uniqueness of items. Under the four assumptions previously described, ω is an accurate estimate of reliability of scale scores, that is, $\rho_{XX'} = \omega$.

Relationship Between ω and α

Over the past half century, the most popular method to estimate reliability of scale scores is coefficient alpha (α). Coefficient α is computed as:

$$\alpha = \frac{J^2 \overline{\sigma}_{jj'}}{\sigma_X^2},$$

where is the mean covariance between item *j* and *j*′. Although numerous authors have cast cautions on or argued against its use, α continues to be widely reported in empirical studies. It has been well documented that α is an undesirable estimate of reliability because it yields a negative bias when the essential tau-equivalency assumption is violated and a positive bias when the error scores are positively correlated across items. The difference between ω and α can be quantified by an index that is derived as follows.

The essential tau-equivalency assumption underlying coefficient α implies that the relationship between the common factor and item scores is the same across items; in other words, all items posit a common factor loading λ. Therefore, in the numerator of Equation 4 is equal to $J^2\lambda^2$; λ^2 is the covariance between any pair of items and is also the mean covariance between items. Equation 4 thus reduces to

$$\omega = \frac{J^2\lambda^2}{\sigma_X^2} = \frac{J^2\sigma_{jj'}}{\sigma_X^2} = \alpha.$$

Subtracting Equation 6 from Equation 4 and solving it algebraically, we could arrive at an index to quantify the difference between ω and α, denoted as *d*:

$$
\begin{aligned}
d &= \omega - \alpha \\
&= \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\sigma_X^2} - \frac{J^2 \overline{\sigma}_{jj'}}{\sigma_X^2} \\
&= \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\sigma_X^2} - \frac{J^2 \lambda^2}{\sigma_X^2} \\
&= \frac{1}{J-1} \frac{\sum_{j=1}^{J} \sum_{j'=1}^{J} \left(\lambda_j - \lambda_{j'}\right)^2}{\sigma_X^2},
\end{aligned}
$$

where $j>j'$. This means α is a lower bound to ω. The equivalence holds when the essential tau-equivalency assumption is met. The magnitude of *d* is trivial when the loadings are similar across items or when the number of items in a measure is large. The magnitude of *d* is nontrivial if factor loadings are greatly heterogeneous across items. For instance, half of the items have very small loadings and the other half of the items have very large loadings.

ω Estimate From Sample Data

Because ω is grounded on factor analysis, an empirical assessment of assumptions underlying ω (as well as α) can be undertaken through a modeling procedure. Specifically, both ω and α require that a single latent factor sufficiently explains the inter-item correlation and there is no residual covariance in the model. Therefore, researchers often begin with a one-factor model with unequal loadings and no residual covariance with an appropriate estimation method. For evaluation of model–data fit, readers are referred to any introductory structural equation modeling textbook. In the classic notion of ω, both factor and item scores are in the interval scale and their relationship is assumed to be linear. Maximum likelihood, generalized least square, or other normal theory-based estimators, is appropriate dependent upon the distributional characteristic of the data. If the one-factor model demonstrates an adequate fit, the parameter estimates from the sample are substituted into Equations 3 and 4 to obtain the ω estimate :

$$\hat{\omega} = \frac{\left(\sum_{j=1}^{J} \hat{\lambda}_j\right)^2}{\left(\sum_{j=1}^{J} \hat{\lambda}_j\right)^2 + \sum_{j=1}^{J} \hat{\psi}_j^2},$$

where the hat on the top of each symbol denotes the sample estimate of the corresponding parameters.

If the one-factor model fails to provide an adequate fit, ω computed by Equation 9 can yield an inaccurate estimate of reliability of scale scores. Researchers often continue with model respecification to determine an optimal model to represent the data. For example, a one-factor model with one or more residual covariances or a two-factor model with a correlation between the two factors may be determined to yield the most meaningful interpretation of data. Through this modeling procedure, researchers not only gain a rich understanding of the interrelationship existing in the item data but also could decide whether an alternative ω to the classic one provides the best estimate of reliability of scale scores. Alternatives to the classic notion of ω are described in the next section.

Beyond the Classic Notion of ω

Although the classic notion of ω assumes that there is only one latent factor underlying all items on a measure, residuals are uncorrelated, and the relationship between items and the underlying factor is linear, ω can be extended for measures that do not conform to these assumptions.

The first assumption underlying the classic notion of ω is that all items in a measure are homogenous (i.e., they measure only one latent construct). However, many instruments designed for educational and psychological constructs measure multiple interrelated subdomains. Examples include measures of attitude, self-esteem, personality, and intelligence. For items that measure more than one latent construct, the linear relationship between factors and item scores is expressed as

$$x_{ij} = \tau_j + \sum_{m=1}^{M} \lambda_{jm} F_{im} + e_{ij}.$$

Different from Equation 1, a subscript *m* is added to both λ_j and F_i indicating that they are associated with the *m*th factor. The factors could be correlated, present a bifactor structure, or posit a hierarchical structure. Following the definition of reliability, ω is the proportion of scale score variance that is attributable to all common factors and is computed as:

$$\omega = \frac{\mathrm{I}\left(\Lambda\Phi\Lambda'\right)\mathrm{I}'}{\sigma_X^2},$$

where Λ is a *j* × *m* factor loading matrix, Φ is a *m* × *m* factor covariance matrix, I is an 1× *j* unit vector, and Λ′ and I′ are the transpose of Λ and I.

Test users are often interested in assessing the degree to which the scale score variance is attributable to one common factor, where the common factor represents the target construct of interest. If a large proportion of scale score variance is explained by this common factor, then individuals’ high-low level on the construct of interest can be indicated by their high–low on scale scores. For this reason, a bifactor model has been advocated by many researchers. In a bifactor model, a general factor underlies all items on a measure, and one or more group factors underlie subsets of items. The general factor represents the latent construct of interest, and group factors explain the additional item covariance that is associated with narrower domains, wording effects, testlet effects, and so on. The general factor is specified as uncorrelated with all group factors. Fixing all factor variances at one, the proportion of variance of the scale factors. Fixing all factor variances at one, the proportion of variance of the scale scores due to the general factor is computed as

$$\omega_H = \frac{\left(\sum_{j=1}^{J} \lambda_{jG}\right)^2}{\sigma_X^2},$$

where λ_{jG} indicates the factor loading of item *j* on the general factor. This index is referred to as ω hierarchical (ω_H). A higher value of ω_H indicates that a large proportion of scale score variance is accounted for by the general factor.

The second assumption underlying the classic notion of ω is that residuals from one item are uncorrelated with residuals from any other items. This assumption is routinely violated with speeded tests and measures containing context-dependent item sets or simply because items are adjacent with each other. Ignoring error covariance from the model leads to model–data misfit. Examination of modification indices and standardized residual covariance matrix could provide statistical suggestion for improving model–data fit. However, researchers should rely on substantive knowledge of measures to decide whether the unexplained covariance among items (after controlling for the common factor) is due to omitting one or more latent constructs or is attributable to nuisance causes. If omitting latent factors from the model is the cause of model–data misfit, the variance of scale scores attributable to the added factors constitutes a reliable component and thus goes into the numerator of Equation 11. If ignoring error covariances is the reason of model–data misfit, the scale score variance attributable to error covariances is a component of the denominator of Equation 11.

The third assumption underlying the classic notion of coefficient ω is that the relationship between item scores and the factor scores is linear. In practice, items are frequently anchored on a limited number of response categories. Applying the linearity relationship to items with at least five categories may not be problematic in applications. However, it can yield considerably inaccurate reliability estimates when the number of categories is fewer and/or the distribution of categorical variables varies across items. Nevertheless, it is not intuitively appropriate to conceptualize that the ordinal score in an item is a linear function of continuous distributed factor scores. For an item with C ≥ 2 categories, it is commonly assumed that the ordinal responses *x* is obtained by applying a set of thresholds to an underlying continuous variable *x** such that

$$x_j = c,\ \mathit{if}\ \tau_{x_j c} < x_j^* < \tau_{x_j (c+1)},$$

where is the threshold for categories $c = 0, 1, \ldots, C - 1$ for item *j*, $\tau_0 = -\infty$, and $\tau_C = +\infty$. Although the relationship between *x** and factor scores remains linear, the relation between the ordered categorical responses and factor scores is nonlinear. In 2009, Samuel Green and Yanyun Yang developed a formula for computing reliability of scale scores that are obtained by summing scores across ordered categorical items. They named the coefficient as nonlinear structural equation modeling reliability coefficient ($\rho_{XX':\,NL}$). Fixing variance of factor(s) at one, $\rho_{XX':\,NL}$ is computed as

$$\rho_{XX':NL} = \frac{\sum_{J=1}^{J}\sum_{J'=1}^{J}\left[\sum_{c=1}^{C-1}\sum_{c'=1}^{C-1}\Phi_2\left(\tau_{x_j c}, \tau_{x_{j'} c'}, \sum_{m=1}^{M}\sum_{m'=1}^{M}\lambda_{x_{j'}^{*} F_m}\lambda_{x_{j'}^{*} F_{m'}}\rho_{F_m F_{m'}}\right) - \left(\sum_{c=1}^{C-1}\Phi_1\left(\tau_{x_j c}\right)\right)\left(\sum_{c=1}^{C-1}\Phi_1\left(\tau_{x_{j'} c}\right)\right)\right]}{\sum_{j=1}^{J}\sum_{j'=1}^{J}\left[\sum_{c=1}^{C-1}\sum_{c'=1}^{C-1}\Phi_2\left(\tau_{x_j c}, \tau_{x_{j'} c'}, \rho_{x_j^* x_{j'}^*}\right) - \left(\sum_{c=1}^{C-1}\Phi_1\left(\tau_{x_j c}\right)\right)\left(\sum_{c=1}^{C-1}\Phi_1\left(\tau_{x_{j'} c}\right)\right)\right]},$$

where is the factor loading linking the continuous item to the *m*th factor F_m, is the correlation between any two factors, is the cumulative probability of given a univariate standard normal cumulative distribution, and is the joint cumulative probability of and given a bivariate standard normal cumulative distribution with a correlation of . The nonlinear structural equation modeling reliability coefficient can be viewed as ω for categorical item data. In a sample, the reliability coefficient can be computed by substituting sample estimates of parameters into Equation 14. The parameter estimates are obtained by fitting polychoric correlation matrix to the model with least square estimators (e.g., diagonally weighted least square estimator) or Bayesian methods.

Yanyun Yang

See also Coefficient Alpha; Reliability

Further Readings

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha.

Psychometrika, 74(1), 155–167.

Heise, D. R., & Bhornstedt, G. W. (1970). Validity, invalidity, and reliability. Sociological Methodology, 2, 104–129.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47, 667–696.

Yang, Y., & Green, S. B. (2015). Evaluation of nonlinear SEM reliability coefficients. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 11(1), 23–34.

Copelan Gammon Copelan Gammon Gammon, Copelan

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Operant Conditioning Operant conditioning

1180

1182

Operant Conditioning

Operant conditioning, also known as instrumental conditioning, is a theory of learning that states that behavior can be modified by its consequences. Unlike classical conditioning, operant conditioning deals with voluntary, rather than reflexive, behavior. Operant conditioning’s effects are maintained or extinguished through reinforcement or punishment. Edward Thorndike’s law of effect first proposed the idea that some consequences of behavior strengthen the behavior. He suggested that satisfying consequences will strengthen a response, but negative consequences will diminish it. B. F. Skinner, commonly referred to as the father of operant conditioning, built on Thorndike’s work but focused exclusively on the empirical study of observable behavior rather than unobservable mental states. This entry includes an evaluation of operant conditioning’s influence on the study of human behavior, the mechanisms by which it functions, and its influence on modern teaching and learning.

Operant conditioning states that behavior is controlled by antecedents, or stimuli, that have previously produced a particular outcome, prompting the actor to repeat behaviors with favorable consequences and avoid those with unfavorable consequences. These outcomes are classified as reinforcement or punishment and further described as positive or negative. Positive consequences refer to the addition of a stimulus, and negative consequences refer to the subtraction of a stimulus. Thus, positive reinforcement occurs when a behavior is followed by a positive stimulus. Negative reinforcement occurs when a behavior is followed by the removal of an aversive stimulus. Positive punishment occurs when a behavior is followed by an aversive stimulus. Negative punishment occurs when a behavior is followed by the removal of a positive stimulus. Reinforcement increases the frequency of a desired behavior, and punishment decreases the increases the frequency of a desired behavior, and punishment decreases the frequency of an unwanted behavior.

Behaviorism

Operant conditioning is part of a greater approach to understand human and animal behavior known as behaviorism. Behaviorism, first coined by John B. Watson, in contrast to cognitive psychology, does not accept unobservable internal states as causes for behavior, but primarily focuses on the effects of environmental factors. In contrast to methodological behaviorism, which dismissed the study of thoughts, feelings, and similar internal states, Skinner's radical behaviorism redefined behavior to include everything that an organism does, including the production of thoughts and feelings. Like his fellow behaviorists, Skinner held that internal states, such as thoughts and feelings, were not valid explanations of behavior. However, he considered them behaviors in their own right, which, like all behaviors, could be explained by environmental factors. In his 1957 book *Verbal Behavior*, Skinner classified language as its own type of operant behavior that functions to interact with and control the surrounding environment, thus making it appropriate for empirical analysis. A primary criticism of behaviorist theories is that they do not sufficiently take into account the mind and personality.

Skinner Box

One of Skinner's valuable contributions to the empirical study of behavior is the operant conditioning chamber, commonly known as the "Skinner box." This chamber allows the placement of an animal, such as a rat or pigeon, in a carefully controlled environment with the ability to perform a simple task and the experimenter means to administer reinforcement or punishment. For example, a rat presses a lever, and food is dispensed as positive reinforcement for the behavior. Over time, the rat presses the lever with greater frequency, signaling that the association between the behavior and the consequence has been learned. Extinction may occur gradually over time if the behavior that had previously been reinforced or punished no longer produces these consequences. The efficacy of operant conditioning on modifying behavior is determined by a variety of factors, including the time interval between operant and reinforcer, and schedule of reinforcement. For example, a shorter interval between action and consequence results in more efficient learning. Frequent, consistent reward or punishment results in faster learning of a behavior, but intermittent of punishment results in faster learning of a behavior, but intermittent enforcement of a consequence produces a slower rate of extinction.

In the Classroom

Operant conditioning has had a significant influence on modern teaching and learning. Operant conditioning has proven to be effective in teaching simple learned behaviors and is frequently utilized with children by parents and teachers. For example, a child learns that a desired behavior results in a reward and increases the frequency of that behavior or an unwanted behavior results in a punishment and reduces the frequency of that behavior. In a school setting this connection might present itself as a certificate given as positive reinforcement for the child achieving all As (a desired behavior). Alternatively, it could present detention as a punishment (the introduction of an aversive stimulus) to certain undesired behaviors like failing to complete assigned work. Reinforcers can include prizes or praise from the teacher, and punishments may manifest in suspension of privileges or being written up. Operant conditioning as an active teaching tool is limited in its usefulness by its restriction to relatively simple behaviors, regulation of acceptable rewards and punishments, and individual differences in the perception of the strength of the desirability or undesirability of the consequences.

Copelan Gammon and Marc H. Bornstein

See also Behaviorism; Classical Conditioning

Further Readings

Schneider, S. M., & Morris, E. K. (1987). A history of the term radical behaviorism: From Watson to Skinner. The Behavior Analyst, 10, 27–39.

Skinner, B. F. (1957). Verbal behavior. New York, NY: Appleton-Century-Crofts.

Skinner, B. F. (1963). Operant behavior. American Psychologist, 18(8), 503–515.

Skinner, B. F. (1984). The evolution of behavior. Journal of the Experimental Analysis of Behavior, 41, 217–221.

Youtz, R. E. P. (1938). The change with time of a Thorndikian response in the rat. Journal of Experimental Psychology, 23(2), 128–140.

Lisa Lee Lisa Lee Lee, Lisa

Order Effects

Order effects

1182

1183

Order Effects

In educational research, an order effect occurs when the order in which research subjects participate in experimental conditions affects the outcome variable being measured. For example, a researcher may be interested in examining the relative effectiveness of two versions of an online reading program on student performance. Students in the study could first complete Program A and then complete Program B, each followed by a standardized assessment. If Program B yields greater improvement than Program A, the researcher might conclude that Program B is more effective. However, without additional evidence, this conclusion may not be valid. Improvement under Program B could have been greater for reasons other than the effects of this version of the reading program. Extraneous factors, such as increased familiarity with the online format of the program or the standardized assessment, could have resulted from completing Program A; these effects could influence the outcomes for Program B. That is, the order in which the participants received the experimental conditions may have affected the measurement outcome. The impact of order effects is important to consider in educational research because of their potential biasing effect on outcome measures and, subsequently, the conclusions drawn from a study. This entry discusses design factors that may lead to order effects in experimental research, methods for addressing and controlling for order effects on outcome measures, and the impact of order effects in questionnaire design, another method commonly used in educational research.

To make causal inferences about what factors may influence the phenomena under observation, researchers design experiments to test the relationship among the factors being studied. In a design that calls for individuals to participate in more than one experimental condition, a phenomenon called sequencing effects more than one experimental condition, a phenomenon called sequencing effects may arise. An order effect is one type of sequencing effect that is a consequence of the order in which participants are administered the experimental conditions. Order effects are distinct from another type of sequencing effect called a carryover effect. A carryover effect is a biasing effect that occurs when the effects of a prior experimental condition continue to influence a participant’s performance in the subsequent condition. For example, in the example previously provided in this entry, it is possible that Program A led to changes in students’ reading strategies. Performance in the second condition, Program B, could be influenced by both order effects (increased familiarity with test conditions or the assessment) and carryover effects (changes in reading strategies).

Broadly speaking, research experiments may incorporate a between-subjects or a within-subjects design. In a between-subjects design, research participants are each assigned to a different condition, and changes in outcome measures between groups are compared. Because research subjects participate in only one experimental condition, order effects do not occur. In a within-subjects design, instead of assigning participants to different experimental groups, participants would be administered more than one and perhaps all of the conditions. Order effects are an issue for any experiment in which research subjects participate in more than one condition.

A common method used for addressing order effects in a research design is counterbalancing. A simple way to counterbalance is to give all research subjects all the experimental conditions, presenting each with a different order while representing all possible orders across subjects or groups. When the number of experimental conditions is small, representing all order combinations across all groups may be feasible. For example, with two experimental conditions, A and B, only two orders are possible, A–B and B–A. However, such an approach becomes more impractical as the number of experimental conditions increases and the total number of possible combinations increases as well. Counterbalanced designs can include a subset of the possible combinations. Such a design would ensure that each experimental condition is adequately represented both in the order in which it is presented and in its appearance both before and after the other experimental conditions included in the research.

Although it is important to counterbalance experimental groups to control for order effects in an experiment, it is also important for the researcher to ensure that the groups within the experiment can be considered equivalent. In the case of the reading study example, careful counterbalancing of the order of presentation of the online reading programs across subject groups will be undermined if the groups themselves are imbalanced on factors, such as baseline reading ability or familiarity with use of computers, that can affect the measurement outcomes. Methods such as randomization can help ensure that experimental groups can be considered comparable.

In addition to the order effects that can occur in experimental settings, order effects are a consideration when using other educational research methods as well. When data from participants are gathered via the use of a survey questionnaire, for example, the order in which the questions appear can influence responses. Order effects in survey research occur when the order in which the survey questions are presented influences the responses to those questions. For example, a question on teacher satisfaction may function differently when it follows questions on workload and pay than when it follows questions on the rewards of being a teacher. A psychological mechanism that is thought to contribute to context effects in a survey setting is priming. Priming refers to the activation of information in memory due to the presentation of a stimulus. Much like a carryover effect, due to priming, the survey questions that precede a particular item will activate information in memory that may in turn influence responses to subsequent items. Questions that activate memories of long work hours and low pay will likely influence ratings of teacher satisfaction differently than questions that elicit memories of the joys of teaching. Survey researchers with an interest in examining the effects of question order on response often conduct experiments to manipulate the order of item presentation across survey respondents and observe the effects of item order on respondent answers.

When conducting research, the design of the study must be carefully planned to ensure that the study yields valid conclusions. To ensure that order effects do not limit the value of a research study, the researcher must consider potential order effects when planning the study and incorporate techniques for reducing the potential impact of order effects on the outcome measures of a study.

Lisa Lee

See also Experimental Designs; Repeated Measures Designs; Survey Methods; Surveys

Further Readings

Further Readings

Alferes, V. R. (2012). Methods of randomization in experimental design. Thousand Oaks, CA: SAGE.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2014). Research methods, design, and analysis. Boston, MA: Pearson.

Johnson, R. B., & Christensen, L. B. (2016). Educational research: Quantitative, qualitative, and mixed approaches. Thousand Oaks, CA: SAGE.

Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys. New York, NY: Wiley.

Paul B. Ingram Paul B. Ingram Ingram, Paul B.

Michael S. Ternes Michael S. Ternes Ternes, Michael S.

Ordinal-Level Measurement Ordinal-level measurement

1184

1185

Ordinal-Level Measurement

Ordinal-level measurement is a method of assigning numbers to values that indicates some hierarchy or order among scores. This is done through the utilization of an arbitrary value system wherein the quantitative categories are differentiated based on the quantity of some variable. Using the arbitrary numbering system in a given set of scores, ordinal-level measurement indicates an order with an uneven spacing assumed between the various response scores. It provides an opportunity to determine whether a value is greater than or less than another point to which it is being compared. As such, ordinal-level measurement constructs a hierarchical structure for the given values of interest. Categorizing responses to include a meaningful and interpretable order contrasts the merely categorical approach to measurement observed in nominal-level scales. However, beyond the establishment of frequency and rank ordering for a given occurrence, ordinal scales do not innately provide significant information for comparing response scaling. For this reason, an ordinal-level scale is sometimes called a ranked scale. The goal of ordinal-level measurement is to provide a method for ranking a set of linear data points such as letter grades in a class. The remainder of this entry further reviews the origin and use of ordinal-level measurement, provides examples of its application, and considers issues related to its use when conducting statistical analyses.

Proposed as one of the four scales of measurement from within Stanley Smith Stevens’s 1946 classification system (often referred to as Stevens’s taxonomy), ordinal-level measurement assumes exclusivity between response groups and a logical, set order. Ordinal scaling does not contain an absolute zero value for responses that indicate a lack of an experience. Although ordinal scaling allows for rank-order comparisons, it does not enable one to discern the degree of for rank-order comparisons, it does not enable one to discern the degree of difference between the various response levels. For instance, for contestants in a race who are awarded first, second, and third place prizes, ordinal-level measurement offers a clear rank-order progression with first place showing a faster race time; however, it cannot be assumed that the time observed between first and second place is the same as the time between second and third. Likewise, finishing last in the race does not presume an absence of time for the event. In addition to rank ordering values, ordinal-level measurement can enable both strongly ordered and weakly ordered information. A strongly ordered measurement would be one that provides information that is not dependent solely upon categorical information and is sequential. Strongly ordered values are represented in the aforementioned example of runners finishing a race. In contrast, a weakly ordered measurement would track information first on a nominal level and then provide a meaningful ranking. An example of this weakly ordered measurement would be when behavior is coded to determine protocol adherence during therapy. A frequency count of behaviors is kept and then, if the protocol prescribes a certain behavior to be employed more than another, the subsequent ranking of the calculated frequencies yields an indication of adherence.

Issues in the Use of Ordinal-Level Measurement

Ordinal-level measurement is a common approach to code many survey questionnaire items, such as questions utilizing both scaled and dichotomous response options. Theorists have suggested that the frequently used Likert and Likert-type items should be considered ordinal in nature as they fail to meet the criteria for an interval scale. For example, assume that an item reads “I am depressed” and offers the opportunity for a respondent to select an answer on a scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*.) An ordinal-level scale assumes that the respondent’s response preference is meaningfully ordered, but that the magnitude between response options may not be the same. An interval-level scale (e.g., temperature) would not have this problem as the difference in item-level information needed to get a respondent to change from indicating 1 (*strongly disagree*) to a 2 (*disagree*) would be equally proportional to that needed to change from a response of 4 (*agree*) to a 5 (*strongly agree*).

Statistical analyses utilizing modern testing theory approaches, such as Rasch modeling using item-response framework, have demonstrated that an interval assumption in response patterns is a frequent and inappropriate one for ordinal scales. Item response levels contain different amounts of information and thresholds for endorsement. Thus, although ordinal-level measurement does not innately provide for sufficient item-level information to meaningfully compare between response categories using classical test theory, modern approaches to measurement have enabled the extrapolation of that information. This modern approach promotes a view that Likert and Likert-type scales can provide sufficient information to place themselves somewhere between the rank-ordered approach originally described in Stevens's taxonomy and the proportionally incremental additions necessary for interval scaling.

Despite this limitation of unequal information between response options, ordinal scale items are frequently treated as though they are interval in nature. They are treated this way in order to take advantage of more advanced statistical analyses (e.g., standard deviation, rank-order and product-moment correlation, and analysis of variance) not available to ordinal-level measurement, according to the definition proposed by Stevens's taxonomy. The assumption that statistics need interval data rests on an assertion that advanced statistics require a knowledge greater than rank-ordered information, which is all that can be presumed with an ordinal scale. Despite this tradition of considering ordinal scaling as inappropriate for many statistical measurements, responses utilizing ordinal scaling are sometimes considered as appropriate for answering many of the questions inherent to behavioral science as other standards of measurement. Some psychometric theorists have even challenged the assumption that ordinal data are inappropriate for more advanced statistics on the basis that the numbers provided in a response to a given item do not enter into consideration for the statistical assumption of the analysis. In other words, these theorists argue that there is no relationship between measurement scales and statistical procedures. The perpetuation of statistical restrictions on the basis of an item being ordinal in nature has come to represent a myth that is difficult to dispel.

Paul B. Ingram and Michael S. Ternes

See also Interval-Level Measurement; Likert Scaling; Nominal-Level Measurement

Further Readings

Coombs, C. H. (1960). A theory of data. Psychological Review, 67, 143–159. doi:10.1037/h0047773

Gaito, J. (1980). Measurement scales and statistics. Psychological Bulletin, 87, 564–567. doi:10.1037/0033-2909.87.3.564

Norman, C. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114, 494–509. doi:10.1037/0033-2909.114.3.494

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103, 677–680. doi:10.1126/science.103.2684.677

Hamish Coates Hamish Coates Coates, Hamish

Organisation for Economic Co-operation and Development Organisation for economic co-operation and development

Organisation for Economic Co-operation and Development

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organization that has grown to play a major role in global education policy research and development. This entry provides an overview of the organization, charts OECD's work in education, reviews two signature programs, and surveys a range of broader initiatives.

The OECD was established in 1961, with foundations formed during the rebuilding of Europe in the decade following World War II. Headquartered in Paris, France, the OECD is a club of mostly rich countries that are democracies with market economies. The 35 members are mostly from Europe, North and South America, and East Asia. A large number of nonmember countries are in accession talks and participate in OECD activities. The OECD Council governs a secretariat, which is structured into a dozen departments, one of which is the Directorate for Education and Skills.

The OECD's work in education has increased substantially since the late 1990s. The Directorate for Education was formed in 2002 under the inaugural directorship of Barry McGaw. Work in education now spans birth to old age and involves consultations, policy studies, large-scale assessments, and innovative research. OECD's work in this field is funded by countries as well as a range of other partnership and agencies. A few signature programs have achieved particular prominence.

The document *Education at a Glance* has been published annually since 1998, growing to provide what is likely to be the most extensive information about education around the world. The analytical presentation of indicator data is structured using the OECD’s Indicators of Education Systems, which frames education as involving a series of actors (e.g., individuals, instructors, providers, and systems) and stages (e.g., outcomes, contexts, and antecedents). The annual publications are now many hundreds of pages long and provide information on dozens of countries sourced from core governmental data as well as a range of discrete initiatives.

The *Programme for International Student Assessment* is the most prominent among the growing suite of OECD’s large-scale assessments. it was first conducted in 2000, with subsequent assessments implemented every 3 years. The program tests the capacity of sampled 15-year-old school students to apply reading, mathematical, and scientific literacy to solve real-life problems. Over 80 countries have participated in the assessment since its inception, with a range of extension initiatives looking to further expand the program’s reach.

OECD conducts a host of other education-related studies. Examples include the Survey of Adult Skills, the Teaching and Learning International Survey, thematic reviews of tertiary education systems, the Assessment of Higher Education Learning Outcomes, a system for benchmarking higher education system performance, and research topics as diverse as innovative learning environments, policy trends shaping education, and open education resources.

Hamish Coates

See also Programme for International Student Assessment; Progress in International Reading Literacy Study; Rankings; Trends in International Mathematics and Science Study

Further Readings

Ball, L. M. (2014). Long-term damage from the Great Recession in OECD countries (No. w20185). Cambridge, MA: National Bureau of Economic Research.

Putnam, R. D. (2016). Education, diversity, social cohesion and “social capital.” Dublin, Ireland: OECD Education Ministers.

Sellar, S., & Lingard, B. (2014). The OECD and the expansion of PISA: New global modes of governance in education. British Educational Research Journal, 40(6), 917–936.

Sharon Brisolara Sharon Brisolara Brisolara, Sharon

Outcomes

Outcomes

1186

1190

Outcomes

One of the contributing factors to the development of program evaluation as a profession has been an interest in learning whether the efforts of social interventions, programs, and other types of innovations had the impact purported by designers. In other words, funding agencies, politicians, and citizens want to know whether program efforts make a difference. What were the results of the investments of time and money? Did the programs accomplish what planners proposed? Were the results achieved intended? What kind of a difference did the program or intervention make and why does it matter? All of these are common questions about outcomes.

Outcomes are often expressed as short-, medium-, and long-term outcomes. Long-term outcomes are sometimes referred to as impact. The time frame associated with what constitutes short-, medium-, and long-term outcomes varies and is often clarified in evaluation documents. In some cases, for example, a short-term outcome may be expected within 1–3 months, with medium-term outcomes expected in 3–6 months. In many cases, the long-term outcomes represent the ultimate change desired and can be expected years later, sometimes outside of the proposed time frame for the program.

This entry examines the various elements and criteria of outcomes in relation to program evaluation and describes how outcomes differ from outputs, indicators, and goals. Next, how outcomes are used in the steps of program evaluation are reviewed, including developing and framing the evaluation and conducting data analysis and interpretation. Finally, the entry concludes by looking at complexity models and other current issues related to outcomes.

Outcomes and Program Evaluation

The focus on outcomes in program planning and evaluation has been shaped by a number of social, cultural, economic, and political forces. Not only are those who design and implement a program concerned about outcomes, but funding agencies, community partners, government officials, and the participants themselves are also interested in the merit, worth, and ultimate results of a program intervention. As previously suggested, one impetus has been a focus on accountability and an interest in the degree to which intended results have been realized through program efforts and investments. From an accountability perspective, a focus on outcomes helps to answer the question of whether funds have been put to good use and if the program warranted the investment of time and money it was given. By examining outcomes, a program evaluation is able to assess whether a program accomplished what it promised to achieve.

Another impetus behind the focus on outcomes has been an interest in effective program design and a deeper understanding of how particular strategies can lead to particular desired results. Part of the work of articulating outcomes includes ensuring that there is a shared understanding of the relationship between activities or strategies and results achieved. With a theory of how the program is intended to work made clear, it is possible to better understand which program elements are critical, the level of investment needed, or perhaps where a program has or could falter or fail.

Most notably, a focus on outcomes is often synonymous with a focus on measurement. Indeed, the use of the term *outcomes* is commonly qualified by authors of funding proposals and program guides as *measurable outcomes*. The focus on being measurable emphasizes the point that the value of outcomes is beyond that of an organizing principle or guide. The usefulness of an outcome lies in one’s ability to measure a program’s progress toward this desired result. Using credible data to measure the degree to which outcomes have been achieved has been seen as critical to effective program improvement, important for informing training and resource allocation, significant in determining future funding and expansion possibilities, and often necessary to the sustainability of program efforts. Sustained evaluation of outcomes across time or program sites leads to greater confidence in the particular activities and strategies chosen as contributing to the types of results desired.

Although there exist differences in how outcomes are written that may depend on disciplinary or organizational norms, outcomes tend to have certain elements in common. One element is the inclusion or clarification of the intended beneficiary. The beneficiary or focus of the desired change can be a person, such as participants, an organization, or system. The intended beneficiary should be specifically described.

Another element of written outcomes is a statement about the desired behavior or change. For participants, for example, an outcome may express the desired changes in behavior, attitudes, awareness, skills, or knowledge. Another aspect may be the direction of the change expected. The desired outcome or result might be an increase, decrease, maintenance, or prevention of a condition or circumstances. Some advocate for including additional statements within the outcome itself regarding conditions or circumstances and standards that qualify the result; for example, these might describe the degree of change expected.

In addition to elements, various criteria have been recommended as characteristics of effective, useful outcomes. One widely used mnemonic that encapsulates such criteria, particularly for effective short and intermediate outcomes, is SMART. SMART is used to remind users creating outcomes to ensure that they are specific, measurable, achievable or attainable, realistic (or sometimes, relevant), and time bound (or timely). *Specificity* refers to the clarity of the elements of the outcome such as the beneficiary and what is to be accomplished. Often *measurable* refers to an outcome being quantifiable, although there is increasing recognition that rigorously collected qualitative data also constitute credible evidence. In either case, the outcome must be crafted in a way that allows it to be measured using reasonable techniques.

SMART and similar guidelines acknowledge that good outcomes are ones that can be *achieved* or accomplished in the given context with the resources available and time frame suggested. The outcome shouldn’t be so lofty that progress toward the outcome would be unreasonable. Similarly, it is important to consider whether the outcomes proposed are *realistic* and can be expected to follow from the design. If outcomes are realistic, it is reasonable to expect that short-term results will lead to long-term results. Finally, outcomes should clarify the length of time in which results are to be accomplished and should be attainable within the suggested time. The criteria of *timeliness* suggest that progress toward an outcome should be determined within a period of time that supports utilization and is not so far removed from implementation that the results or finding don’t matter or cannot be used. Other guidelines for outcome characteristics have been suggested. Often they contain variations of the characteristics already noted. One recent addition has been the criteria of credibility, meaning the measurement or types of data produced are credible to the intended audiences of the results.

Outputs, Outcomes, Indicators, and Goals

In common usage, the term *outcomes* can be used loosely and confused with other terms. Outcomes are commonly confused with outputs, indicators, and goals. From a measurement and program evaluation perspective, these remaining terms can be defined as follows:

- *Outputs*: Outputs refer to the products produced as a result of the activities, support, or service interactions participants received. Examples might include the number of workshops offered or the number of participants served. Note that an output does not mention the result desired from the product of the activities. However, outputs can provide a sense of the program dosage or the breadth of program intervention planned.
- *Indicators*: Indicators can be defined as specific and measurable characteristics used to demonstrate progress toward an outcome. As such, one might include more than one indicator per outcome. In some cases, outcomes may be framed as the desired change for the intended beneficiary without including specific measurement criteria. The specific standard might then be expressed in the indicator. That said, indicators are not exclusively used for outcomes; indicators can be used for outputs and processes as well.
- *Goals*: Most often, goals are broad, general statements of future desired states. In such cases, they do not need to be measurable. Sometimes, long-term outcomes or impacts are used to express the goal of a program.

It should be noted that the term *objective* is sometimes used in place of *outcomes*, particularly in the field of education, but the term has its own specific meaning. Often, objectives are used to refer to course objectives, which are the goals and intentions of a particular academic course. The clear, measurable criteria that demonstrate what students should learn or achieve are increasingly referred to as student learning outcomes in keeping with the meaning of outcome noted in this entry.

Use of Outcomes in Program Design and Evaluation

Use of Outcomes in Program Design and Evaluation

The use of outcomes can be found at every stage of the evaluation enterprise. A few examples of evaluation stages and use of outcomes are provided in this section.

Developing the Evaluation Plan

The development of an evaluation plan often specifies what will be evaluated; the criteria for the evaluation and key questions; as well as how data will be collected, analyzed, interpreted, and reported. Key evaluation questions may be linked to specific program outcomes that the questions seek to elucidate. Alternatively, and perhaps more commonly, the evaluation plan may be based on a logic model or program theory/theory of change that describes how particular inputs, activities, and/or strategies lead to chosen outcomes. These program models are sometimes presented in table form and sometimes as visual representations. The description of causal links, contributions, or influences expected to lead to further results informs evaluation and monitoring processes, as well as program design.

Framing and Describing the Evaluation

Articulation of outcomes, as a process and product, helps evaluation planners and users to determine standards for success (e.g., what success would look like). Creating outcomes is critical to determining appropriate measurement strategies and data collection tools. The creation of outcomes and selection of data collection tools or measurement strategies can be an iterative process with an awareness of what can be measured in a particular context (e.g., given cultural and societal norms, evaluation budget, and program capacity) and available data and tools informing eventual selection of outcomes and measurement strategies.

Data Analysis and Interpretation

Selected outcomes and, specifically, the theory underlining how and why a short-term outcome will lead to intermediate-and long-term outcomes are examined again during the process of data analysis and interpretation. Data findings may be reported in relation to intended outcomes. Both intended and unintended outcomes and results are examined during this stage of an evaluation, although efforts to collect data on unintended outcomes should be integrated throughout the stages previously mentioned.

Current Issues

Although an examination of the range of different approaches or models of program evaluation is beyond the scope of this entry, it should be noted that evaluation models differ in their stance on the role of and use of outcomes in evaluation. For example, in earlier years, goal-free evaluation required that the evaluator conduct the evaluation without having prior knowledge of the intended goals or outcomes in order to better ascertain the actual outcomes produced by the intervention. Appreciative inquiry principles have been integrated into program evaluation models through a deep investigation of what works best in an organization or system and to hone in on what enlivens or animates individuals and the institution. This can constitute more of a focus on what *is* rather than what *was* intended to be. Recently, complexity models have emerged in response to programs that are expected to significantly change and develop over time. Such programs, and sometimes the contexts in which they are situated, are themselves responsive to new challenges, opportunities, and partners. Strategies and the outcomes themselves emerge and may not be the ones potentially identified at the beginning of the process. Complexity models have garnered attention, given the messy nature of community work, particularly in times of political and economic change, and the interest in innovation.

However, even in models where measurable outcomes are established in advance and monitored, evaluation practitioners note the importance of not only attending to intended and articulated outcomes but also exploring unintended outcomes. These unintended outcomes may be positive or negative in nature. Multiple strategies have been suggested for systematically exploring the existence of unintended consequences. In general, including qualitative methods and open-ended questions can be useful in elucidating unexpected results. It is important to take into account the experiences of various subpopulations in exploring both intended and unintended outcomes as consequences may differ significantly across populations. Indeed, unintended consequences might occur as the result of changes in interactions or dynamics with disparate impacts when considered by gender, class, racial or ethnic identification, area of residence, ability, sexual orientation, age, religious affiliation, and the like.

Sharon Brisolara

See also Ethical Issues in Evaluation; Evaluation Versus Research; Logic Models; Process Evaluation; Program Evaluation; Program Theory of Change

Further Readings

Bamberger, M., Rugh, J., & Mabry, L. (2012). Real world evaluation: Working under budget, time, data, and political constraints. Thousand Oaks, CA: SAGE.

Brisolara, S., for JBS International. (2014). Gender-sensitive evaluation and monitoring: Best and promising practices in engendering evaluation. USAID. Retrieved from http://pdf.usaid.gov/pdf_docs/PA00K43P.pdf.

Innovation Network. Evaluation plan workbook. Washington, DC: Author. Retrieved from http://www.pointk.org/client_docs/File/evaluation_plan_workbook.pdf.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). Program evaluation: Alternative approaches and practical guidelines (4th ed.). Old Tappan, NJ: Allyn and Bacon.

Mertens, D. M., & Wilson A. T. (2012). Program evaluation theory and practice: A comprehensive guide. New York, NY: Guilford Press.

Patton, M. Q. (2011). Developmental evaluation: Applying complexity concepts to enhance innovation and use. New York, NY: Guilford Press.

Preskill, H., & Catsambas, T. (2006). Reframing evaluation through appreciative inquiry. Thousand Oaks, CA: SAGE.

Rossi, P. H., Lipsey, M., & Freeman. H. E. (2004). Evaluation: A systematic approach (7th ed.). Thousand Oaks, CA: SAGE.

W. K. Kellogg Foundation Evaluation Handbook. Retrieved from https://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook.

David M. Hansen David M. Hansen Hansen, David M.

Out-of-School Activities

Out-of-school activities

1190

1194

Out-of-School Activities

A large portion of children's and adolescents' time in the United States is devoted to discretionary or leisure activities outside of normal school hours; some estimates suggest they spend half of their waking hours in these activities. Historically, society has held a somewhat negative view of the importance of these out-of-school activities. However, research over the past 25 years has led to a shift in society's and scholars' view of out-of-school activities. There is now widespread recognition that out-of-school activities play an important role in promoting child and adolescent learning.

This entry first provides a clear definition of out-of-school activities, then looks at approaches researchers have taken to study these activities. Next, the participation patterns of youth in out-of-school activities are investigated. Finally, what children and adolescents learn from out-of-school activities is considered, and methodological challenges of studying the impact of these activities are explored.

Defining Out-of-School Activities

As used both here and in research, out-of-school activities refer to activities that are organized and structured occurring outside of the formal school day with some degree of adult supervision or leadership. Thus, unstructured activities, such as time spent "hanging out with friends," are excluded. The following are other terms that are frequently used for out-of-school activities: organized youth activities/programs, afterschool activities, extracurricular activities, and community-or school-based activities.

In general, children’s and adolescents’ participation in out-of-school activities is considered voluntary, differentiating out-of-school activities from formal education, which typically consists of required activities. It should be noted that for children, participation in out-of-school activities may be somewhat less voluntary and more directed by a parent than in adolescence, particularly if a working parent needs child care during afterschool hours. However, even in such cases, children tend to perceive greater control over what they do in out-of-school activities versus the regular school day.

Out-of-school activities are tied to particular youth-serving programs or organizations. Youth programs run the gamut from local, grassroots programs to long-established national and international programs (e.g., YMCA). A program’s philosophy and perception of what youth need guides the aims and purposes of the program activities. These aims and purposes can vary widely across programs, as well as within a program over time. Not surprising then, the range of activities reflects the diversity of programs’ aims and purposes.

Approaches to Studying Out-of-School Activities

Researchers have created several activity taxonomies in order to study out-of-school activities—a taxonomy represents the particular way a researcher classifies activities into meaningful units that can then be used to understand how learning differs across these units. Understanding these taxonomies, then, is important for interpreting research of out-of-school activities.

One common taxonomy is to group by different types of activities. For example, some research has used a six-group model: sports, performance and fine arts, academic clubs and organizations, community-oriented programs, service programs, and faith-based groups. Other research has used more or fewer categories of activity types, for example, team versus individual sports or service, faith-based, and community-oriented activities combined into one group. The rationale for grouping by type of activity is that participants should have similar learning experiences because the types of activities are also similar.

Another taxonomy for studying the impact of out-of-school activities is to group activities by their purpose. Given that organized youth activities have a set of guiding aims and goals, it makes conceptual sense that activities with common aims and goals would offer similar learning experiences. Common purpose-based groupings include those focused on academic, enrichment, physical/sport, visual art, culinary, or specific occupation-related skills.

Organization-based taxonomies have also been used. The rationale for this classification is that particular organizations (e.g., YMCA) have distinct learning aims and goals that guide a site or location's activities. This method is conceptually similar to the purpose-based classification, but instead of grouping by a common purpose across activities, the organization becomes the de facto purpose. It is important to note that most organizations have multiple aims and purposes; thus grouping by organization focuses on evaluating the overall impact of a program on participants rather than on learning specific to a common aim or purpose.

The choice of a particular taxonomy has implications for understanding what participants learn in out-of-school activities—it provides a lens through which to view learning. The particular taxonomy, then, necessarily excludes potential differences that are beyond its focus. Thus, as attention next turns to summarizing research on out-of-school activities, bear in mind that the findings are based on a given taxonomy with its particular focus and limitations.

Out-of-School Activity Participation Patterns

Obviously, in order for children and adolescents to benefit from out-of-school activities, they need to participate in them. But, how much participation is needed to experience these benefits? The idea of "dosage" is a useful way to conceptualize how much children and adolescents are exposed to the learning opportunities of the activities. Dosage has been conceptualized and quantified in different ways: dichotomous *yes/no* participation, *intensity* (hours), *frequency* (number of days in a given period), *duration* (participation in same activity/program over time), *breadth of participation* (number of different types of activities), and *total number* of activities.

At the broadest definition of dosage—dichotomous *yes/no* participation—research indicates that a vast majority (>75%) of adolescents and children participate in at least one activity during a given period (e.g., over 3 months), suggesting that out-of-school activity participation has become the norm and not the exception. Given that participation is the norm, this yes/no level of dosage measurement has little ability to differentiate learning between youth.

Intensity and *frequency* of participation vary widely depending on the type of activity, with sports associated with greater intensity and frequency. That said, across all types of activities, some estimates indicate that, on average, adolescents participate between 2 and 5 hours per week and attend an activity at least once per week. Greater intensity and frequency of participation are both generally associated with more positive outcomes and experiences. Greater participation in an activity over time—*duration*—has also been associated with increased positive academic and psychological (e.g., self-worth) outcomes. However, this pattern of association may not be as strong for some types of activities.

Another measure of participation is *breadth*—the range of different types of activities in which a child or adolescent participates. There are two ways to conceptualize breadth: concurrent participation in multiple activities or the range of different activities participated in over time. Overall, participating in a greater breadth of activities has been associated with more positive outcomes. Closely related to breadth of participation is the *total number* of activities. Unlike breadth of participation, the total number of activities is a summation of all activities irrespective of their types. Thus, the total number of activities could represent a total in the same type of activity (e.g., multiple sport activities) or a total across the types of activities. In general, a greater total number of activities has been associated with more positive outcomes.

Participation in out-of-school activities also differs by important sociocultural factors, including indicators of socioeconomic status, race, and sex. Higher parental income, education level, and occupational prestige (e.g., indicators of socioeconomic status) are associated with higher rates of child and adolescent activity participation. Research has also found differences in activity participation by race, with lower rates of participation among minority children and adolescents compared to their White counterparts. However, part of this difference is due to activity availability associated with income disparities; lower income, urban communities and schools tend to have fewer activities than middle-and upper-income, suburban communities. Finally, research findings regarding sex differences in out-of-school activity participation do not allow for decisive conclusions, although there is some evidence that females may participate in a greater number of activities.

Learning in Out-of-School Activities

What do children and adolescents learn in out-of-school activities? The answer to that question can be as varied as the different types and aims of the activities in which they participate. Given this variation then, this section provides a summary of learning within more general domains (e.g., academics) rather than on more nuanced findings (e.g., math grades). Additionally, a large portion of research on learning in out-of-school programs is based on adolescent samples, thus much of what is reviewed next comes from this research field; when possible, findings based on research with child samples will be noted.

Participation in out-of-school activities is associated with learning that can be collectively referred to as “21st-century competencies.” There are three domains of 21st-century competencies: cognitive, interpersonal, and intrapersonal. Participation in out-of-school activities has been associated with the development of *cognitive competencies*, including learning to engage in system-level thinking to address complex, real-world problems, increased creativity, and increased content-specific skills (e.g., technology). Participation has also been associated with learning *interpersonal competencies*, including learning teamwork, leadership, communication, and social skills. Similarly, research indicates that youth learn *intrapersonal skills* from participation, such as initiative, self-direction, time management, and enhanced psychological well-being.

Academic competencies are considered separately from 21st-century skills, although they could reasonably fit within the cognitive domain. Research in general indicates that participation in out-of-school activities is positively correlated with academic achievement and engagement, but the association may be attributable to preexisting differences (e.g., those doing well academically may be more likely to participate). Finally, participation in out-of-school activities has been associated with reduced problem and risk behaviors, such as lower delinquency rates and substance use.

It would be easy to get the impression that participation in any out-of-school activity promotes a host of competencies. However, there are several things to keep in mind. First, learning that occurs within out-of-school activities should, in theory, be connected to the aims and purposes of a program or organization, for example, academic activities are designed to enhance academic learning. Thus, the specific impact of participation on a competency is best conceptualized as the result of the particular activity aims. However, there is often an association between participation and learning not directly related to stated aims and purposes. This could be attributable to a program having multiple aims and purposes, youth participating in multiple programs or activities and thus they purposes, youth participating in multiple programs or activities and thus they may not differentiate learning across settings, or inadequate research designs and measures that allow for this level of specificity.

Second, there is evidence that different types of activities differ in their “profile” of learning experiences and opportunities. For example, sport activities appear to promote the development of initiative and perseverance, whereas community-oriented activities may promote the development of leadership and collaboration skills. Thus, it may be better to focus on the set of learning opportunities in an activity rather than any single one type of learning. Third, the potential impact of participation is moderated by the quality of the activity environment. Scholars have identified key features (i.e., quality) of activities that promote positive development, such as psychological safety. Research has only recently begun to investigate how quality relates to learning, but existing research suggests higher quality may indeed promote learning. However, this conclusion is based on relatively few published studies, which makes this conclusion tentative.

Methodological Challenges

Attributing specific learning to participation in out-of-school activities is fraught with challenges. Although grouping by activity types has provided a useful metric to examine learning experiences, it remains unclear how fine grained a distinction should be made (e.g., differentiating team and individual sports or simply grouping them all together as sports). Youth often participate in multiple types of activities during a given period or in different activities over time. Thus, isolating the effect of one particular type of activity is difficult and not possible without extensive assessments and research designs that would be cost prohibitive. Research to date has primarily been descriptive, pointing to a need for theory-driven studies. However, recent emerging qualitative ground theory research has started to provide testable hypotheses for future research.

At present, there exist measures of learning experience in activities and of the quality of the activities, but there are few strong outcome measures (e.g., teamwork skills) for many 21st-century competencies. Researchers also face statistic challenges, as many statistical procedures require large numbers of participants per activity (e.g., >40) to satisfy underlying test assumptions. However, many, and perhaps most, activities have far fewer participants, which excludes the types of analyses that could isolate activity effects. As researchers continue to find ways to address these, and many other challenges, researchers will be better able to differentiate how and what types of learning occurs within out-of-school activities.

David M. Hansen

See also Adolescence

Further Readings

Durlak, J., & Weissberg, R. (2007). The impact of afterschool programs that promote personal and social skills. Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Hansen, D. M., Skorupski, W. P., & Arrington, T. L. (2010). Differences in developmental experiences for commonly used categories of organized youth activities. Journal of Applied Developmental Psychology, 31(6), 413–421.

Larson, R., & Angus, R. (2011). Adolescents' development of skills for agency in youth programs: Learning to think strategically. Child Development, 82(1), 277–294. doi:10.1111/j.1467-8624.2010.01555.x

Larson, R. W., Hansen, D. M., & Moneta, G. (2006). Differing profiles of developmental experiences across types of organized youth activities. Developmental Psychology, 42, 849–863.

Mahoney, J. L., Larson, R. W., & Eccles, J. S. (Eds.). (2005). Organized activities as contexts of development: Extracurricular activities, after school and community programs. Mahwah, NJ: Erlbaum.

Marsh, H., & Kleitman, S. (2002). Extracurricular school activities: The good, the bad, and the nonlinear. Harvard Education Review, 72, 464–511.

National Research Council. (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. Washington, DC: The National Academies Press.

National Research Council, & Institute of Medicine. (2002). Community programs to promote youth development. Washington, DC: National Academy Press.

P

Hyung Won Kim Hyung Won Kim Kim, Hyung Won

p Value

p Value

1195

1198

p Value

In education disciplines, quantitative studies conducted via statistical hypothesis testing often depend on the calculation of a probability value or *p value*. Statistical hypothesis testing is a procedure of evaluating suppositions, which are assumptions about certain characteristics of a population, using a sample from that population. The goal of testing a statistical hypothesis is to determine whether the sample evidence challenges the study’s *null hypothesis* (i.e., there is no observed effect) and supports the *alternative hypothesis* (i.e., there is an observed effect).

The idea of *p* value arises in conjunction with the α (or *significance*) level associated with a statistical test. The α level is the threshold probability value that is chosen for the test; it is commonly .05 or .01 (equivalently, 5% or 1%). The *p* value of the sample test is compared with the chosen α level. A small *p* value indicates the unlikeliness of obtaining the given result if the null hypothesis was true. Therefore, if the associated *p* value is less than the α level, the sample data provide evidence for rejecting the null hypothesis for the entire population.

The use of a *p* value typically involves the following steps:

1. state the null hypothesis (H_0) and alternative hypothesis (H_a),
2. set the α level,
3. collect data,
4. compute the test statistic and associated *p* value from the collected data,
5. compare the *p* value with the significance level α, and

6. make conclusions as to whether the null hypothesis should be rejected.

To understand statistical inference, it is important to understand how to find, use, and interpret p values in a given context.

Applications

Hypothesis tests (z test, t test, F test, and chi-square test) use p values regardless of the test type. In many fields of social science, including education, p values are commonly used to test statistical hypotheses for inferential statistics. Most researchers use statistical software such as SPSS, SAS, Minitab, Excel, and R, or tools available on many websites, to calculate p values.

A Simplified Example

Suppose the following as an example: An innovative teaching approach has been developed for a college-level elementary statistics course. The efficacy of this teaching approach over the traditional approach is analyzed via data gathered from student examination scores. Traditionally, the mean score is 72 with a standard deviation of 9.6. The goal is to provide evidence that the mean score of the students who learned elementary statistics by the new approach is higher than 72. Taking a study sample of 64 exam scores from students taught with the new teaching approach, a sample mean (μ) of 73.5 is obtained.

Note that the sample mean score of 73.5 is greater than 72 by 1.5. To determine whether the sample provides evidence for claiming that this (seeming) improvement is significant, the following steps can be used:

> *State the hypotheses*: First, define a random variable X to be the exam score for a student who learned elementary statistics with the new approach. Then, the null hypothesis states that $\mu = 72$ (the mean exam score for those who learned elementary statistics via the new teaching approach is 72), and the alternative hypothesis states that $\mu > 72$.
> *Set the* α *level*: Choose $\alpha = .05$. (*Note:* Choose a different α level. However, once the α value is fixed, it should not be changed after find the p value is obtained.)
> *Compute the test statistic*: Consider the sampling distribution of the variable X. The central limit theorem allows the normality of the sampling

distribution because the sample size of 64 is greater than 30. In the sampling distribution, the test statistic (*z* statistic) of 73.5 is:

$$\frac{73.5-72}{9.6/\sqrt{64}}=1.25.$$

Compute the p value: In the context of the problem, the *p* value is the probability that a replicated sample mean will be greater than 73.5 (or equivalently, the probability that a test statistic is higher than 1.25). The *p* value associated with the *z* statistic of 1.25, which can be found using a standard normal table or statistics software, is .1057 (equivalently, 10.57%).

Compare the p value with the α *level*: To reject the null hypothesis, a *p* value has to be less than the α level. In this example, the *p* value of .1057 is greater than the α level of .05. Therefore, at α = .05, the *p* value of .1057 is considered "not small enough to reject the null hypothesis." That is, the data do not provide evidence to support the alternative hypothesis.

The study suggests that the sample data do not provide significant evidence to conclude that the new teaching approach for elementary statistics is more effective than the traditional approach.

How *p* Values Are Interpreted

A small *p* value, then, indicates that the statistic from the sample in hand diverges significantly from the parameter value suggested in the null hypothesis. This serves as evidence against the null hypothesis. To determine how small a *p* value would be considered significant, the *p* value has to be compared with an α level. If the *p* value is less than the α level, one can reject the null hypothesis. Alternatively, if the *p* value is greater than the α level, the statistic from the sample would not be considered to have diverged significantly from the parameter suggested by the null hypothesis.

The p Value in Relation to the Type of Test

As seen earlier, the *p* value of a sample directly depends on the test statistic. It also depends on the type of hypothesis testing: whether the test is one tailed or two tailed. (The earlier example is of a one-tailed test.) Customarily, the two types of tests are defined as follows:

- A significance test is *one tailed* if the alternative hypothesis opens in a single direction—either upward (e.g., μ > 72) or downward (e.g., μ < 72).
    - If the alternative hypothesis opens *upward* (right-tailed test), the *p* value is indicated by the area to the right of the test statistic under the curve of the sampling distribution.
    - If the alternative hypothesis opens *downward* (left-tailed test), the *p* value is indicated by the area to the left of the test statistic under the curve of the sampling distribution.
- A significance test is *two tailed* if the alternative hypothesis states that a parameter differs from the value suggested by the null hypothesis (e.g., μ = 72).
    - The *p* value is indicated by the sum of the two areas on each tail under the curve of the sampling distribution.

Example of a p Value in a Two-Tailed Test

In another example, a team of researchers is interested in finding out if bilingual fourth graders perform differently on mathematics word problems than fourth graders in general. Suppose that the mean score for a particular exam that measures a student's ability for word problems is 72.8 with a standard deviation of 8.4. Data were collected from 144 bilingual fourth graders to see whether their performance was significantly different from that of fourth graders in general. The data can be statistically analyzed following the steps below.

> *State the hypotheses*: Define a random variable *X* to be the exam score for the fourth graders who are bilingual. Then, the null hypothesis states that μ = 72.8, and the alternative hypothesis states that μ ≠ 72.8.
> *Set the* α *level*: Choose α = .01.
> *Compute the test statistic*: Consider the sampling distribution of the variable *X*. The normality of the sampling distribution is assumed by the central limit theorem. In the sampling distribution, the test statistic of the sample mean 74.7 is computed as $\dfrac{74.7 - 72.8}{8.4 / \sqrt{144}} = 2.714.$
> *Compute the p value*: The sample data gave the mean score of 74.7, which is different than 72.8 by 1.9. You have to determine whether the distance of 1.9 standard deviations is significant enough to reject the null hypothesis. The test statistic of 2.714 directly determines the *p* value. This test is two

tailed because the alternative hypothesis (μ ≠ 72.8) states “different from” rather than “less than” or “greater than.” This means you consider the sum of the two areas by taking both the left and the right tails of the density curve. The left and right tails would be marked off at 2.714 and 2.714, respectively. Using software, one finds that the area of each tail is 0.042 (the two areas are the same by symmetry). The sum of the two areas, .0074 (.0037 + .0037), is the *p* value of this sample data.

Compare the p value with the α *level*: The *p* value of .0074 is less than the α level of .01. This means that a mean score of 74.7 is considered unlikely to occur when the actual mean of scores for bilingual fourth graders is 72.8. Therefore, at α = .01, the *p* value of .0074 is understood as “small enough to reject the null hypothesis,” that is, the data in this example provide enough evidence to support the alternative hypothesis.

The study suggests that the sample data provide statistically significant evidence for a difference in performance on math word problems between fourth graders who are bilingual and fourth graders in general.

Misinterpretations and Negative Views

Students often develop misconceptions regarding the *p* value, such as it being the probability of the null hypothesis or the probability of the data having arisen by chance. The statistics education reformers of the late 1900s and early 2000s criticized the use of *p* values as an inadequate measure of evidence in hypothesis testing. The criticism revolved around the dichotomous nature of the *p* value and its lack of precision with large samples. For example, *p* values can inflate the evidence against the null hypothesis if samples are large due to the improbability of the null hypothesis being exactly true. Some scholars have suggested that inferential statistics–based research should depend less on *p* values and more on confidence intervals. In the culture of inferential statistics–based research in education, *p* values have been used widely in hypothesis testing. Thus, the inferential method of using *p* values has played a significant role in knowledge development in education.

Hyung Won Kim

See also Alpha Level; Hypothesis Testing; Inferential Statistics; Significance; Type I Error

Further Readings

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence. Journal of the American Statistical Association, 82(397), 112–122.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? Journal of Statistics Education, 17(2). Retrieved from www.amstat.org/publications/jse/v17n2/castrosotos.html

Fidler, F., & Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. Paper presented at the 55th session of the World Congress of the International Statistical Institute, Sydney, Australia. Retrieved from http://iaseweb.org/documents/papers/isi55/FidlerCumming.pdf

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. Theory & Psychology, 18(1), 69–88.

Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. In C. Reading (Ed.), Proceedings of the Eighth International Conference on Teaching Statistics [CDROM]. Ljubljana, Slovenia: IASE.

Leon, G. A., & Frankfort, N. C. (2014). Essentials of social statistics for a diverse society. Los Angeles, CA: SAGE.

Marden, J. H. (2000). Variability in the size, composition, and function of insect flight muscles. Annual Review of Physiology, 62(1), 157–178.

Nelder, J. A. (1999). From statistics to statistical science. Journal of the Royal Statistical Society: Series D (The Statistician), 48(2), 257–269.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Series A, 231, 289–337.

Dan He Dan He He, Dan

Hongling Lao Hongling Lao Lao, Hongling

Paper-and-Pencil Assessment Paper-and-pencil assessment

1198

1200

Paper-and-Pencil Assessment

Paper-and-pencil assessment refers to traditional student assessment formats such as written tests and also to standardized tests that ask students to use pencils to fill in bubbles on a scannable answer sheet. Standardized tests are now commonly administered on computers, but classroom assessment usually requires students to submit written responses on paper. In the classroom, paper-and-pencil assessment frequently refers to tests scored objectively, which are meant to measure memorized knowledge and lower levels of understanding, as compared with performance-based assessment, which is meant to measure deeper understanding through skills and ability. The following sections discuss how assessments are developed and describe the current and future applications for paper-and-pencil assessment.

Assessment Development

Assessments are needed to collect information for making decisions, and the design of an assessment depends on the intended use of the assessment results. The typical cycle of assessment begins by specifying its purpose, including the information to be collected and how it will be used. Often based on a table of specifications or a blueprint that identifies the objectives to be covered and the relative weights for each objective or domain, a paper-and-pencil assessment is essentially a test that is a collection of items, questions, or tasks.

Item development is a crucial step that determines the validity and reliability of an assessment. Assessment developers, whether teachers or psychometricians, face the challenge of making an invisible trait or characteristic (what social scientists call a *construct*) visible through test takers' responses. Consequently, an underlying assumption is made for each item that the response is due only to the level of that construct; nothing else is contributing to the item response. This is a strong assumption that requires justification and empirical support.

The test medium (e.g., paper-and-pencil, computer, and tablet) needs to be taken into consideration as an assessment is developed. This is because the property of the test medium sets the behavior boundary that is permissible and recordable for collecting the information of interest. For example, it is inappropriate to use paper-and-pencil assessment to evaluate a procedural ability such as swimming or typing. Neither swimming nor typing ability is measurable via paper and pencil. For information that is difficult to be presented via language or symbols (e.g., knowledge of violin playing), paper-and-pencil assessment is rarely the medium of choice.

Paper-and-Pencil Assessment in the Classroom

Paper-and-pencil assessment formats in the classroom include multiple-choice questions, matching items, true–false questions, fill in the blanks, short answer, and sometimes essay questions. They are usually

- scored objectively: There is a single correct answer, and no judgment is required to determine whether an answer is acceptable, and
- "selection" items: The correct answer is there to be chosen from among several possibilities.

But they may also be

- scored subjectively: There are different possible ways to answer, and a teacher's expert judgment determines whether an answer is acceptable, and
- "supply" items: The answer is not on the paper to be selected, and students must provide it themselves.

An advantage of paper-and-pencil assessment in the classroom is the straightforwardness of the test medium, especially compared with the initial expense and maintenance and training issues associated with computers and tablets. Moreover, the majority of students have experience interacting with paper and pencil, which serves as transferrable experience in taking tests using this medium.

Paper-and-Pencil Assessment in Standardized Testing

When a standardized test uses a paper-and-pencil assessment format, examinees read items printed on a test booklet and write down or mark their responses on an answer sheet or test booklet. A “Number 2” pencil (referring to the softness grade of the lead) is usually required to ensure that scoring machines can adequately detect the responses. Standardized tests can be administrated to a large group of examinees at the same time, and many thousands of answer sheets can be scored quickly and almost completely without error. In some cases (e.g., on college admissions tests), paper-and-pencil tests are used to assess skill proficiency as well as aptitude, but typically they are used for quick and fairly direct assessment of knowledge.

The Future of Paper-and-Pencil Assessment

Paper-and-pencil assessment has existed for hundreds of years. As a classroom format, this tried-and-true approach remains popular, even as tablets, smart phones, and computers have found their way into most schools, and is used by teachers to support or deliver their instruction. With the rapid development and popularization of Internet and computer technology in recent years, however, a growing number of schools and educational assessment agencies are adopting new formats within computer-based assessment to deliver tests. The 2016 academic year in the United States marked the first time that more states administrated statewide summative assessments for Grades 3–8 by computer-based rather than paper-and-pencil format. In the world of large-scale standardized testing, paper-and-pencil assessment may soon be extinct.

One reason for the growing popularity of computer-based assessment is test security. The administration of paper-and-pencil assessment includes making printed copies, transporting and storing these copies before the testing session, distributing them to test takers, and collecting the responses for scoring. It is a long, difficult-to-monitor process, and unauthorized access and tampering can occur during any step. With computer-based assessment, the electronic test data are transmitted and stored in a relatively secure system. One security threat unique to computer-based testing, however, is hacking. Experiences from the recent administration of state educational assessments in the United States suggest that testing networks are at risk to attacks by hackers. Such attacks have resulted in the temporary shutting down of computer-based assessment systems in some states.

in some states.

Other advantages to computer-based assessment include ease of scoring and reporting as well as lower overall cost. Paper-and-pencil assessment scoring, once performed by hand, is now a batch-scanning process by a machine that can process thousands of answer sheets in a short time frame. But collecting answer sheets and preparing them for scanning is a time-consuming process. With computer-based assessment, a test is scored automatically and immediately after an examinee completes it, and results can be reported online. The short turnaround time for test results helps the examinee, teachers, parents, and educators get quick feedback. The costs of paper-and-pencil assessment include printing test paper, shipping, storage, machine scoring, and personnel training and administration. Computer-based tests eliminate many of these expenses. Once the technology infrastructure of a testing system is built and ready to use, computer-based testing is a less expensive alternative to traditional paper-and-pencil assessment.

Dan He and Hongling Lao

See also Computerized Adaptive Test; Computer-Based Assessment; Test Development; Test Security; Tests

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: AERA.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments. (1986). Guidelines for computer-based tests and interpretations. Washington, DC: Author.

Kingston, N. M. (2008). Comparability of computer-and paper-administered multiple-choice tests for K–12 populations: A synthesis. Applied Measurement in Education, 22(1), 22–37.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. International Journal of Testing,

6(1), 1–24.

Wang, H., & Shin, C. D. (2009). Computer-based & paper-pencil test comparability studies. Test, Measurement & Research Services Bulletin, 9, 1–6.

Arthur J. Cunningham Arthur J. Cunningham Cunningham, Arthur J.

Paradigm Shift Paradigm shift

Paradigm Shift

A *paradigm shift* is a fundamental conceptual transformation that accompanies a change in accepted theory within a scientific field. The term was introduced by the historian and philosopher of science Thomas S. Kuhn in his influential 1962 book, *The Structure of Scientific Revolutions*. Kuhn referred to the network of conceptual, theoretical, and methodological commitments shared by scientists in a given field as a *paradigm*. He argued that a significant change in accepted theory is accompanied by profound changes in this network of group commitments. The result is a fundamental transformation in the way scientists view the world and pursue their research. This entry explains the notions of paradigm and paradigm shift as Kuhn presented them.

Kuhn’s Definition of Paradigm

Kuhn observed that any developed field in the natural sciences rests upon a body of established theory. In part, this means that a field’s practitioners accept an array of time-tested laws, models, and methods that serve as shared tools for investigating and explaining natural phenomena. But Kuhn emphasized that a body of established theory encompasses a wider range of group commitments as well. In learning to employ the laws and methods that are central to their field, scientists also absorb a specialized conceptual framework, together with a broad set of convictions about what the world is like and about how to practice science. A discipline’s reliance on a body of established theory thus implies a whole network of shared commitments that decisively shape the character of scientific research and education within the field.

As an example of how these group commitments are instilled by scientific training, consider the physics student learning Newton's second law of motion, typically expressed by the equation *f* = *ma* (force = mass × acceleration). What this law means to a physicist, and what a student of physics must learn to understand it, is not adequately captured by this equation alone. A physicist's grasp of this law includes the ability to apply it to a range of different physical scenarios—that is, to see a variety of different phenomena as manifestations of the same theoretical pattern. Students develop that ability by learning to emulate standard examples that show how the law can be applied to particular problems. Through such training, Kuhn said, scientists acquire a highly specialized and discipline-specific way of seeing: They come to see nature in terms of the theoretical concepts and relationships exemplified in the discipline's standard examples.

It should be noted that in 1962, when *The Structure of Scientific Revolutions* was published, *paradigm* was a fairly obscure word that meant "shared model" or "exemplar." Kuhn had this established meaning in mind when he first employed the word in his book: He referred to standard examples of a theory's applications as *paradigms* to emphasize that such examples serve as shared models for scientists to emulate. In the course of the book, however, Kuhn inadvertently gave the word a new meaning by using it to refer to the entire network of group commitments that underlie a discipline.

Paradigm Shifts

In Kuhn's view, the relationship between scientific theory and observation is a two-way street. Theory is informed by and responsive to observed facts, but observation does not occur in a vacuum. What scientists observe is necessarily informed by their disciplinary training. Indeed, every aspect of an experiment—from which instruments are used and how they are employed, to which data are collected and how they are analyzed—is influenced by the conceptual framework and the outlook on nature that a scientist brings to the scientist's work.

For this reason, Kuhn thought that any significant change in theory implies a corresponding change in the way science is practiced. A change in the laws a scientific community accepts, for example, implies a change in the community's conceptual framework and a corresponding change in the way that community sees the world. Seeing the world differently, scientists will afterward interact with it differently, employing methods and instrumentation befitting their altered understanding of nature. Even their ideas about the problems to be solved in their field and the standards for an acceptable solution will be altered. In short, theory change involves a significant alteration in the entire network of group commitments that characterizes a discipline. Kuhn referred to this collective transformation in the way scientists understand their discipline and the world as a *paradigm change* or *paradigm shift*.

Kuhn’s Legacy

Although Kuhn discussed paradigms and paradigm shifts exclusively in connection with the natural sciences, these notions subsequently came to be used much more widely. The term *paradigm shift* in particular is now used quite generally to refer to any episode in which a conceptual framework that was formerly dominant within a field, industry, institution, or group is replaced by a new conceptual framework, leading to a corresponding change in outlook and practice.

Important parts of Kuhn’s legacy include the idea that there is no such thing as empirical research in the absence of some theoretical framework, and the idea that the collection and interpretation of data is always informed by researchers’ disciplinary training. The moral is sometimes drawn from Kuhn’s work that empirical research is unavoidably *biased*, but Kuhn would have rejected this way of putting the point. Talk of bias suggests a failure of some kind, as though in letting their research be shaped by existing knowledge, researchers are failing to be properly objective. But Kuhn thought there need be no failure involved. He insisted that a background of established theories, concepts, and methods that researchers can confidently employ is a prerequisite for systematic empirical study in any field; researchers who could not take for granted any ideas or methods could hardly do meaningful research. Any conception of scientific objectivity that fails to make room for the influence of background ideas is therefore not a useful ideal but a misplaced and unhelpful picture of what empirical research can and should be.

Arthur J. Cunningham

See also Epistemologies, Teacher and Student; Objectivity; Positivism

Further Readings

Further Readings

Kuhn, T. S. (1977). Second thoughts on paradigms. In T. S. Kuhn (Ed.), The essential tension: Selected studies in scientific tradition and change (pp. 293–319). Chicago, IL: University of Chicago Press.

Kuhn, T. S. (2012). The structure of scientific revolutions (4th ed.). Chicago, IL: University of Chicago Press. (First edition published in 1962. Editions since the second include an important postscript by Kuhn; the fourth edition includes a useful introductory essay by Ian Hacking.)

Bruno D. Zumbo Bruno D. Zumbo Zumbo, Bruno D.

Parameter Invariance Parameter invariance

1201

1203

Parameter Invariance

Although parameter invariance may at first glance appear to be an arcane mathematical or statistical concept, in practice, it is far from that. Partly because of parameter invariance, statistical model–based measurement using item response theory (IRT) is one of the most popular current methodological frameworks for modeling data from assessments.

As is widely appreciated in statistics courses, the word *parameter* indicates that parameter invariance refers to population quantities, whose values are to be estimated with data collected within a random sampling design. Parameters can in this context refer to the set of item parameters (item difficulty, discrimination, and/or guessing) and the set of examinee parameters (the examinee test scores, or theta [θ] scores, implied by the IRT model) that are tied to a particular measurement model. The word *invariance* indicates that parameter values are identical in separate examinee populations or across separate measurement conditions, which is commonly investigated through estimated parameter values from different calibration samples. What this implies is that parameter invariance is only relevant when comparing groups or measurement conditions. If there is only one population or condition, invariance is not relevant. That is, the matter of parameter invariance addresses the question of whether test scores or item parameters are equally valid for different populations of examinees or different measurement conditions. If parameters are not invariant, the statistical foundation for inferences is not identical across the populations or measurement conditions and, hence, the inferences are not generalizable across those to the same degree. Note, however, that parameter invariance denotes an absolute ideal state that holds only for perfect model fit, and any discussion about whether there are “degrees of invariance” or whether there is “some invariance” is technically inappropriate. And as noted earlier, the question of whether parameter invariance exists in any single population is illogical because parameter invariance requires at least two examinee populations or two measurement conditions for parameter comparisons to be possible and meaningful.

Implications of Parameter Invariance

Some of the most important advances and advantages of IRT over other statistical models of measurement is its direct application of parameter invariance in test equating or linking, computer adaptive testing, and cognitively diagnostic assessment. In fact, testing populations are often inherently heterogeneous, and invariance becomes important in this context as well. For example, a population of schoolchildren may consist of discernible subpopulations involving background, culture, or languages that are relevant to the construct being measured (e.g., oral language expression or spelling). Parameter invariance implies that an item in a test has identical difficulty and discrimination for each of the discernible subpopulations. In short, parameter invariance implies that the same IRT model, with identical item parameter values, holds true for all corners of the data in the population.

The concept of item parameter invariance then stipulates that with a sufficiently large pool of examinees, item parameters are independent of the ability distribution of the examinees. Likewise, the concept of person parameter (ability, or θ) invariance stipulates that with a sufficiently large set of items, respondents' ability score and overall distribution of the ability score are independent of the set of test items.

Parameter Invariance With Other Statistical Practices

Parameter invariance is a concept that is also possible in commonly used statistical models such as ordinary linear regression. That is, if we imagine a scatterplot with an *x* and *y* axis, we can also imagine that there is a cloud of population data points and a linear regression line defined for that cloud of data points in the *x*–*y* scatter diagram. If the population regression equation is $y = 10 + 3x$, then one will always get the same regression line regardless of the subpopulation of points along the *x* axis. This is a simple consequence of the definition of a regression line fitting perfectly in a population. In fact, this concept is so fundamental and idealized that many professors do not teach it in their regression courses. A student can ask, “Do we ever have a perfectly fitting regression line in a population?” The response is that we do not, so although parameter invariance applies in regression analysis, it is of no practical consequence to data analysts.

Parameter Invariance in IRT Involves the Latent Variable, θ

This notion of parameter invariance in regression is of the same mathematical flavor that is discussed in IRT. Mathematically, parameter invariance is a simple identity for parameters that are on the same scale. The essential phrase in the previous sentence is “on the same scale.” IRT involves an arbitrarily scaled latent random variable: the examinees’ scale scores (θ). The latent scale in IRT models is arbitrary, so that unequated sets of model parameters will be invariant only up to a set of linear transformations specific to a given IRT model. When estimating these parameters in unidimensional IRT models with calibration samples, this indeterminacy is typically resolved by requiring that the latent variable be normally distributed with a mean of 0 and standard deviation of 1. Once estimated values of the parameters for different populations are available on their respective scales, it is of interest to determine the type of relationship that exists between them as a yardstick to assess whether the same IRT model with the same parameter values is likely to hold in both examinee populations and measurement conditions (i.e., whether parameter invariance is likely to hold). In addition, work in score equating, differential item functioning, and item parameter drift deals with lack of invariance and the effects introduced thereby.

Bruno D. Zumbo

See also Computerized Adaptive Testing; Item Response Theory

Further Readings

Breithaupt, K., & Zumbo, B. D. (2002). Sample invariance of the structural equation model and the item response model: A case study. Structural Equation Modeling, 9, 390–412. doi:10.1207/S15328007SEM0903_5

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (pp. 147–200). New York, NY: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. Alberta Journal of Educational Research, 49, 264–276.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether invariance holds for IRT models: When Pearson correlations are not enough. Educational and Psychological Measurement, 64, 588–599.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. Educational and Psychological Measurement, 66, 63–84.

Zumbo, B. D. (2013). On matters of invariance in latent variable models: Reflections on the concept, and its relations in classical and item response theory. In P. Giudici, S. Ingrassia, & M. Vichi (Eds.), Statistical models for data analysis (pp. 399–408). New York, NY: Springer.

Bo Hu Bo Hu Hu, Bo

Parameter Mean Squared Error Parameter mean squared error

1203

1205

Parameter Mean Squared Error

The parameter mean squared error (MSE), also known as empirical mean squared error, indicates the deviation of an estimated value from the expected value of a given parameter. The lower the MSE is, the better accuracy an estimated value or an estimation method presents. Mathematically, it is formulated as the average of the squared deviations across a certain number of estimations; thus, the MSE is always a positive value. To calculate the MSE, one needs to know the expected values of the parameters, which normally are unknown in statistical analysis. For this reason, the MSE is commonly used as an evaluation criterion in conjunction with the *Markov chain Monte Carlo* (MCMC) method, in which data are randomly sampled from probability distributions rather than collected from the real world. This entry introduces the definition and calculation of MSE and, through an example, discusses its usefulness within MCMC methods.

Calculation of the MSE

If an estimation procedure is repeated a number of times, one should be able to calculate the average squared deviation of an estimator from the expected value of a given parameter across all replications. Let x denote the expected value of a parameter and x_i denote the estimator of x from the ith, $i = 1, \ldots , T$, replication. Then, the MSE for the estimator can be written as

$$\mathrm{MSE} = \frac{1}{T} \sum_{i=1}^{T} \left(\hat{x}_i - x \right)^2 .$$

At times, one may need to estimate a set of parameters. For example, in item response theory calibration, the ability parameters for a group of examinees need to be estimated. Let *X* be a vector of the expected values of *N* parameters and X_i be a vector of estimators for the *i*th, *i* = 1, … , *T*, replication. Then, the MSE for the estimators can be written as

$$\text{MSE} = \frac{1}{T}\frac{1}{N}\sum_{i=1}^{T}\sum_{j=1}^{N}(\hat{x}_{ij} - x_j)^2,$$

where x_{ij} and x_j are the *j*th elements in X_i and *X*, respectively.

There are also occasions in which researchers are interested in the accuracy of the output of a function with respect to its expected value. For example, a psychometrician attempts to examine the accuracy of the equating results obtained by using the estimated linking coefficients and related equating functions. Let f_s be the “true” function of *s*, which is built upon the expected values for all related coefficients. Let f_{is} denote the estimated function of *s*, in which all coefficients are estimators from the *i*th, *i* = 1, … , *T*, estimation. Then, the MSE for the estimated function can be expressed as

$$\text{MSE} = \frac{1}{T}\sum_{i=1}^{T}\left[\hat{f}_i(s) - f(s)\right]^2.$$

In a study using MCMC simulation methods, the MSE is typically viewed as an index that summarizes the total errors occurring during a given statistical process (e.g., estimation, equating, and scaling). In fact, based on the types of source, there exist two types of error: *systematic* and *random*. The former relates to constant inaccuracy and is also known as *bias*, whereas the latter is unpredictable and occurs only by chance. Correspondingly, the MSE can be further broken down into two components to represent the systematic and the random error, respectively. Equation 1, for instance, can be rewritten as

$$\mathrm{MSE} = \frac{1}{T}\sum_{i=1}^{T}(\hat{x}_i - x)^2$$

$$= (\bar{x}_i - x)^2 + \frac{1}{T}\sum_{i=1}^{T}(\hat{x}_i - \bar{x}_i)^2,$$

where

$$\bar{x}_i = \frac{1}{T}\sum_{i=1}^{T}\hat{x}_i.$$

As shown in Equation 4, the MSE is the sum of squared bias, $(x_i - x)^2$, and sample variance, . The square roots of the sample variance are the empirical standard errors of the estimator, which indicate the consistency of the estimator and estimation methods.

The MSE in an MCMC Simulation

The MCMC method is named after the gambling location in Monaco because the key to the technique is to draw random outcomes (i.e., outcomes occurred by chance) from probability distributions. The objective of drawing random outcomes is achieved by a stochastic process with the Markov property, termed the *Markov chain*. One of the crucial properties of the Markov chain is that once the chain reaches its equilibrium, the outcomes of a variable follow its specified distribution. Therefore, the outcomes constructing the equilibrium distribution of the Markov chain can be treated as the random data sampled from the desired distribution of the variable. The MCMC method can simply be called a stimulation.

In a study using the MCMC method, the characteristics of the data (e.g., mean, variance, skewness, and missing) and the values for parameters are known. The MCMC method enables researchers to draw random samples from a specified distribution for the parameters of interest. Data can be generated based on the parameters and the particular statistical model. The fact that researchers know the expected value of the parameters beforehand makes it possible for them to calculate the MSE.

As an example, a researcher conducts an MCMC simulation to compare two estimation methods, expectation–maximization and maximum a posteriori, in estimating a Rasch model. The assessment data under a Rasch model are explained by two parameters: examinee ability (θ) and item difficulty (*b*). The researcher creates a setting in which a 30-item test is administered to 500 examinees. To obtain the data, he first generates 500 ability parameters and 30 item-difficulty parameters from a normal distribution *N*(0,1). Now, he has a vector θ for the expected values of 500 ability parameters and a vector *B* for the expected values of 30 difficulty parameters. Based on the Rasch model and the values of each parameter, he further generates the response data of 500 examinees on the 30-item test. Then, he conducts the calibrations using the expectation–maximization and MAP methods, respectively. These analyses give him the estimated vectors for both θ and *B*.

To evaluate the efficacy of the two estimation methods, the researcher examines how accurately the results from each method can recover their expected values. Therefore, he repeats the entire process 100 times and then calculates the MSEs for both parameters. Specifically,

$$\mathrm{MSE}_\theta = \frac{1}{100}\frac{1}{500}(\hat{\theta} - \theta)^2$$

and

$$\mathrm{MSE}_b = \frac{1}{100}\frac{1}{30}(\hat{B} - B)^2.$$

The estimation method that produces a smaller MSE_θ and MSE_b may be seen as having the advantage in terms of estimation accuracy and stability.

Bo Hu

See also Estimation Bias; Markov Chain Monte Carlo Methods; Parameter Random Error

Further Readings

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97–109. doi:10.2307/2334940

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices (3rd ed.). New York, NY: Springer-Verlag.

Lehmann, E. L., & Casella, G. (2006). Theory of point estimation (3rd ed.). New York, NY: Springer.

Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2007). Mathematical statistics with applications (7th ed.). Belmont, CA: Thomson Higher Education.

Alex Brodersen Alex Brodersen Brodersen, Alex

Can Shao Can Shao Shao, Can

Ying Cheng Ying Cheng Cheng, Ying

Parameter Random Error Parameter random error

1205

1206

Parameter Random Error

When fitting a statistical model and estimating parameters, the variability in parameter estimates due to random sampling is called parameter random error. This is in contrast to systematic bias in parameter estimates, which may arise from model misspecification or convenience sampling. This entry describes parameter random error in the context of educational measurement, specifically as it relates to item response theory (IRT).

Source

In most uses of IRT, item parameters are estimated from a calibration sample prior to operational use. If we denote the population item parameters as Γ, then item parameter estimates Γ that arise from samples 1 to *m* that are randomly drawn from a population of test takers will be Γ_1, Γ_2, … , Γ_m. The variability among these sets of estimates is the random error for Γ.

Item parameter estimates are often treated as the population values and are used in test construction, ability estimation, linking/equating, or item selection in computerized adaptive testing. Traditionally, IRT was primarily used for the development of large-scale educational achievement and ability tests in which items were calibrated on large samples (e.g., several thousand test takers). In these cases, parameter random error was assumed to be negligible. Recently, interest in and use of IRT has increased dramatically, and the applications of IRT models have extended to settings where large sample sizes may be unavailable.

In these cases, parameter random error should not be ignored.

Effects

Whether items are part of a fixed-form assessment or a computerized adaptive testing model, they are often selected for use partially on the basis of their item parameters. For example, *maximum Fisher information* is a commonly used item selection algorithm that tends to select items with a large discrimination parameter. Likewise, items with high discrimination parameters are also frequently selected in fixed-form assessments in order to build a desirable test information function. Because of the relationship of test information with the standard errors of maximum likelihood ability estimates, neglecting parameter random error can lead to underestimation of standard errors. Although several methods have been proposed to correct for parameter random error in ability estimation, neglecting this additional error may lead to misinterpretation of ability estimates.

Outside of ability estimation and test construction, parameter random error also has implications in linking and equating. Studies have found that random error in parameter estimates can produce poor estimates of linking coefficients using common equating methods. However, other studies have found that certain methods (e.g., test response function linking/equating method) are fairly robust to item calibration error in the case of a sufficiently large number of common items between the forms.

Alex Brodersen, Can Shao, and Ying Cheng

See also Equating; Item Information Function; Item Response Theory; Score Linking; Simple Random Sampling

Further Readings

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. Applied Measurement in Education, 7(3), 171–186. doi:10.1207/s15324818ame0703_1

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. Applied

Psychological Measurement, 25(1), 39–52. doi:10.1177/01466216010251003

Patton, J. M., & Cheng, Y. (2014). Effects of item calibration error on applications of item response theory. In Y. Cheng & H. H. Chang (Eds.), Advancing methodologies to support both summative and formative assessments (pp. 89–105). Charlotte, NC: Information Age Publishing.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. Psychometrika, 55(2), 371–390. doi:10.1007/BF02295293

Joan Newman Joan Newman Newman, Joan

Parenting Styles Parenting styles

1206

1207

Parenting Styles

Parenting style refers to the behaviors and relationship that parents display in a relatively constant manner as they interact with their offspring throughout their development. Although a parent’s behavior can fluctuate according to the situation, researchers have identified and measured some consistencies that characterize a particular parent and differentiate one parent from another. These consistencies can be termed the *parenting style*. Studies have related these to developmental outcomes of children and adolescents such as delinquency, self-esteem, substance abuse, and academic achievement.

In the 1960s, Diana Baumrind conceptualized parenting style according to two dimensions: the amount of control exerted and responsiveness to the child. Three types of parenting style were identified using these dimensions: authoritarian (high control and low responsiveness), authoritative (high control and high responsiveness), and permissive (low control and high responsiveness). Later researchers added the category of neglectful (low control and low responsiveness).

Although the original research was based on interviews and observation of mothers with their children, later researchers have developed questionnaires with which parenting style can be reported and measured. Data from the questionnaires can be analyzed either as parenting style categories or can be converted into continuous scales by which one or all parenting styles can be examined. Using these parenting style categories, many researchers have reported that an authoritative parenting style is most often related to positive outcomes.

Some psychologists have advocated training to help parents employ authoritative parenting. However, the characteristics of the children have a great influence on parenting. However, the characteristics of the children have a great influence on the type of parenting that their parents employ. Children with behavioral difficulties or achievement problems may require or elicit particular types of parenting. Also children being raised in at-risk settings may benefit from particular styles of parenting.

Cross-cultural research has challenged the accounts of parenting style effects that have been based on both theoretical models and data gained from American families. In some cultures, especially East Asian cultures, authoritarian parenting has been shown to be more prevalent and also often related to desired outcomes. Alternative conceptualizations of parenting style have been published that reflect non-Western values and practices. These show that loving care for a child can be shown by a variety of parent behaviors. Because of this, teachers and other practitioners have been advised to avoid recommending a particular parenting style as desirable for all children.

Modifications of the concept and measurement of parenting style have been made. For example, the specific parenting behaviors that a parent might employ in a particular context have been distinguished from a parenting style that conveys the more general and enduring manner in which those behaviors are displayed.

Joan Newman

See also Cross-Cultural Research; Multicultural Validity

Further Readings

Baumrind, D. (1971). Current patterns of parental authority. Developmental Psychology Monographs, 4(1, Pt. 2), 1–103.

Chao, R. K. (1994). Beyond parental control and authoritarian parenting style: Understanding Chinese parenting through the cultural notion of training. Child Development, 65, 1111–1119.

Steinberg, L., Lamborn, S. D., Darling, N., Mounts, N., & Dornbusch, S. (1994). Over-time changes in adjustment and competence among adolescents from authoritative, authoritarian, indulgent and neglectful families. Child Development, 65, 754–770.

Kyung (Chris) T. Han Kyung (Chris) T. Han Han, Kyung (Chris) T.

PARSCALE

PARSCALE

1207

1210

PARSCALE

PARSCALE is a software tool designed to analyze test response data based on item response theory (IRT) models. In the early 1990s, Eiji Muraki and R. Darrell Bock first developed the initial version of PARSCALE for the DOS platform. The latest version, PARSCALE 4.1 for the Microsoft Windows platform, is commercially available from Scientific Software International, Chicago, IL. There also is a modified version of PARSCALE for the UNIX platform, which is customized mainly for internal use at Educational Testing Service.

PARSCALE was one of the few software tools available in the 1990s and early 2000s for analyzing ordered polytomous response data, such as Likert-type rating scale responses and graded response data. The program supports several different IRT models for polytomous data including the graded response model (GRM), partial credit model (PCM), and generalized partial credit model (GPCM). It also supports dichotomously scored data based on one-or two-parameter logistic models as a special case of GRM or partial credit model or data based on a three-parameter logistic model. PARSCALE can handle a mixture of items with different numbers of response categories and IRT models.

In addition to estimating the item parameters and person scores, PARSCALE offers useful functionalities for multiple-group analysis, rater effect analysis, and differential item and functioning analysis.

IRT Models Supported in PARSCALE

GRM in PARSCALE is, in the logistic form, given by

$$P_{jk}(\theta) = \frac{\exp\left(Da_j\left(\theta - b_j + d_{jk}\right)\right)}{1+\exp\left(Da_j\left(\theta - b_j + d_{jk}\right)\right)} - \frac{\exp\left(Da_j\left(\theta - b_j + d_{jk+1}\right)\right)}{1+\exp\left(Da_j\left(\theta - b_j + d_{jk+1}\right)\right)},$$

where D is the scaling coefficient, which is either 1 (for logistic metric) or 1.7 (for the normal metric); a_j and b_j are a common slope parameter (i.e., item discrimination parameter) and a threshold parameter (i.e., item difficulty parameter) for item j, respectively; and d_{jk} is the score-category parameter for category k. Unlike PARSCALE, in Samejima’s original GRM work in 1969 and 1972, $–b_j + d_{jk}$ was simply expressed as $–b_{jk}$. PARSCALE also supports the normal-ogive version of GRM, which is

$$P_{jk}(\theta) = \frac{1}{\sqrt{2\pi}} \int_{a_j\left(\theta - b_j + d_{jk}+1\right)}^{a_j\left(\theta - b_j + d_{jk}\right)} \exp\left(\frac{-t^2}{2}\right) dt.$$

Another IRT model for polytomously scored response data supported by PARSCALE is the Masters partial credit model and its generalized version, GPCM, by Muraki. The GPCM is given by

$$p_{jk}(\theta) = \frac{\exp\left(\sum_{v=0}^{k} Da_j\left(\theta - b_{jv}\right)\right)}{\sum_{w=0}^{mj} \exp\left(\sum_{v=0}^{w} Da_j\left(\theta - b_{jv}\right)\right)},$$

where $k = 1, 2, \ldots, m_j$.

For dichotomously scored response data, PARSCALE supports the three-parameter logistic model, in which the lower asymptote of item characteristic function is addressed by *c*-parameter in

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + \exp(-Da_j(\theta - b_j))}.$$

Estimation Methods and Processes

PARSCALE divides the model estimation process into four “phases.” Phase 0 is for processing the syntax input and other data input files and calibration settings. Phase 1 is for summarizing the response data matrix (reporting frequencies of response categories for each item) and for calculating the mean and standard deviation of response for each item as well as the Pearson product-moment correlation and polyserial correlation coefficients, which are used for computing the initial values for *a*- and *b*-item parameters for each item.

Phase 2 is for calibrating item parameters. PARSCALE uses the marginal maximum-likelihood estimation (MMLE) method, which is an instance of the expectation–maximization (EM) algorithm. In the expectation (E) step of the EM algorithm with MMLE at the beginning, given the initial item parameter values from Phase 1, the expected posterior distribution is computed based on marginalized θ distribution, which is from an integral of probabilities of different response patterns across quadrature nodes with preset weights. In PARSCALE, the integral is approximated using the Gauss–Hermite quadrature method. In the maximization (M) step of Phase 2, using the expected values from the E step, item parameter values that maximize the log-likelihood function are searched using the Newton–Gauss (i.e., Fisher scoring) iterative procedure. By default, PARSCALE assumes a log-normal prior distribution for *a* parameters (i.e., slope parameter), a normal prior distribution for *b* parameters (i.e., threshold parameter), and a β prior distribution for *c* parameters (i.e., lower asymptote parameter). Once the M step search finds the item parameter estimates, the E and M steps are repeated until either the number of iterations reaches a maximum level or the convergence criterion is met. PARSCALE allows users extensive controls over the MMLE setting, including the number and weight of quadrature points, the maximum number of EM iterations, the maximum number of Newton–Gauss iterations, and conversion criteria for the EM and Newton–Gauss procedures.

The Phase-3 model estimation process of PARSCALE is used for estimating θ parameters (i.e., person score) given the item parameters calibrated from Phase 2. PARSCALE offers three different ways to estimate θ: (1) maximum-likelihood estimation, (2) Warm’s weighted maximum likelihood (WML) estimation, and (Bayes) expected a posteriori (EAP) estimation. For MLE and WML estimation, the Newton–Raphson iterative procedure is used to find θ that maximizes

$$\text{In}\, L\left(U_{jk} \mid \theta\right) = \text{In} f\left(\theta\right) + \text{In} \prod_{j=1}^{m} \prod_{k=0}^{m_j} \left[P_{jk}\left(\theta\right)\right]^{U_{jk}},$$

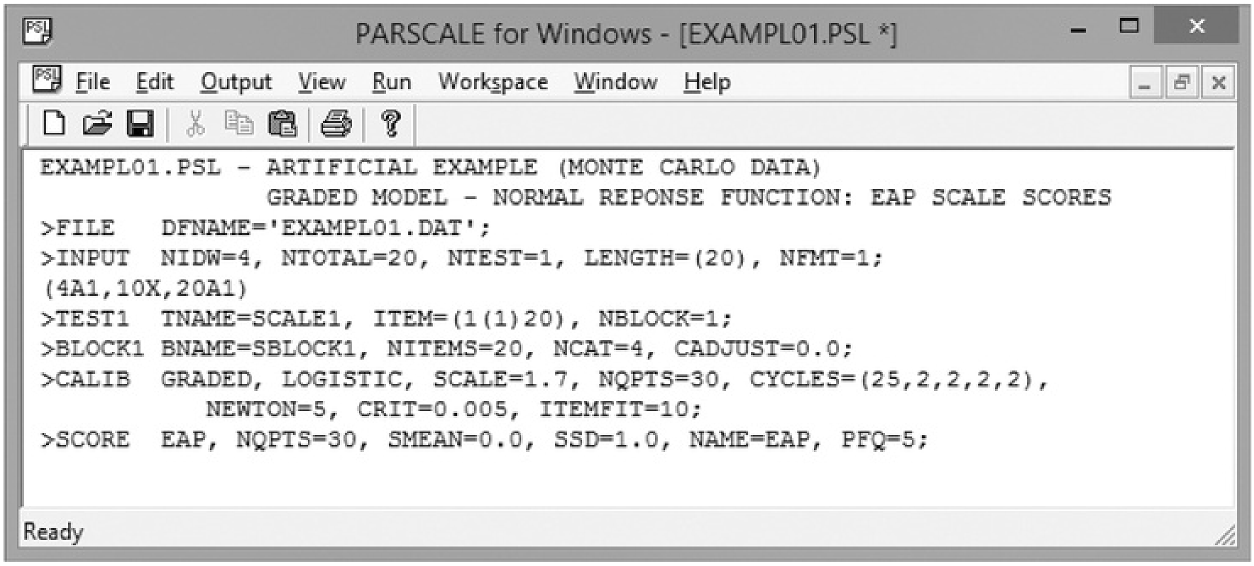
where $f(\theta)$ is 1 when MLE is used, and it is the square root of the test information function when WML estimation is used. For (Bayes) EAP, the mean of the posterior distribution of θ is computed, given the observed responses and a prior distribution. For the prior distribution of EAP, PARSCALE supports (a) uniform distribution, (b) normal on equidistance points, and (c) normal on Gauss–Hermite points. Users also can specify the weights across quadrature points.

User Interface and Syntax

Although the latest commercial version of PARSCALE, Version 4.1, runs on Microsoft Windows operating systems, its input and output user interface is mostly text based (except for the plotting features). The core computational programs for processing Phases 0 through 3 are console applications (PSL0.EXE, PSL1.EXE, PSL2.EXE, and PSL3.EXE for Phases 0, 1, 2, and 3, respectively). Because results from earlier phases are required to run later phases, users should execute each phase sequentially. All phases use a single command syntax file.

A command syntax must have the following command lines: TITLE, FILES, INPUT, TEST, BLOCK, CALIB, and SCORE. All other commands are optional. As shown in Figure 1, in the example syntax file (that comes with the PARSCALE 4.1), the first two lines of the syntax are for the title. All other commands should start with the “>” character, so that PARSCALE can recognize each command line.

Figure 1 Example of PARSCALE syntax file, Adapted from EXAMPLE01.PSL included in PARSCALE 4.1


PARSCALE for Windows - [EXAMPL01.PSL *]
File Edit Output View Run Workspace Window Help
Ready

```
EXAMPL01.PSL - ARTIFICIAL EXAMPLE (MONTE CARLO DATA)
               GRADED MODEL - NORMAL REPONSE FUNCTION: EAP SCALE SCORES
>FILE   DFNAME='EXAMPL01.DAT';
>INPUT  NIDW=4, NTOTAL=20, NTEST=1, LENGTH=(20), NFMT=1;
(4A1,10X,20A1)
>TEST1  TNAME=SCALE1, ITEM=(1(1)20), NBLOCK=1;
>BLOCK1 BNAME=SBLOCK1, NITEMS=20, NCAT=4, CADJUST=0.0;
>CALIB  GRADED, LOGISTIC, SCALE=1.7, NQPTS=30, CYCLES=(25,2,2,2,2),
           NEWTON=5, CRIT=0.005, ITEMFIT=10;
>SCORE  EAP, NQPTS=30, SMEAN=0.0, SSD=1.0, NAME=EAP, PFQ=5;
```

With the FILE command, users must specify the response data file with the "DFNAME" key word. As an option, users can specify data files for existing item parameter values and codes for omitted or not-presented items in the FILE command. The INPUT command is where users can define the specifications of response data. It is important to note that the format of the data file in FORTRAN programming language must immediately follow the INPUT command. In the example shown in Figure 1, "(4A1, 10X, 20A1)" indicates that PARSCALE expects a string having four characters, 10 placeholders, and 20 characters. According to the number of subtests specified in the INPUT command with the key word "NTEST," PARSCALE expects the same number of TEST command lines. In the example shown in Figure 1, NTEST equals 1, so there is only one TEST command line (">TEST1"). In each TEST command line, users should supply the number of test blocks within the subtest. Each TEST command line should be followed immediately by as many BLOCK command lines as specified with the "NBLOCK" key word. In each BLOCK command line, users can enter the number of response categories, recoding definition of response categories, and estimation options for the lower asymptote (*c* parameter) and slope parameter. In Figure 1, "NITEMS = 20, NCAT = 4" indicates that the block expects 20 items with polytomous responses with four categories.

Users can supply the input needed for item calibration in the CALIB command. A user must input either the "GRADE" key word to use GRM or the "PARTIAL" key word to use GPCM. For GRM, the normal-ogive metric is available with the "NORMAL" key word. In Figure 1, the user has chosen the logistic metric with the “LOGISTIC” key word. Users can specify the details of MMLE/EM iterative procedure, for example, the maximum number of iterations and conversion criteria in the CALIB command line.

Finally, users select the θ estimation method among MLE, WML, and EAP methods in the SCORE command line. In PARSCALE, EAP is the default method.

Outputs

Each phase produces an output file. The output file from Phase 0 has a file name with the extension “PH0” and reports recognized syntax commands, quadrature points with prior weights, and information from the response data file. The output from Phase 1 contains the frequency table and summary statistics of responses for each item and initial *a*- and *b*-parameter values for item calibration. The output from Phase 2 contains the *a*-, *b*-, and *c*-item parameter estimates with their standard errors. If requested in the syntax, the item-fit statistics are also provided in the Phase 2 output. The Phase 3 output file contains θ estimates (i.e., person score) with standard error of estimation.

Once users complete Phases 0–3, they can use the plotting feature for drawing item characteristic curves, item information functions, and/or test information functions.

Kyung (Chris) T. Han

See also Item Response Theory

Further Readings

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of EM algorithm. Psychometrika, 46, 443–459; *47*, 369 (Errata).

du Toit, M. (Ed.). (2003). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Chicago, IL: Scientific Software International.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159–176.

Muraki, E., & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data. Chicago, IL: Scientific Software International.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores (Psychometrika Monograph Supplement, No. 17). Richmond, VA: Byrd Press.

Stroud, A. H., & Secrest, D. (1966). Gaussian quadrature formulas. Englewood Cliffs, NJ: Prentice Hall.

A. Alexander Beaujean A. Alexander Beaujean Beaujean, A. Alexander

Part Correlations

Part correlations

1210

1212

Part Correlations

A part correlation is a measure of linear association between one variable and another after the variability in at least one other variable is removed from only one of the original variables. The traditional formula for the part correlation between, say, variables *X* and *Y* after controlling for *Z* in *Y*, denoted $r_{X(Y \cdot Z)}$, is

$$r_{X(Y \cdot Z)} = \frac{r_{XY} - r_{XZ} \times r_{YZ}}{\sqrt{1 - r_{YZ}^2}},$$

where *r* is the Pearson correlation between two variables.

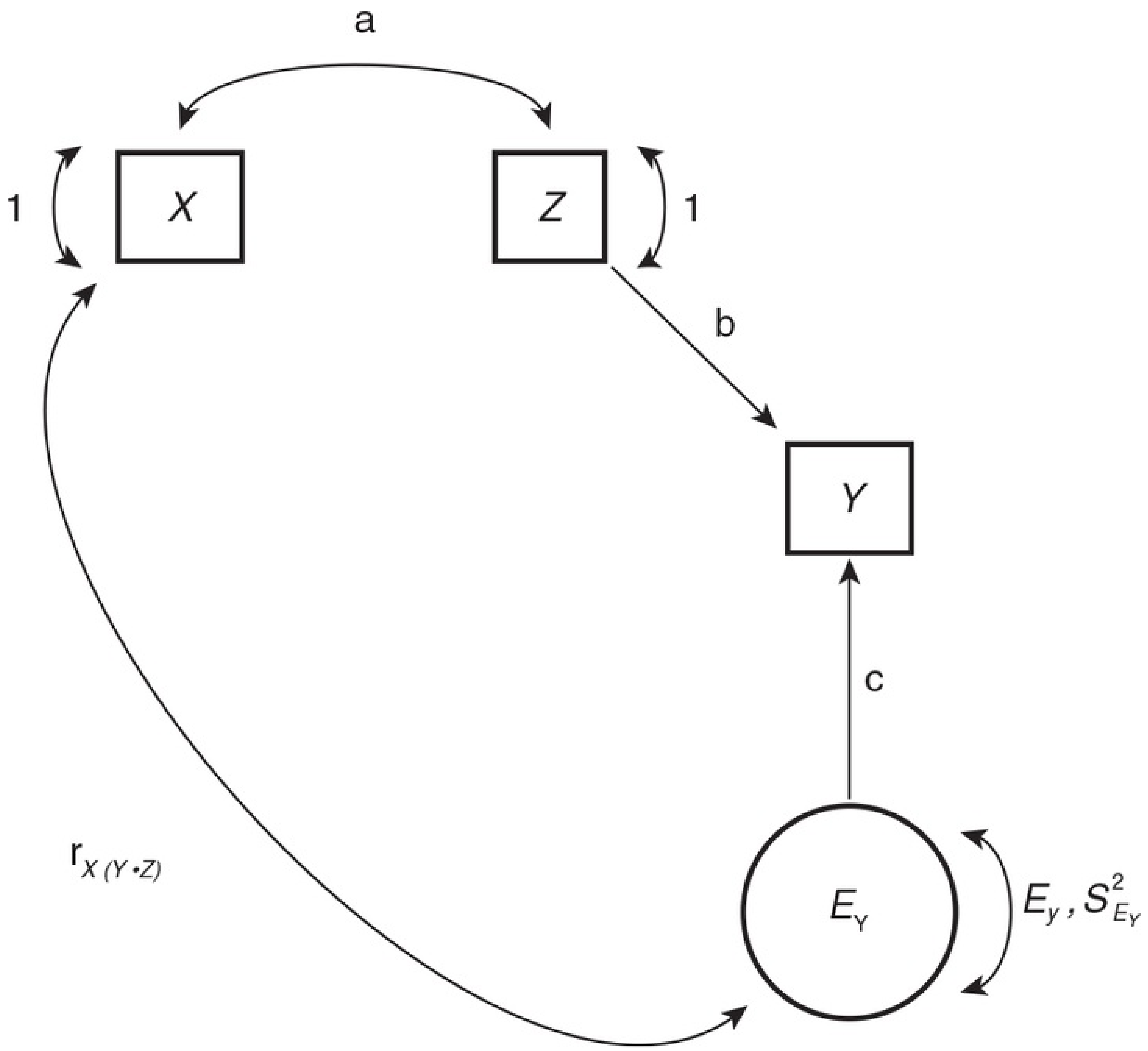
$r_{X(Y \cdot Z)}$ is sometimes referred to as a *first-order* part correlation, to note that the correlation only controls for one other variable. If controlling for two variables, it would be a *second-order*, and so on. Zero-order correlations do not control for any other variables.

While not obvious from Equation 1, the part correlation is really the correlation between *X* and the residual of *Y* after regressing *Y* on *Z*. A path model representing this interpretation is shown in Figure 1, which assumes *X*, *Y*, and *Z* are standardized variables.

In Figure 1, the zero-order correlation between *X* and *Y*, r_{XY}, can be found using the tracing rules for path diagrams. It is

$$r_{XY} = a \times b + c \times s_{E_Y}^2 \times r_{XY \cdot Z}.$$

Figure 1 Path diagram of part correlation. Variables *X*, *Y*, and *Z* are standardized



The values for paths *a* and *b* are the zero-order correlations of *X* with *Z* and *Y* with *Z*, respectively. To make the model identified (i.e., be able to find unique estimates for the paths), either *c* or the variance of E_y, , needs to be constrained. To make $r_{X(Y \cdot Z)}$ a correlation (as opposed to a covariance), constrain the variance of E_y to one (i.e., standardize the residual). This makes path *c* equal the square root of the variance in *Y* not explained by *Z* (i.e.,).

After substituting terms, Equation 2 now becomes

$$r_{XY} = r_{XZ} \times r_{YZ} + \sqrt{1 - r_{YZ}^2} \times r_{X(Y \cdot Z)}.$$

Equation 3 can be rearranged as

$$r_{X(Y \cdot Z)} \times \sqrt{1 - r_{YZ}^2} = r_{XY} - r_{XZ} \times r_{YZ}.$$

Dividing both sides of this rearrangement by produces Equation 1.

Part Correlation Extensions

As with zero-order correlations, part correlations can be squared to give the amount of variance in common between two variables. indicates the increment in *Y*’s variance explained by *X* beyond that explained by *Z*. It is analogous to the eta squared (η^2) effect size often used in analysis of variance models.

If removing the variance of more than one variable (e.g., second order, third order), then Equation 1 can be generalized to

$$r_{X(Y \cdot A)} = \sqrt{R^2_{X|YA} - R^2_{X|A}}\,,$$

where ***A*** is set of variables, and is the squared multiple correlation from a regression analysis with *X* being the outcome and *Y* and ***A*** being the predictors (likewise for , except only ***A*** are the regression predictors).

There are different ways to calculate the standard error of the part correlation. One method is to use Fisher’s *z* transformation, as is often done with zero-order correlations. Another method is to apply the standard error formula for standardized regression coefficients:

$$\sqrt{\frac{1-r_p^2}{n-k-1'}},$$

where *n* is the sample size, *k* is the number of variables in the correlation minus 1, and is the squared part correlation.

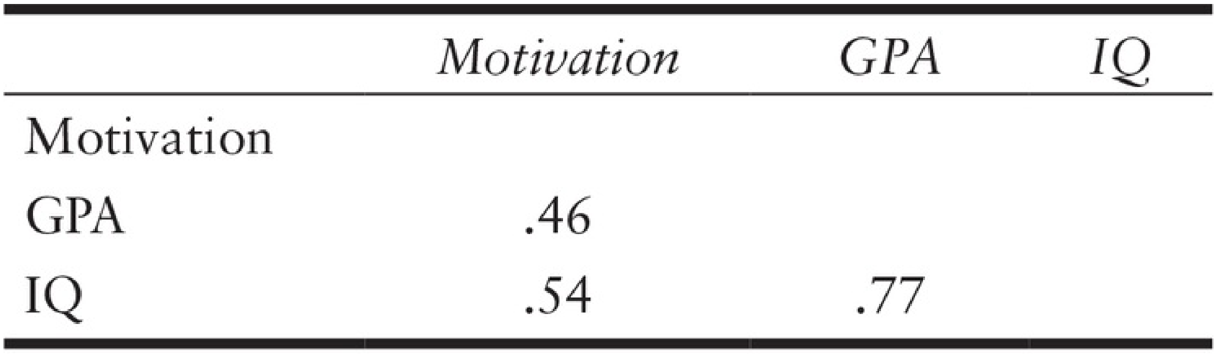
Example

Some didactic data are provided in Table 1. The variables are IQ, academic motivation, and GPA for six students. All variables were scaled to have a population mean of 10 and standard deviation of 3. The zero-order correlations are in Table 2. For example, the motivation-GPA correlation is .46. To calculate the motivation-GPA correlation after controlling for IQ in GPA only, plug the appropriate values into Equation 1:

$$r_{\text{Motiv}(\text{GPA}\cdot\text{IQ})} = \frac{r_{\text{Motiv,GPA}} - r_{\text{Motiv,IQ}} \times r_{\text{GPA,IQ}}}{\sqrt{1 - r^2_{GPA,IQ}}}$$

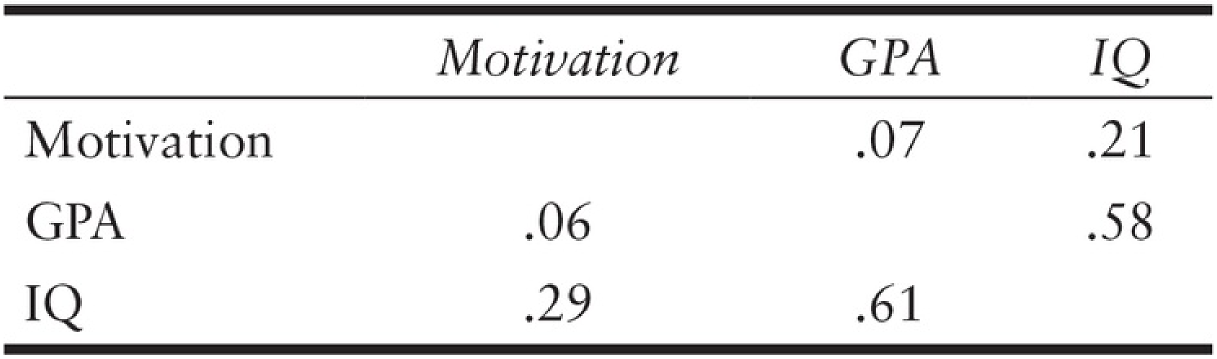
$$= \frac{.46 - .54 \times .77}{\sqrt{1 - .77^2}} = \frac{.05}{.64} = .07.$$

Student	*Motivation*	*GPA*	*IQ*
1	13	11	13
2	16	10	10
3	7	12	11
4	5	7	5
5	10	10	15
6	14	12	15



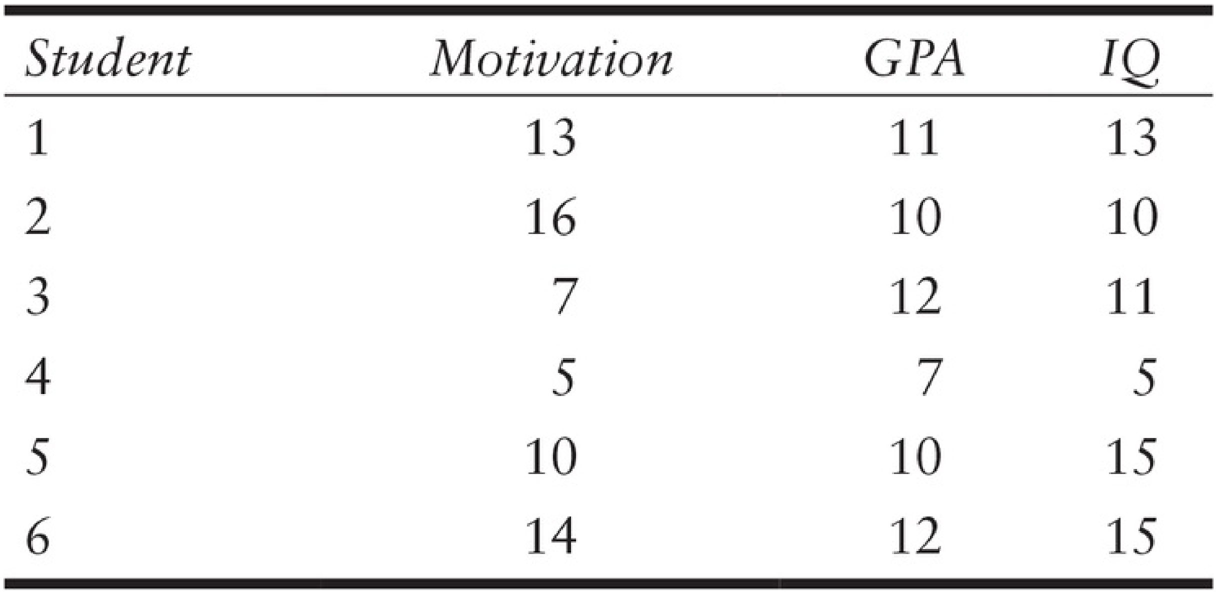
	Motivation	*GPA*	*IQ*
Motivation			
GPA	.46		
IQ	.54	.77	

Table 3 Part Correlations of Example Data



	Motivation	*GPA*	*IQ*
Motivation		.07	.21
GPA	.06		.58
IQ	.29	.61	

The correlation between motivation and GPA, after controlling for IQ in GPA, is .07. The other part correlations for the data are given in Table 3.



Student	*Motivation*	*GPA*	*IQ*
1	13	11	13
2	16	10	10
3	7	12	11
4	5	7	5
5	10	10	15
6	14	12	15

Note: Coefficients have the column variable partialed.

Using Equation 5, the standard error for the part correlation between motivation and GPA after controlling for IQ in GPA is

$$\sqrt{\frac{1-r_p^2}{n-k-1}}=\sqrt{\frac{1-.07^2}{6-2-1}}=.58.$$

Relation to Other Statistical Methods

Part correlations are closely related to standardized coefficients in multiple regression. Moreover, as can be seen by Equation 4, the squared part correlation is equivalent to the change in the R^2 after removing the partial variable from the regression. For the didactic data, the . Consequently, the difference in R^2 values from a regression model predicting motivation from IQ and GPA and a regression model predicting motivation from IQ is .006.

A. Alexander Beaujean

See also Multiple Linear Regression; Partial Correlations; Path Analysis; Pearson Correlation Coefficient

Further Readings

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Loehlin, J. C., & Beaujean, A. A. (2017). Latent variable models: An introduction to factor, path, and structural equation analysis (5th ed.). New York, NY: Routledge.

Preacher, K. J. (2006). Testing complex correlational hypotheses with structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 13, 520–543. doi:10.1207/s15328007sem1304_2

A. Alexander Beaujean A. Alexander Beaujean Beaujean, A. Alexander

Partial Correlations Partial correlations

1212

1214

Partial Correlations

A partial correlation is a measure of linear association between two variables after variability in at least one other variable is removed from both variables. The traditional formula for the partial correlation between, say, variables *X* and *Y* after controlling for *Z*, denoted $r_{XY \cdot Z}$, is

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} \times r_{YZ}}{\sqrt{1 - r_{YZ}^2} \times \sqrt{1 - r_{YZ}^2}},$$

where *r* is the Pearson correlation between two variables.

$r_{X(Y \cdot Z)}$ is sometimes referred to as a *first-order* partial correlation to note that the correlation only controls for one other variable. If controlling for two variables, it would be a *second-order*, and so on. Zero-order correlations do not control for any other variables.

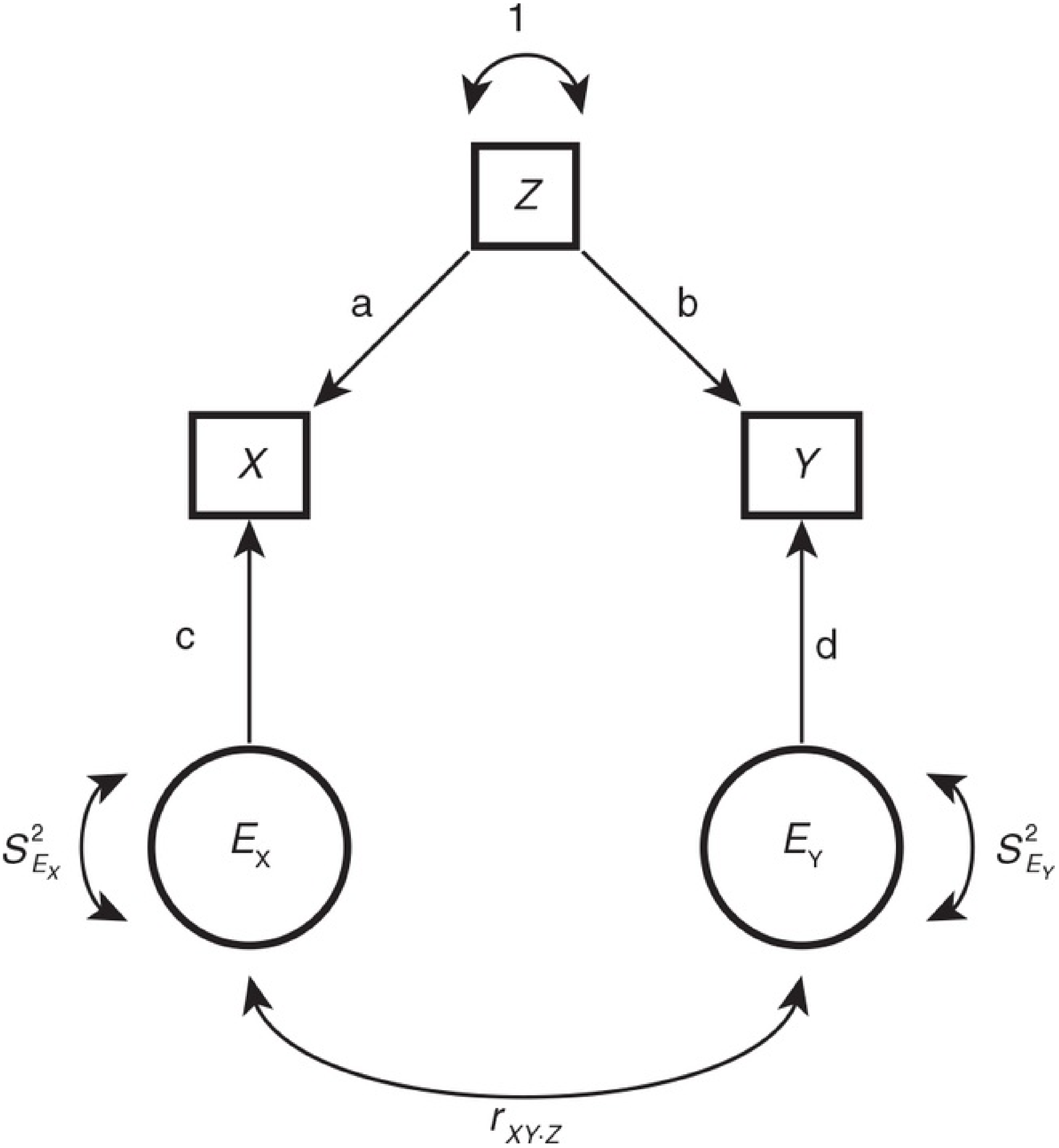
While not obvious from Equation 1, the partial correlation is really the correlation between the residuals of *X* and *Y* after regressing both variables on *Z*. A path model representing this interpretation is shown in Figure 1, which assumes *X*, *Y*, and *Z* are standardized variables.

In Figure 1, the zero-order correlation between *X* and *Y*, r_{XY}, can be found using the tracing rules for path diagrams. It is

$$r_{XY} = a \times b + c \times s_{E_X}^2 \times r_{XY \cdot Z} \times s_{E_Y}^2 \times d.$$

Figure 1 Path diagram of partial correlation. Variables *X*, *Y*, and *Z* are

standardized



The values for paths *a* and *b* are the zero-order correlations of *X* with *Z* and *Y* with *Z*, respectively. To make the model identified (i.e., be able to find unique estimates for the paths), either *c* and *d* or the variances of E_x and E_y, and , need to be constrained. To make $r_{XY \cdot Z}$ a correlation (as opposed to a covariance), constrain the variances of E_x and E_y to one (i.e., standardize these residuals). This makes the paths *c* and *d* equal the square root of the variance in *X* and *Y* not explained by Z (i.e., and .

After substituting terms, Equation 2 now becomes

$$r_{XY} = r_{XZ} \times r_{YZ} + \sqrt{1 - r_{YZ}^2} \times r_{XY \cdot Z} \times \sqrt{1 - r_{YZ}^2}.$$

Equation 3 can be rearranged as

$$r_{XY \cdot Z} \times \sqrt{1 - r_{XZ}^2} \times \sqrt{1 - r_{YZ}^2} = r_{XY} - r_{XZ} \times r_{YZ}.$$

Dividing both sides of this rearrangement by produces Equation 1.

Partial Correlation Extensions

As with zero-order correlations, partial correlations can be squared to give the amount of variance in common between two variables. indicates the variance in common between *X* and *Y* after removing what they both share with Z. It is analogous to the partial eta squared effect size often used in analysis of variance models.

If removing the variance of more than one variable (e.g., second order, third order), then Equation 1 can be generalized to

$$r_{XY \cdot A} = \sqrt{\frac{R_{X|YA}^2 - R_{X|A}^2}{1 - R_{X|A}^2}} = \sqrt{\frac{R_{Y|XA}^2 - R_{Y|A}^2}{1 - R_{Y|A}^2}},$$

where ***A*** is set of variables, is the squared multiple correlation from a regression analysis with *X* being the outcome and *Y* and ***A*** being the predictors, and squared multiple correlation with only ***A*** as the predictors (likewise for and except *Y* is the outcome variable).

There are different ways to calculate the standard error of the partial correlation. One method is to use Fisher's *z* transformation, as is often done with zero-order correlations. Another method is to apply the standard error formula for standardized regression coefficients:

$$\sqrt{\frac{1-r_p^2}{n-k-1'}},$$

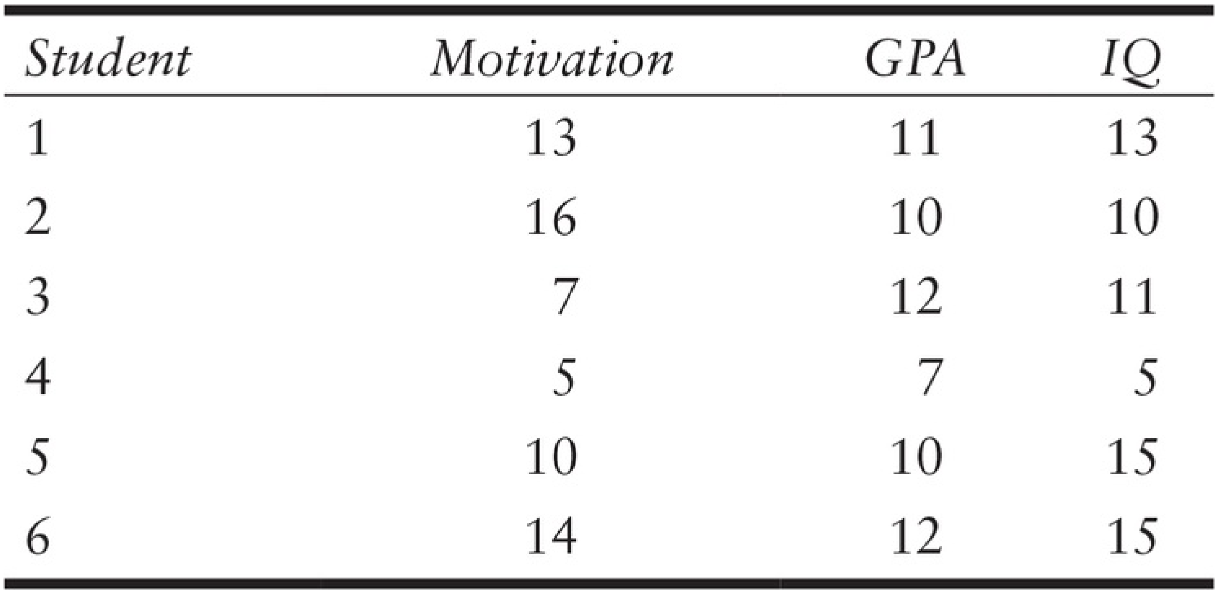
where *n* is the sample size, *k* is the number of variables in the correlation minus 1, and is the squared partial correlation.

Example

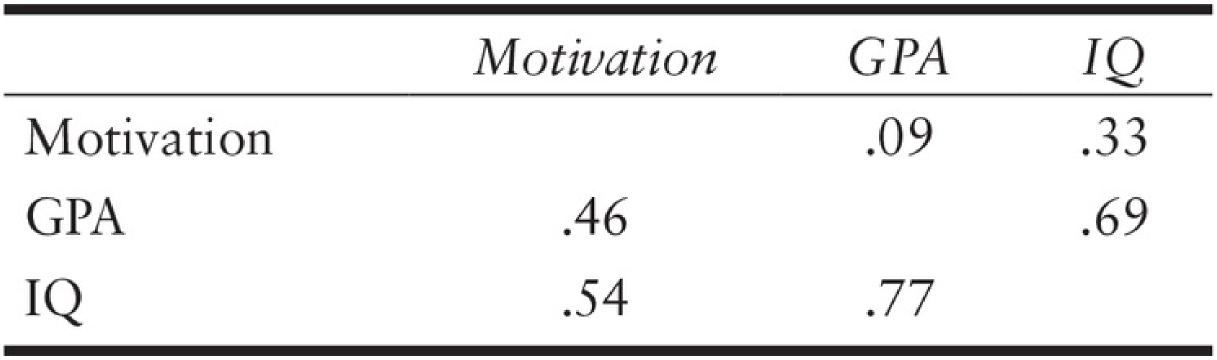
Some didactic data are provided in Table 1. The variables are IQ, academic motivation, and GPA for six students. All variables were scaled to have a population mean of 10 and standard deviation of 3. The zero-order correlations are below the diagonal in Table 2. For example, the motivation-GPA correlation is .46. To calculate the motivation-GPA correlation after controlling for IQ, plug the appropriate values into Equation 1:

$$r_{\text{Motiv,GPA}\cdot\text{IQ}} = \frac{r_{\text{Motiv,GPA}} - r_{\text{Motiv,IQ}} \times r_{\text{GPA,IQ}}}{\sqrt{1-r^2_{Motiv,IQ}} \times \sqrt{1-r^2_{GPA,IQ}}}$$

$$= \frac{.46-.54\times.77}{\sqrt{1-.54^2}\times\sqrt{1-.77^2}} = \frac{.05}{.84\times.64} = .09.$$



Student	*Motivation*	*GPA*	*IQ*
1	13	11	13
2	16	10	10
3	7	12	11
4	5	7	5
5	10	10	15
6	14	12	15



	Motivation	*GPA*	*IQ*
Motivation		.09	.33
GPA	.46		.69
IQ	.54	.77	

Note: Zero-order correlations below the diagonal, partial correlations above diagonal.

The correlation between motivation and GPA, after controlling for IQ in both variables, is .09. The other partial correlations for the data are given above the diagonal in Table 2.

Using Equation 5, the standard error for the partial correlation between motivation and GPA after controlling for IQ is

$$\sqrt{\frac{1-r_p^2}{n-k-1}}=\sqrt{\frac{1-.09^2}{6-2-1}}=.58.$$

Relation to Other Statistical Methods

Partial correlations are foundational to other statistical techniques, such as path analysis, structural equation models, and exploratory factor analysis. They are also used in the traditional method of examining mediation—that is, without the presence of a third variable, the correlation between two other variables should disappear (i.e., $r_{XY \cdot Z} = 0$), or at least be substantially attenuated (i.e., $r_{XY \cdot Z} \ll r_{XY}$).

A. Alexander Beaujean

See also Mediation Analysis; Multiple Linear Regression; Part Correlations; Path Analysis; Pearson Correlation Coefficient

Further Readings

Blalock, H. M., Jr. (1963). Making causal inferences for unmeasured variables from correlations among indicators. American Journal of Sociology, 69,

53–62.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Loehlin, J. C., & Beaujean, A. A. (2017). Latent variable models: An introduction to factor, path, and structural equation analysis (5th ed.). New York, NY: Routledge.

Preacher, K. J. (2006). Testing complex correlational hypotheses with structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 13, 520–543. doi:10.1207/s15328007sem1304_2

Satoko Siegel Satoko Siegel Siegel, Satoko

Participant Observation Participant observation

1214

1216

Participant Observation

Participant observation is a qualitative data collection methodology that provides rich descriptive information on human behaviors and experiences in a particular context. This approach enables a researcher to participate in a social group and observe people as well as the environment. In doing so, the researcher develops a holistic understanding of how people make sense of their experiences and what is occurring around them. Participant observation is used in disciplines such as anthropology, sociology, and education.

This entry describes the process of participant observation, focusing on (a) the researcher’s level of involvement, (b) reflexivity of the participant observer, (c) types of observation conducted, (d) observation record, and (e) fieldwork analysis. The entry concludes with an overview of the advantages and disadvantages of the participant observation method.

The Researcher’s Level of Involvement

During the participant observation process, the researcher is the primary data collection tool. As a participant observer, the researcher goes into the field of study to observe people, events, and social contexts. A researcher’s involvement can be divided into five levels: (1) nonparticipation, (2) passive participation, (3) moderate participation, (4) active participation, and (5) complete participation. Nonparticipation refers to a situation in which a researcher observes activities from outside of the field (e.g., viewing activities captured on video). Passive participation is when researchers are present in a particular social context but not actively involved in the activities. They observe an event and take notes without being immersed in the situation. Because of the researchers’ position as an outsider to the social group, they get to know the people and the environment and learn how to act appropriately in the setting as an addition to the group dynamic. Moderate participation is when the researcher is between a passive outsider and an active member of the social group in the context. The researcher occasionally joins the social activities and events while also observing the phenomena and taking observation notes at the site. Active participation refers to situations in which a researcher is engaged in the activities of the social group. Although the researchers must still learn the culturally constructed rules for observed behavior of the group, they are considered engaged member. Complete participation occurs when the researcher obtains the full membership of the social group. The researcher and the group have an open relationship, and the people who are being observed sometimes forget that the observer is there to do research.

Reflexivity of the Participant Observer

Reflexivity refers to the researchers’ awareness of their role in the field of the study. More specifically, it is the idea that who the researcher is in terms of gender, ethnicity, age, and cultural and historical background will influence the process of the study. This is important because nearly every aspect of the data collection process (e.g., how to enter the field of the study, who becomes the informant of the study, how to perceive and interpret the phenomena, and how to interpret the data) will be influenced by the researcher’s reflexivity. In addition, reflexivity involves an awareness of the relationship between the researcher and the research participants. This relationship constantly affects the fieldwork and negotiates the power relationships in the field. For example, if there are informants who share their insider perspectives on a social event, the researchers need to be aware of the relationship between the informants and themselves, as well as the informant’s position within the social group.

Types of Observation

The approaches used while conducting participant observation typically change over time. There are three types of observation: (1) descriptive, (2) focused, and (3) selective. Descriptive observation is conducted at the beginning phase of the observation process. The researcher tries to obtain as much information in the field as possible in order to get a general sense of a naturally and socially occurring situation. After analyzing the initial observation, the researcher finds reoccurring events and phenomena, or *themes*. Focused observation occurs when a researcher, after conducting a descriptive observation and deciding on several themes, concentrates on specific subject areas. After conducting more observations in the field and analyzing the focused themes further, the researcher is able to narrow the inquiry further and observe selectively. Selective observation is highly focused and involves the researcher investigating and trying to make sense of what is being observed.

Observation Record

The observation record is an essential bridge between the observation and the analysis of a studied field. Although observation records are usually documented by field notes, other means, such as taking photographs, making maps, and filming, can be used as additional recording tools.

Field notes should be recorded during and/or after each observation. It may or may not be easy for the researcher to take fields notes while being a participant observer. When the researchers conduct passive observation to gain general ideas of the studied field, they take intensive field notes. In the case of moderate, active, and complete observations, the researchers take notes in the field whenever possible. If note-taking is impossible, the researchers document events from memory later.

Field notes include descriptions of the activities, interactions, conversations, and contexts of the field. In addition, the observer’s reflections and reactions are documented. Field notes are entered typically on the same day or no later than 48 hours after the observation while the memory is fresh.

Fieldwork Analysis

Participant observation methodology uses recursive ongoing analysis, meaning the researcher develops thematic categories and organizes the data based on recurring incidents, people’s behaviors, and conversation topics. Once the themes are determined, the researcher looks for further insights and perspectives, focusing on the themes generated.

Advantages and Disadvantages

The advantages of participant observation methodology are as follows. First, this methodology is beneficial for holistically exploring the everyday lives of people. It allows the researcher to provide an in-depth understanding and rich description of a situation from a viewpoint of a person within the culture. Second, for sensitive research topics such as drugs and gambling, participant observation can uncover the meanings of situations that participants are unwilling or unable to talk about. Third, participant observation enables the researcher to check whether people do what they believe they do. Finally, this methodology allows the researcher to build direct connections and trust with the group. The rapport that is established increases the likelihood that the participants act naturally around the researcher, so in-depth and authentic information can be gained.

There are some disadvantages to participant observation methodology. First, it is time-consuming. The researcher spends time with the people of the social group, taking observation notes, analyzing the observation, developing the themes, and going back into the field. Although the period of the fieldwork can range from a few hours to a few years, depending on the types of study and focus, each process of participant observation takes time. Second, because the researchers are the primary data collection instruments, their experiences and biases may influence the results. The data might lose objectivity, especially when the researcher is personally involved with the group. On the other hand, if the researchers’ level of involvement is minimal, their interpretation and analysis might not be authentic enough to the social group’s perspectives. As solutions of this limitation, researchers need to state clearly their reflexivity as a part of methodology and conduct “member checks” with the studied group to increase the validity of the study. Third, due to the nature of qualitative research, this method lacks generalizability. In other words, results from participant observation studies may not be applicable to other events. Lastly, there are some ethical issues that a study utilizing participant observation methodology may face. When the study’s focus involves illegal or immoral acts and crimes, protecting participants’ confidentiality and ensuring that they do not experience any negative effects because of participating in the study are the researcher’s responsibilities.

Satoko Siegel

See also Field Notes; Qualitative Data Analysis; Qualitative Research Methods

Further Readings

Ager, M. (1980). The professional stranger: An informal introduction to ethnography. New York, NY: Academic Press.

DeWalt, K. M., & DeWalt, B. R. (2011). Participant observation. A guide for fieldworkers. Lanham, MD: AltaMira Press.

Geertz, C. (1984). From the native’s point of view: On the nature of anthropological understanding. In R. A. Sheweder & R. LeVine (Eds.), Culture theory: Essays on mind, self, and emotion (pp. 123–136). New York, NY: Cambridge University Press.

Spradley, J. P. (1980). Participant observation. New York, NY: Holt, Rinehart and Winston.

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

B. Venkatesh Kumar B. Venkatesh Kumar Kumar, B. Venkatesh

Participatory Evaluation

Participatory evaluation

1216

1220

Participatory Evaluation

A participatory approach to evaluation analyzes contextualized data that answer questions about the successes, contextualizing realities, and failures of a program or policy and examines ways to make the program or policy work better through the active involvement of representative and relevant local stakeholder groups. In participatory evaluation, multiple stakeholders actively coconstruct and participate in the research and evaluation process through its various stages—from research question formation and research design through to data collection, data analysis, and the sharing and dissemination of findings.

Participatory evaluation is developed through locally constructed approaches to democratic and collaborative decision making and engagement, with a diverse range of stakeholders contributing to shared processes, learnings, and outcomes. This approach to evaluation is built upon the fundamental ethic of equity-oriented inclusion and equal participation and voice of all stakeholders. This entry first discusses the guiding principles of participatory evaluation, how it differs from conventional evaluation, and the elements of a participatory evaluation framework. It then provides a template for the methodology and process flow of participatory evaluations and describes the range of data collection methods that can be used. Finally, the entry describes the benefits and challenges of participatory evaluation.

Guiding Principles

The guiding principles of participatory evaluation include

- collective problem, need, and resource identification and shared decision making;
- equity and inclusivity, especially of the most vulnerable and marginalized groups, in the design, implementation, and analysis of data;
- democratic collaboration and engagement that create the conditions for people to be agentic and empower themselves;
- transparency and accountability in decision-making and evaluation processes based on transparent criteria;
- multiple stakeholder consultation to develop the process and then guide all choices;
- coconstruction of knowledge and learning that positions stakeholders as sharers of knowledge and active teachers and learners;
- responsive design and process adaptability that allow for methodologies to align with the context, resources present, and skills of the research team participants;
- diversity-oriented process that engenders engagement with a diverse group of stakeholders and participants;
- nonimpositional, bottom-up approach that actively resists power asymmetry and hegemony;
- people-centered approach to evaluation that focuses on participants' lived realities, experiences, resources, and needs;
- local knowledge construction that attends to the range and variation of “locals,” that is, the array of diverse individuals and groups that exist in any context and, further, that does not essentialize members of communities or organizations; and
- focus on the development of effective communication pathways that include the development of critical dialogue and conflict resolution skills.

Differences From Conventional Evaluation

Participatory evaluation is a specific form of and approach to evaluation that differs from conventional evaluation approaches in significant ways, including that participatory evaluation

- views stakeholders and participants as experts of their own experiences, collaborators, partners, and cocreators of knowledge and understanding (rather than as passive recipients or learners);

creates the conditions in which stakeholders are encouraged and supported to participate throughout each stage of the evaluation process;

helps create the conditions for participants to build ownership and contribute to the sustainability of the project even after the funders or development partners have exited;

creates the conditions in which participants can act as watchdogs, thereby creating more accountability for effectiveness and assessment;

creates the conditions in which communities and stakeholders who are vulnerable, marginalized, and/or underserved become active, agentic participants;

is a process through which stakeholder–researchers can build the skills and capacities they identify as necessary for their own development;

provides a fuller, more contextualized understanding of any system, organization, policy, or program through formally engaging and triangulating multiple perspectives; and

builds social capital and strengthens the initiative, policy, or program in focus through bonding and bridging a range of stakeholder groups.

Participatory evaluation requires collective identification of core issues, evaluation constructs, scope and terms of the evaluation, as well as criteria for assessment. These elements result in the development of clear and contextually relevant indicators, which are crucial to assess the performance of a program or policy. The indicators, which are what the evaluation is measured against, are developed through a collaborative consultative process in which the participants, representing multiple stakeholder groups such as beneficiaries (e.g., students and/or teachers), implementers (e.g., teachers and/or school leaders), and partners (e.g., funders and/or national-level policy leaders) identify needs, goals, objectives, outputs, and outcomes according to specific, well-articulated criteria. Participatory evaluation must be rigorous, informed by the host communities or institutions, as well as by the guiding goals and objectives of the programs or policies in focus. This helps to create a textured, robust, and well-informed framework and theory of action that can stand the test of time and resist external pressures or influences that might alter or interrupt contextually relevant learnings and authentic evaluation/program outcomes. Thus, a layer of accountability is built into participatory evaluation because many people who have a real stake are involved from the beginning.

Developing an Evaluation Framework

At the start of participatory evaluation, an evaluation framework is developed that positions the researchers as collaborators in the development, design, and implementation of the evaluation broadly and the evaluation indicators specifically.

For any program or policy to be implemented and evaluated, there must be a structured organizational mechanism that supports effective implementation while representing all relevant constituencies. It is important to develop this institutional/community architecture through a bottom-up, consultative process that involves those individuals and groups who have (or can build) requisite knowledge, competence, and dispositions for the institution/community to perform an evaluation to its optimal level.

The systems and processes of participatory evaluation include the logistical, pragmatic, human capital, and financial resources necessary to design and conduct an evaluation. This includes the use of needs–resource assessments, implementation of findings to drive programming, and the human capital and financial resources involved in the policy or program. To develop an evaluation framework, evaluators must consider how the project was implemented including the challenges, realities, and successes of implementation. As well, the team has to consider impact in terms of who, how many, and in what ways people were impacted by the policy or program. There must be a focus on emergent challenges and addressing those challenges.

Evaluators must be able to show that the program or policy has specific, articulated goals and that there are tangible, measureable results and specific outcomes that map onto those goals. Most often these outputs can be measured quantitatively, but qualitative exploration of the mediating factors is vitally important. Outcomes are comprised of medium-to long-term changes, often qualitative in nature, that can serve to foment systemic and macrolevel change. Outcome evaluations assess the effectiveness of a program or policy in producing desired change. They often focus on difficult questions about what happened to and for program participants and the nature and extent of change or improvement that the program made for them relative to program objectives.

A results framework broadly reflects the stated goals and objectives of the program or policy in focus and develops an achievable target with a definitive timeline to assess a program’s performance against these targets. Finally, dissemination and knowledge sharing are vital to the effectiveness and sustainability of evaluation findings and related recommendations for change sustainability of evaluation findings and related recommendations for change. The learnings help in the formulation and implementation of future programs and policies.

Participatory Evaluation Methodology and Process Flow

The following is a broad template for the methodology and process flow of participatory evaluations. Clearly, context will mediate these steps, and they are often iterative rather than linear.

A participatory approach is used to create and orient a locally based research evaluation team (i.e., local stakeholders). This team should ideally have select members who have appropriate qualifications that include experience conducting conventional and participatory evaluation research. Once the team is formed, members are sensitized to the values of the approach and oriented for the specific contextualized process.

A needs–resource assessment, typically garnered through interviews and focus groups, serves as a necessary data-based foundation for an evaluation. Both needs and existing human resources are assessed, so that baseline data are not deficit oriented (i.e., focused only on needs). This often includes collection and review of archival data.

Ideally, the evaluation team and key stakeholders engage in deep processes of dialogic engagement through which they identify the foci and construct guiding evaluation/research questions together. Intentional and diverse stakeholder participation is vital as the group frames the guiding evaluation questions and then maps the methods of the evaluation onto these questions. This includes creating a detailed timeline. The team and other local stakeholders collaboratively decide upon the appropriate methodological approach and related data collection methods and tools.

The data collection tools, or instruments, are pilot tested through engaging a sample of the population in pilot interviews, focus groups, and so on. Analysis of these data is shared within the team, so that learnings can be formatively implemented into the evaluation. The research team works together to plan for and engage in rigorous data collection that includes multiple data sources such as interviews, focus groups, town hall meetings, transect walks, questionnaires, and/or community mapping. Attention is paid to rationales for why methods and/or community mapping. Attention is paid to rationales for how methods are sequenced.

The team engages in collaborative analysis of the data. This ideally includes member checks, which are participant validation strategies that help the evaluators to challenge and verify their analyses of the data by having participants vet and comment on their interpretations of the data and findings. The team and stakeholders collaboratively work to develop a shared understanding around the findings and how they relate to the goals and outcomes of the evaluation and the policy or program in focus. This can require debate that is viewed as generative dialogue where all viewpoints are shared.

Once the findings have been discussed with all appropriate stakeholders and research partners and consensus is reached, an action plan is formulated and then vetted by a range of stakeholders, refined, and then disseminated to all relevant parties.

Evaluations ideally are conducted at a midterm point and then at the end of a program's implementation. A midterm evaluation helps inform midcourse corrections, so that deviations or problems in implementation, if any, can be addressed. An endterm evaluation examines the entire process to assess whether and how a policy or program was implemented with fidelity, whether and how it met its key objectives, and what forces mediated these outcomes.

Methods of Participatory Evaluation

There is a range of possible data collection methods that can be used in a participatory evaluation, and these are ideally strategically combined and sequenced.

Focus groups consist of three or more participants interviewed in groups to allow for group interaction, engagement, and the emergence of generative groupthink (i.e., the ways that members of a focus group build upon and/or challenge each other's responses). In *talking circles*, the goal is to equalize communication and encourage stakeholders to speak openly about ideas and/or concerns. These circles create a symbolic understanding that no one within a group is in a position of prominence. There is often an object passed around for the speaker in focus to hold while the speaker is talking.

Interviews are 1:1 conversations that are on a continuum from fixed questions to open-ended questions. The goal is focused engagement with a set of individuals to learn about their lived realities, experiences, ideas, and concerns, as they relate to the program or policy under evaluation.

Questionnaires or surveys can be administered on paper, online, or read to individuals who are not literate; they serve to create a broad sense of experience that can be compared and contrasted to identify patterns across individuals or subgroups. As well, these data generated can be situated in relation to the rest of the data set which includes more in-depth, individual-centered responses. Interested and affected parties can also be given an opportunity to share in *written representations* their opinions, viewpoints, suggestions, concerns, and/or ideas for improving the program and/or evaluation process publicly or confidentially through anonymous suggestion boxes or other methods.

In *town hall meetings*, all members of a specific community, organization, or region are invited to a public forum to discuss ideas, concerns, and wishes for a program or policy and/or the evaluation. These can also support the identification of evaluation team members at the outset and, throughout, are a forum to share concerns or ideas that otherwise might not emerge.

Gallery walks get people moving around rather than sitting in listening mode. The goal is to actively involve participants in representing, discussing, and synthesizing important concepts, in processes of consensus building, in writing out and sharing thoughts and questions, and in dialogic engagement. *Social network, community, and institutional mapping* is an efficient and inexpensive tool for collecting community or institutional data (descriptive, diagnostic, and/or analytic) with the guidance and narration of insiders. Mapping allows for the collection of contextualized data that can be used in a range of ways.

Secret ballots allow for the participation of stakeholders through an anonymous process, wherein participants can provide honest views on issues that deserve attention without fear of backlash. *Priority ranking* is a mixed-methods approach to data collection, in which selected participants are guided in generating responses to a specific question or issue in ways that visually represent their priorities. This draws on both quantitative and qualitative methods, so that data are contextualized and can be ranked and compared within and across stakeholder groups.

Transect walks introduce evaluators to a community context. Local community members serve as guides and orient evaluators to the geographic areas and localities within them by walking them to places and narrating why these places are relevant and important.

Photovoice is a method that developed from concern about authentic representation(s) of the perspectives and experiences of marginalized populations (including those who are nonliterate). Participants are provided with cameras and asked to chronicle daily experiences, contexts, and/or events; they then engage in group discussions about the images and their contexts, using the photos to narrate their importance. *Pictorial depictions* are largely used among stakeholders who are semi-or nonliterate as a way to share their points of view through drawing or other forms of artistic expression.

Performance audits are independent examinations of a program, policy, or organization to assess if it is achieving efficiency and effectiveness in the enactment of available resources.

Benefits and Challenges of Participatory Evaluation

There are myriad benefits of using a participatory approach to evaluation. It can provide deeper insight into a program or policy because the problems, foci, findings, and solutions emerge from a range of diverse, local stakeholders rather than external actors. Engaged groups can gain agency and a sense of ownership because the problems and solutions emerge from them. This helps with continuity and sustainability of projects, program, and policies.

Participatory evaluation creates structures and processes to bring forward the voices of a range of local actors who otherwise might be silenced or unheard, including those from marginalized, stigmatized, and underserved groups. It encourages engagement and collaboration within, outside, and across stakeholder groups and allows for genuine knowledge sharing and knowledge transfer. It elevates the reliability and validity of the data because representative stakeholder groups are engaged at each step of the research process.

At the same time, there can be challenges in and obstacles to participatory approaches to evaluation. It requires facilitation by individuals who are trained in participatory methodologies, which requires time and resources. Without this, the evaluations can yield unreliable data.

Bringing together multistakeholder groups requires active planning and coordination and can be time and resource intensive. Conflict resolution is crucial, as there are often situations and/or populations that are at odds, and disagreements can result in conflict or derail projects. Resolving conflicts requires specific skills and understanding the sociopolitical terrain. Establishment of rapport and trust among various stakeholder groups can be a time-consuming and intensive process. At times, it can be fraught with given power asymmetries.

Sharon M. Ravitch and B. Venkatesh Kumar

See also Action Research; Collaborative Evaluation; Data; Data Mining; Evaluation; Focus Groups; Interviews; Member Check; Methodology; Qualitative Data Analysis; Qualitative Research Methods; Quantitative Research Methods; Trustworthiness; Validity

Further Readings

Bamberger, M., Rugh, J., & Mabry, L. (2011). RealWorld evaluation: Working under budget, time, data, and political constraints (2nd ed.). Thousand Oaks, CA: SAGE.

Chilisa, B. (2012). Indigenous research methodologies. Thousand Oaks, CA: SAGE.

Cousins, B., & Chouinard, J. A. (2012). Participatory evaluation up close: An integration of research-based knowledge. Charlotte, NC: Information Age.

Hacker, K. A. (2013). Community-based participatory research. Thousand Oaks, CA: SAGE.

Nancy N. Boyles Nancy N. Boyles Boyles, Nancy N.

Partnership for Assessment of Readiness for College and Careers Partnership for assessment of readiness for college and careers

1220

1222

Partnership for Assessment of Readiness for College and Careers

The Partnership for Assessment of Readiness for College and Careers (PARCC) is a consortium of states that is developing new assessments for mathematics and English language arts based on the Common Core State Standards. The assessments are administered in Grades 3 through 8 and high school in states that have selected PARCC as their assessment system. PARCC assessments measure whether students' performance is on track for them to succeed academically in K–12 schools and ultimately in college and careers.

PARCC tests are designed to measure deep thinking across standards. The tests measure not only lower level knowledge and skills, which characterized many state assessments in the past, but also critical thinking, problem-solving, and communication. The tests are administered annually during a window of approximately 30 days at the end of the school year. Schools typically complete their testing in 1–2 weeks.

In English language arts, the PARCC test asks students to read, analyze, and write about a variety of fiction and nonfiction passages and includes three literacy task types: research simulation, literary analysis, and narrative writing. For the research task, students analyze an informational topic presented through several texts or multimedia stimuli, answer a series of questions, and write an analytic response to a prompt, synthesizing information from multiple sources.

For the literary analysis task, students read and analyze two pieces of literature, which could include short stories, novels, or poems. Students write an analytic response to a prompt based on the literary texts.

For the narrative writing task, students read a literary text from a short story, novel, poem, or other type of literature and then write a narrative response to a prompt based on this literary text.

PARCC mathematics items measure critical thinking, mathematical reasoning, and the ability to apply skills and knowledge to real-world problems. Students are asked to solve problems involving the key knowledge and skills for their grade level based on the Common Core State Standards, using mathematical practices, reasoning, and modeling.

PARCC has developed scoring rubrics to evaluate student responses and has set performance levels that indicate the level of performance a student's work demonstrates. Level 1 indicates the greatest need for improvement. Level 5 shows the strongest performance.

Assessment results are intended to help teachers customize learning plans to better meet student needs. Detailed score reports with information about the extent to which students are mastering knowledge, skills, and deep-thinking processes are provided online to schools and teachers for this purpose. Easy-to-understand printed reports are available for parents.

To support students' individualized learning plans, PARCC offers instructional tools that teachers may use throughout the school year. These include formative tasks for Grades K–2, 3–8, and high school; diagnostic tools; speaking and listening tools; and performance-based modules. One tool available to teachers to guide classroom instruction is the released assessment items from previous PARCC tests.

Nancy N. Boyles

See also Achievement Tests; Common Core State Standards; Formative Assessment; Smarter Balanced Assessment Consortium; Summative Assessment

Further Readings

Chingos, M. M. (2013). Standardized testing and the Common Core Standards: You get what you pay for? Washington, DC: Brookings Institution Press, Brown Center on Education Policy. Retrieved from https://www.brookings.edu/research/standardized-testing-and-the-common-core-standards-you-get-what-you-pay-for/

Doorey, N., & Polikoff, M. (2016). Evaluating the content and quality of next generation assessments. Washington, DC: Fordham Institute. Retrieved from https://edexcellence.net/publications/evaluating-the-content-and-quality-of-next-generation-assessments

The Partnership for Assessment of Readiness for College and Careers. (n.d.). Assessments. Retrieved from www.parcconline.org/assessments

The Partnership for Assessment of Readiness for College and Careers. (n.d.). The PARCC tests. Retrieved from http://www.parcconline.org/about/the-parcc-tests

The Partnership for Assessment of Readiness for College and Careers. (n.d.). Parent resources. Retrieved from http://www.parcconline.org/resources/parent-resources

The Partnership for Assessment of Readiness for College and Careers. (n.d.). Resources. Retrieved from www.parcconline.org/resources

The Partnership for Assessment of Readiness for College and Careers. (n.d.). Score results. Retrieved from http://www.parcconline.org/assessments/score-results

The Partnership for Assessment of Readiness for College and Careers. (n.d.). Test design. Retrieved from http://www.parcconline.org/assessments/test-design

Partnership Resource Center. (n.d.). PARCC released items. Retrieved from https://prc.parcconline.org/assessments/parcc-released-items

Yanyun Yang Yanyun Yang Yang, Yanyun

Path Analysis Path analysis

1222

1225

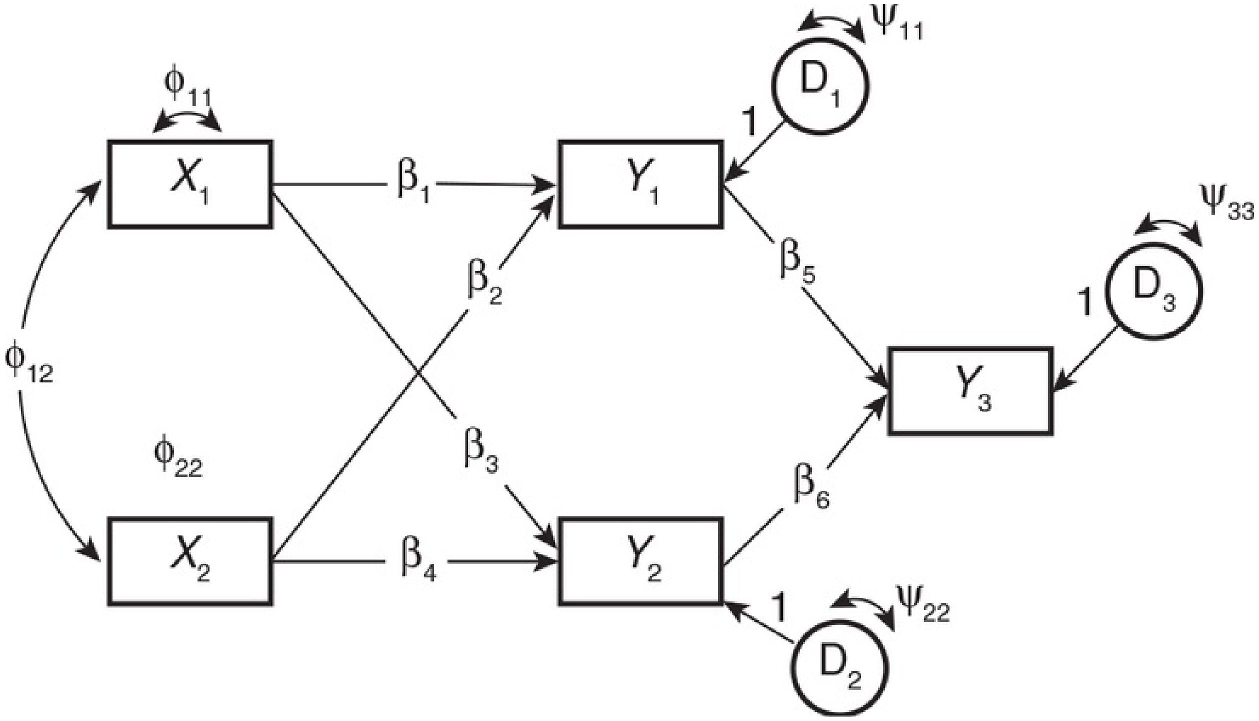
Path Analysis

Path analysis is a statistical procedure for testing the causal relationship between observed variables. In a path analysis model, this cause–effect relationship is not discovered via data analysis but instead is formulated based on the researcher’s knowledge or on previous studies. Path analysis was initially developed by Sewall Wright in 1921 for examining the direct and indirect effects of variables on other variables; a century later, it continues to be a popular statistical procedure. Since the rapid development starting in the 1970s of a more comprehensive family of statistical techniques known as structural equation modeling (SEM), path analysis has been viewed as a special type of SEM in which only observed variables are involved in the analysis.

Example of a Path Analysis Model

As an example, a researcher formulates the following hypotheses: X_1 and X_2 are the common causes of Y_1 and Y_2, and both Y_1 and Y_2 are the causes of Y_3. This path analysis model can be represented in a path diagram (Figure 1).

Figure 1 An Example of a Path Analysis Model



The corresponding regression equations are

$$Y_1 = \tau_1 + \beta_1 X_1 + \beta_2 X_2 + D_1,$$

$$Y_2 = \tau_2 + \beta_3 X_1 + \beta_4 X_2 + D_2,$$

$$Y_3 = \tau_3 + \beta_5 X_1 + \beta_6 X_2 + D_3,$$

where β_1 to β_6 denote the path coefficients from a predictor to an outcome variable, and D_1 to D_3 are the residuals for the corresponding outcome variable. An intercept term τ is included in each equation. Therefore, this set of equations corresponds to unstandardized regression models. Although it is possible, the intercept terms are typically not shown in the path diagram when the purpose of the analysis is the cause–effect relationship between variables.

Key Components

In a path analysis model, an observed variable is presented within a square or rectangle. An observed variable is either *exogenous* (X_1 and X_2) or *endogenous* (Y_1, Y_2, and Y_3), which corresponds to a predictor or outcome variable, respectively, in a regression model. The cause of an exogenous variable is not included in the model, whereas the cause of an endogenous variable is explicitly specified. The causal relationship is indicated by a single-headed arrow (e.g., $X_1 \rightarrow Y_1$ in Figure 1), with the variable at the tail of the arrow being the cause and the variable at the head of the arrow being the effect. The *direct effect* from one variable to another is quantified by the path coefficient, similar to the slope in regression analysis. In a path analysis model, exogenous variables may also affect endogenous variables through some intermediate variables ($X_1 \rightarrow Y_1 \rightarrow Y_3$). The intermediate variables are sometimes called *mediators*, and the mediating effect is called the *indirect effect*. The double-headed curved arrow at the top of an exogenous variable indicates the *variance* of the exogenous variable (φ_{11}, φ_{22}), and the double-headed curved arrow connecting two exogenous variables is the *covariance* between the two variables (φ_{12}). Each endogenous variable has an unobserved *disturbance* (D_1, D_2, and D_3), represented within a circle or oval. A disturbance contains two confounded components: the measurement error of the endogenous variable and all the causes of the endogenous variable that are not explicitly specified in the model. If it is known that a pair of endogenous variables share some common missing causes, their disturbances can be correlated. The path coefficient from a disturbance to its endogenous variable is typically fixed at 1 to assign a metric to the disturbance.

Identification Issues

When the mean structure (including the means of the exogenous variables and the intercepts of the endogenous variables) is not of interest in a path analysis, the number of freely estimated model parameters is the sum of (a) the direct effects from one variable to another (β_1–β_6), (b) the variance and covariance of the exogenous variables (φ_{11}, φ_{22}, and φ_{12}), and (c) the variance and covariance of the disturbances (ψ_{11}, ψ_{22}, and ψ_{33}). This number should be less than or equal to the number of observations provided in the data. The number of observations is the number of variance and unique covariance among the observed variables, computed as $v(v+1)/2$, where v is the number of observed variables involved in the path analysis. The model degrees of freedom (*df*) is equal to the number of observations minus the number of freely estimated model parameters. In this example, $df = 15 - 12 = 3$.

A path analysis model should be theoretically identified in order to be testable. Theoretical identification is a property of a model and not of data. Two basic assumptions for determination of model identification are that (1) each disturbance has a metric and (2) $df \geq 0$. An *underidentified* path model has $df < 0$; it is not testable and should be respecified. A *just-identified* path model has $df = 0$; it always perfectly describes the relationship between the observed variables. In other words, the covariance matrix among the observed variables is reproduced perfectly by using the model parameters including path coefficient, variance and covariance among exogenous variables, and variance and covariance among the disturbances. The covariance matrix produced by the model parameters is called a *model-implied covariance matrix*. An *overidentified* path analysis model has $df > 0$. The positive number in *df* results from the number of omitted direct causal relationship between variables posited in the hypothesized model.

In this example, the direct effects connecting these three pairs of variables (therefore, $df = 3$) are omitted: (1) X_1 and Y_3, (2) X_2 and Y_3, and (3) Y_1 and Y_2. If the parameters from this model still will reproduce the covariance matrix among the observed variables, the researcher can gain confidence that the hypothesized model is indeed a reasonable representation of the relationship between variables. For this reason, an overidentified path analysis model is more interesting than a just-identified model. A path analysis model that is theoretically identified does not necessarily converge to proper solutions. A model may encounter empirical identification due to some unexpected characteristics of data, for example, extreme collinearity between variables.

The classic path analysis considers only unidirectional causal relationships between variables. Such a path model is *recursive*, which is always theoretically identified with $df \geq 0$. A nonrecursive path model contains a bidirectional causal relationship, feedback loop, or covariance between disturbances with direct effects between the corresponding endogenous variables. The two basic requirements described earlier are not sufficient to identify a nonrecursive path model. Other requirements such as rank condition and order condition should be met.

Parameter Estimation and Evaluation of Model–Data

Fit

For a given sample size *N*, model parameters are commonly estimated using the maximum likelihood estimation method through an iterative procedure. A set of estimates is chosen to maximize the probability of generating the observed data; at the same time, the difference between the sample covariance matrix (**S**) and the model-implied covariance matrix is minimized. The minimization of the difference between the two matrices is quantified by the fit function (*F*). The fit function for a maximum likelihood estimation is

$$F_{MLE} = \log\left|\hat{\Sigma}\right| + \mathrm{trace}\left(\mathbf{S}\hat{\Sigma}^{-1}\right) - \log\left|\mathbf{S}\right| - \nu,$$

where log|.| is the natural logarithm of the determinant of the corresponding matrix.

For a just-identified path model, the fit function is zero because the model-implied covariance matrix is identical to the observed covariance matrix. Evaluation of model–data fit is not of interest for a just-identified model. For an overidentified path model, the fit function is usually greater than zero. A higher value for the fit function indicates a greater discrepancy between the two matrices, suggesting model–data misfit.

The *t* statistic is computed as $(n - 1)F$ and is used for testing the null hypothesis, , where **Σ** denotes the covariance matrix among the observed variables in the population. There are two sources of a positive value of *t*: (1) sampling fluctuation and (2) model misspecification. In the example given earlier, if the three pairs of variables (X_1 and Y_3, X_2 and Y_3, and Y_1 and Y_2) indeed have zero direct relationships in the population, then sampling fluctuation is the only source of a positive value of *t*. The sampling distribution of the *t* statistic would then follow asymptotically a central chi-square distribution, with the expected value being the model *df* when the distributional assumption of data is met. However, if at least one of these three pairs of variables actually has a nonzero direct relationship in the population but is formulated as zero in the model, the sampling distribution of the *t* statistic would then follow a noncentral chi-square distribution, with the expected value being the model *df* plus the noncentrality parameter λ, where λ quantifies the degrees to which the path analysis model is misspecified in the population. SEM software denotes the *t* statistic as a chi-square statistic. A *p* value greater than the predefined α level means failure to reject H_0, which is an indication of good model–data fit.

In addition to a chi-square test, a number of fit indices are often used, including but not limited to the root mean square error of appropriation, comparative fit index, Tucker–Lewis index, and standardized root mean square residual. Each index sheds some light on model–data misfit. The purpose of using multiple fit indices is to obtain a diagnosis of the source of the model–data misfit.

Interpretation of Model Parameters From a Path Analysis

An overidentified model often does not demonstrate an adequate fit to the data in applications. If a hypothesized path model demonstrates a poor fit to the data, parameter estimates should not be interpreted and the model is often respecified. Once a theoretically meaningful path model yields a reasonable fit to the data, parameter estimates can be reported in either the unstandardized or standardized form, along with their standard errors (*SE*s). Null hypothesis testing can be conducted to examine whether the sample estimate of a parameter comes from a population with the specified value (often zero). The unstandardized and standardized path coefficients are interpreted in a similar way to the slope in multiple regression: the expected unit (or standard deviation) change in the endogenous variable for a single-unit (or standard deviation) change in the predictor, controlling for all other predictors. The standardized variance of disturbance is the percentage of variance in an endogenous variable that is not explained by its predictors; one minus this value is similar to R^2 in multiple regression.

In a path analysis, the direct and indirect effects of variables on other variables are of primary interest. The direct effects are indicated by the path coefficients, and a two-tailed or one-tailed t test can be performed to test for statistical significance. The coefficient indicating indirect effect from one variable to another through a mediator is estimated by the product of the two direct effects. Taking $X_1 \rightarrow Y_1 \rightarrow Y_3$ as an example, the indirect effect is estimated by $b_1 \times b_5$, where b_1 and b_5 are the sample estimates of β_1 (i.e., for $X_1 \rightarrow Y_1$) and β_5 (i.e., for $Y_1 \rightarrow Y_3$), respectively. The coefficient of indirect effect is interpreted in the same way as a direct effect. However, the *SE* associated with the indirect effect is difficult to estimate because of the complexity of the distributions. A well-known method for approximating the *SE* of indirect effects is a Sobel test using the delta method: where follows a *z* distribution. Consequently, a statistical significance test can be conducted and confidence intervals can be constructed for each indirect effect. Bootstrapping is another popular method to empirically approximate the sampling distribution of the indirect effect. When mediating variables are involved, the sum of the direct effect and all indirect effects forms the total effect of an exogenous variable on an endogenous variable. Total effect is interpreted in the same way as a path coefficient.

Assumptions in Path Analysis

Similar to other statistical procedures, path analysis requires a certain number of assumptions. Although the classic path analysis requires all relationships between variables to be linear and recursive and no correlated disturbance in the model, these assumptions can be relaxed when conceptualizing the path analysis in an SEM framework. Two key assumptions follow.

First, all variables, particularly the exogenous variables, are measured without error; that is, reliability = 1. This assumption is routinely violated because nearly all measures are imperfect. Violation of this assumption may lead to estimation bias. A more flexible family of statistical techniques, SEM, can take measurement error into account. Another key assumption is that the correlation between each predictor and the disturbance is zero, that is, $\mathrm{Cov}(X,e)=0$. This assumption is unlikely to hold when the path analysis model is misspecified. As previously noted, the disturbance of an endogenous variable contains both measurement error and all predictors that are not explicitly specified in the model. If these predictors share some commonality with any of the predictors in the model, then $\mathrm{Cov}(X,e)\neq 0$. Researchers are advised to include all important predictors in the path analysis to avoid violating this assumption.

Yanyun Yang

See also Bootstrapping; Correlation; Structural Equation Modeling

Further Readings

Bollen, K. A. (1989). Structural equations with latent variables. New York, NY: Wiley.

Kenny, D. A., & Milan, S. (2012). Identification: A non-technical discussion of a technical issue. In R. Hoyle (Ed.), Handbook of structural equation modeling (pp. 145–163). New York, NY: Guilford.

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, XX(7), 557–585.

Kimberly Capp Kimberly Capp Capp, Kimberly

Kimberly Ethridge Kimberly Ethridge Ethridge, Kimberly

Anthony Odland Anthony Odland Odland, Anthony

Peabody Picture Vocabulary Test Peabody picture vocabulary test

1225

1228

Peabody Picture Vocabulary Test

The Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4) assesses the receptive vocabulary for American English among children and adults from 2½ through 90 years and older. It is a useful aid in the diagnosis of reading disorders and language impairments across age ranges. This entry describes the testing procedures, historical development, and test validity and discusses the applications and limitations of the PPVT-4.

Testing Procedures

Test materials include an easel, protocols for two alternate forms (A and B), and a manual. The PPVT-4 Forms A and B are individually administered; each consists of training items and 228 test items. Each item has four enlarged full-color pictures that are presented on the easel as response options for the examinee. The administrator states a word from the protocol, and the examinee selects the picture that best illustrates that word’s meaning; examinees may make their selection by either pointing to the corresponding illustration or by verbally stating the number coinciding with the illustration. Test content consists of a range of receptive vocabulary reflecting different parts of speech and 20 content areas. Item difficulty increases as the examinee proceeds with the test. There is no time limit for the test or the individual items, though the average administration time is approximately 15–20 minutes. If an examinee takes longer than 10 seconds to respond to an item, the examiner may prompt the examinee to attempt an answer.

Test administration begins with examinee training: Training Page A is for examinees younger than 4 years and training Page B for examinees 4 years and older. These training items are intended to teach the examinee how to respond to each administered item. Training items also provide the examiner with information regarding examinee capability. Administration starting points are provided for various ages based on an expected ability level that was established during test standardization. The basal set has been established when the examinee makes one or zero errors on the examinee’s first set of 12 items. If an examinee makes more than one error at the starting point corresponding to the examinee’s age, the examiner administers items in reverse sequence until the basal set has been established. The ceiling set is established when an examinee makes eight or more consecutive errors. If an examinee makes eight or more consecutive errors in a set of 12 items, testing should be discontinued.

The PPVT-4 can be scored either by hand or online using the Pearson Q-global scoring and reporting system. Computer scoring is performed by entering individual item responses or raw score data. Scores can be reported using either age-or grade-based standard scores that range from 20 to 160. Scores have a mean of 100 and a standard deviation of 15. Percentiles, normal curve equivalents, stanines, age and grade equivalents, and growth scale values are also available for reporting.

Revisions

The PPVT was published in 1959 by Lloyd M. Dunn and Leota M. Dunn. In 1981, a revised version was published that expanded upon its predecessor, and a third edition (published in 1997) featured updated content and expanded norms. In 2007, the current edition, the PPVT-4, was released. Added features in the PPVT-4 include new stimulus words, updated and expanded norms, additional easy items, streamlined administration procedures, a scale for measuring change over time, expanded interpretive options that include analysis of item content by part of speech, and larger and full-color illustrations that have been evaluated to ensure they can be perceived by individuals who are color blind.

Standardization Procedure

From fall 2005 to spring 2006, standardization data for the PPVT-4 were collected by 450 examiners at 320 test sites throughout the United States. The normative sample consisted of 3,540 cases normed by age (2½ years through 90 years and older) and 2,003 cases normed by grade (kindergarten through Grade 12). The examinees were from four geographic regions: the Northeast (17.5% of age-normed individuals and 18.0% of grade-normed individuals), North Central (23.1% age-normed individuals and 22.9% grade-normed individuals), South (38.0% age-normed individuals and 37.1% grade-normed individuals), and West (21.3% age-normed individuals and 22.0% grade-normed individuals).

The normative sample was also stratified by gender, race/ethnicity, and level of education (self or parent). The age-normed sample consisted of 50.6% females and 49.4% males; 15.1% were Black, 15.4% Hispanic, 63.4% White, and 6.1% qualified as Other. The education level for this sample consisted of 12.1% who achieved Grade 11 or lower, 27.9% with Grade 12 or GED certification, 31.5% with 1–3 years of college, and 28.5% with 4 or more years of college. The grade-normed sample consisted of 50.1% females and 49.9% males, with 15.8% being Black, 15.9% Hispanic, 62.0% White, and 6.3% considered as Other. The education level for the grade-normed sample consisted of 10.4% with 11th-grade education or lower, 27.3% with Grade 12 or GED, 33.8% with 1–3 years of college, and 28.5% with 4 or more years of college.

In addition, representative samples of special populations for examinees aged 2 through 18 years were included: speech and language impairment, mental retardation and developmental delay, specific learning disability, emotional and behavioral disturbances, attention-deficit/hyperactivity disorder, and autism.

Psychometrics

The PPVT-4 possesses high reliability (i.e., between .80 and 1.00). Reliability estimates for the PPVT-4 used internal consistency, alternate-form reliability, and test–retest reliability. Split-half reliability for the PPVT-4 was calculated for each age-group and ranged from .89 to .97. Alternate-form reliability for Forms A and B ranged from .87 to .93, suggesting homogeneity between forms. Test–retest reliability for the PPVT-4 ranged from .92 to .96.

In terms of content validity, stimulus words were based on a review of over 12 published sources to represent 20 content areas. When compared to the Comprehensive Assessment of Spoken Language, the PPVT-4 correlated .50 for basic concepts, .41 for antonyms, and .54 for sentence completion for ages 3–5 years. Additional correlations with the Comprehensive Assessment of Spoken

Language include synonyms .65, antonyms .78, sentence completion .63, and lexical/semantic composite .79 for ages 8–12 years. The average correlation was .82 across all age ranges when compared to the Expressive Vocabulary Test, Second Edition. When the PPVT-4 was correlated with the Clinical Evaluation of Language Fundamentals, Fourth Edition, core language was .73, receptive language was .67, and expressive language was .72 for ages 5–8 years. Additionally, the PPVT-4 correlated with core language (.72), receptive language (.75), and expressive language (.69) on the Clinical Evaluation of Language Fundamentals, Fourth Edition, for ages 9–12 years. When correlating the PPVT-4 to scores on the Group Reading Assessment and Diagnostic Evaluation test, the average correlation was .63 across grade levels. When the PPVT-4 was compared to the PPVT-3, the correlation was .84.

Applications

The PPVT-4 can be used to characterize a broad range of language-based difficulties, assist with diagnosis of conditions and disorders related to language impairment, and help with treatment planning. There are a variety of difficulty levels; therefore, the test is a useful measure of aphasias and vocabulary deterioration in adults. Because no reading or writing is required, the PPVT-4 is useful in screening individuals who are unable to read and those with written-language difficulties as well as preschool-aged children. Additionally, because the PPVT-4 measures vocabulary acquisition and understanding of spoken words in standard American English, it can be used to assess an individual’s understanding of the language. Likewise, the PPVT-4 can be used as a screening tool for occupations in which good listening comprehension skills are required. The PPVT-4 provides a good assessment of an examinee’s receptive vocabulary knowledge for English in individuals whose primary language is not English.

Qualitative analysis of an examinee’s response style on the PPVT-4 can provide insight on potential neurological damage. For example, if the examinee cannot detect or describe the pictures in the stimulus book, visual agnosia may be present. If the examinee ignores answer choices on one side of the response easel, a visual field cut, visual neglect, hemispatial inattention, or impulsivity may be present.

The PPVT-4 is also a useful assessment for diverse populations that may exhibit difficulties being assessed with other measures. For example, the black outline on the full-color illustrations enables the PPVT-4 to be used with individuals who have moderate visual disabilities such as visual–perceptual problems and color blindness. The PPVT-4 is also advantageous in assessing individuals with cerebral palsy or other major physical disabilities. If the examinee is unable to point to the examinee’s selection, it is acceptable for the administrator to point to each option and ask the person to shake or nod the head to indicate yes or no to each selection. Because the PPVT-4 is administered individually and does not require that the examinee interact directly with the examiner, it may be a good assessment to use on those who perform poorly on group tests, individuals with autism, withdrawn individuals, or individuals who exhibit symptoms of psychosis.

In combination with other neuropsychological/psychological measures and extratest data, the PPVT-4 can assist with different aspects of treatment planning by allowing the clinician to observe examinee strengths and weaknesses. Qualitative data collected during administration can enable the clinician to determine whether low scores are due to a deficit in receptive vocabulary or caused by other physical or language barriers. In addition, the growth scale value metric provided by the PPVT-4 allows a clinician to measure change over time to see whether a particular intervention has been successful in improving the receptive vocabulary abilities of an examinee. An increase of 8 points on the growth scale value score indicates that vocabulary has improved.

Limitations

The PPVT is only suitable for examinees who are able to hear the administrator call out the stimuli words and those who are able to visually distinguish between the stimuli presented on the easel in order to make a selection. Without appropriate adaptation, this measure is not suitable for deaf or blind examinees given the requisite of some level of vision and hearing. Those individuals with moderate visual impairment or colorblindness should still be able to undergo assessment using the PPVT-4 because of the larger size and black outlines of the full-color illustrations. As with any psychological instrument, appropriate consideration must be made with respect to the background characteristics of the examinee. For example, it should be noted that the majority of the normative sample consisted of individuals identifying as White. Therefore, caution should be exercised when applying results and interpretations to those from different racial backgrounds. The PPVT-4 is only available in standard American English, which should be taken into consideration for examinees who speak other varieties of English. It is important to note that the PPVT-4 is not meant to serve as a diagnostic tool but rather to aid in the formulation of a diagnosis and to inform treatment for the examinee. It is important that clinicians conduct a thorough assessment using multiple measures and employ their clinical judgment when formulating a diagnosis.

Kimberly Capp, Kimberly Ethridge, and Anthony Odland

See also Achievement Tests; Wechsler Intelligence Scales

Further Readings

Community–University Partnership for the Study of Children, Youth, and Families. (2011). Review of the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4). Edmonton, Canada.

Dunn, L. M., & Dunn, D. M. (2007). PPVT-4 manual. Bloomington, MN: NCS Pearson.

Dunn, L. M., & Dunn, D. M. (2015). Peabody picture vocabulary test: PPVT 4. Bloomington, MN: NCS Pearson

Golden, C. J. (1979). Clinical interpretation of objective psychological tests. New York, NY: Grune & Stratton.

Golden, C. J., Espe-Pfeifer, P., & Wachsler-Felder, J. (2000). Neuropsychological interpretation of objective psychological tests. New York, NY: Springer Science & Business Media.

Mervis, C. B., & Pitts, C. H. (2015, June). Children with Williams syndrome: Developmental trajectories for intellectual abilities, vocabulary abilities, and adaptive behavior. American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 169(2), 158–171.

Percy-Smith, L., Busch, G., Sandahl, M., Nissen, L., Josvassen, J. L., Lange, T.,

& Cayé-Thomasen, P. (2013). Language understanding and vocabulary of early cochlear implanted children. International Journal of Pediatric Otorhinolaryngology, 77(2), 184–188. doi:10.1016/j.ijporl.2012.10.014

Phillips, B. A., Loveall, S. J., Channell, M. M., & Conners, F. A. (2014). Matching variables for research involving youth with Down syndrome: Leiter-R versus PPVT-4. Research in Developmental Disabilities, 35(2), 429–438. doi:10.1016/j.ridd.2013.11.016

Shaw, S. R. (2008). Review of the peabody picture vocabulary test (4th ed.). In K. F. Geisinger, R. A. Spies, J. F. Carlson, & B. S. Plake (Eds.), The eighteenth mental measurements yearbook (pp. 143–144). Lincoln, NE: The Buros Institute of Mental Measurements.

Spaulding, T. J., Hosmer, S., & Schechtman, C. (2013). Investigating the interchangeability and diagnostic utility of the PPVT-III and PPVT-IV for children with and without SLI. International Journal of Speech-Language Pathology, 15(5), 453–462. Retrieved from http://dx.doi.org/10.3109/17549507.2012.762042

Matthew Gordon Ray Courtney Matthew Gordon Ray Courtney Courtney, Matthew Gordon Ray

Pearson Correlation Coefficient Pearson correlation coefficient

1228

1233

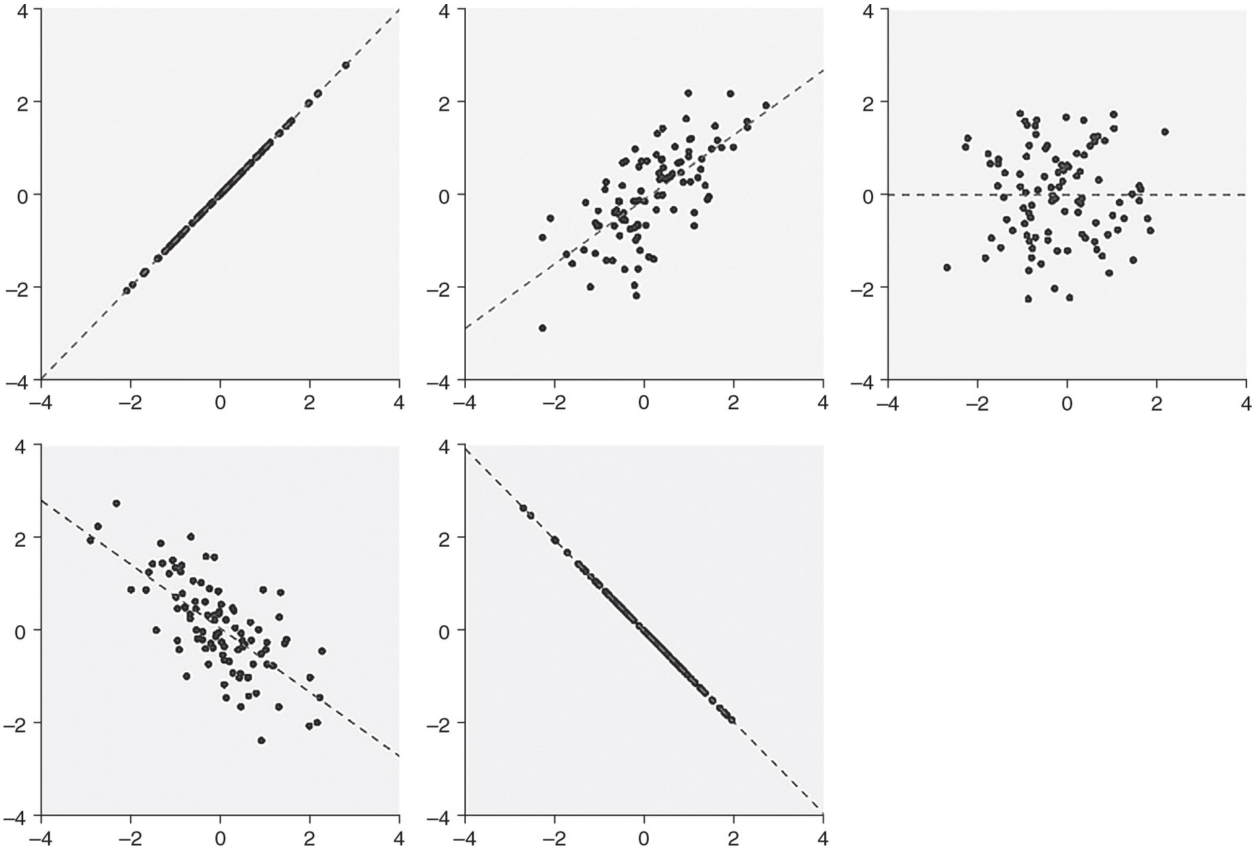
Pearson Correlation Coefficient

Karl Pearson (1857–1936) is credited with establishing the discipline of mathematical statistics. Building on earlier work by Francis Galton (1822–1911), one of Pearson’s major contributions to the field was the development of the Pearson product-moment correlation coefficient (or Pearson correlation, for short), which is often denoted by *r*. The correlation is one of the most common and useful statistics. The Pearson correlation, a measure of the relationship often between two continuous variables, is utilized throughout quantitative research in education and the social sciences. This entry is devoted to describing what the Pearson correlation is; the steps used to calculate it; the interpretation of its size, direction, and level of statistical significance; its data assumptions; and its limitations.

What Is the Pearson Correlation?

Put simply, the Pearson correlation is a measure of the linear relationship between two variables, *X* and *Y*, giving a value between +1.0 and −1.0, where 1.0 is a perfect positive correlation, 0.0 (zero) is no correlation, and −1.0 is a perfect negative correlation. Examples of the possible data distributions associated with five Pearson correlations are illustrated in Figure 1.

Figure 1 Example of five Pearson product-moment correlation coefficients

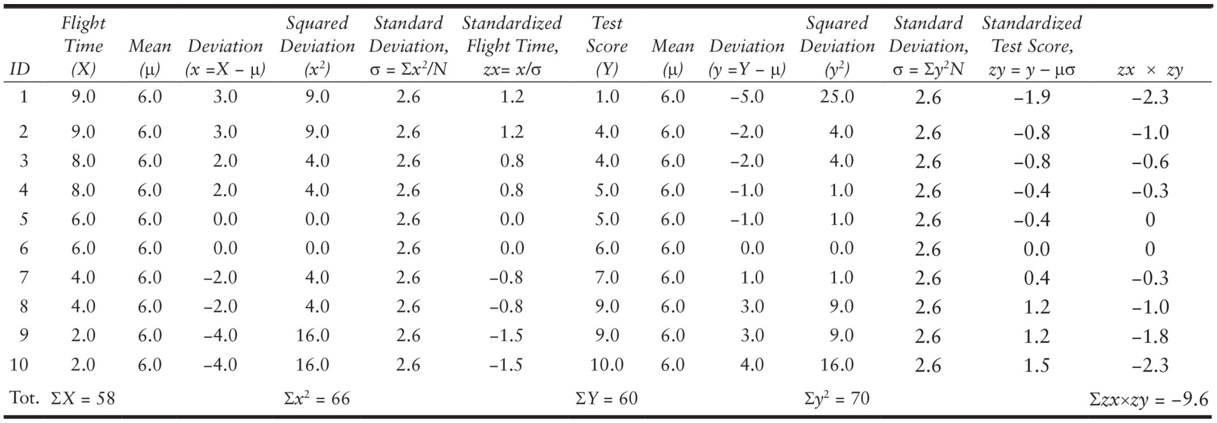


Importantly, where correlational estimates are concerned, there is no attempt to establish one of the variables as independent and the other as dependent. Therefore, relationships identified using correlation coefficients should be interpreted for what they are: associations, not causal relationships. To arrive at a Pearson correlation value (*r*) between two variables of interest, a number of calculations and logical steps are made. To illustrate these steps, a fictional example of two educational variables of interest is now provided.

Calculation of the Pearson Correlation Coefficient

Suppose you are the head of curriculum at a small English as a Second or Other Language institute in Auckland, New Zealand. A new cohort of intermediate-level English as a Second or Other Language students arrives every 10 weeks to participate in your program. The cohort flies to Auckland from various spots in the Asia-Pacific region: nearby in Polynesia, farther away in Micronesia, even farther in Southeast Asia, and at points beyond in East Asia. Being one of the teachers on the course, you notice a trend whereby, despite exhibiting equivalent levels of English fluency, the students originating from farther abroad tend to have more limited knowledge of New Zealand, its culture, and its customs, and often struggle with course material integrating such content. For the purpose of trying to better understand and tailor to the needs of particular student groups, you would like to explore the statistical relationship between (a) the distance that students travel to get to New Zealand and (b) their general knowledge of New Zealand.

To illustrate the steps taken to calculate a Pearson correlation, a fictional educational data set that includes a sample of one intake, namely 10 (N = 10) international students (Table 1, ID column) will be used. The time that it takes each student to fly directly to Auckland, New Zealand, the location of the course, can be used as a proxy measure of each student's distance of travel to New Zealand. The flight times are presented in Table 1, flight time column (X). On the first day of the course, the students sit a 10-item general knowledge test about New Zealand. The results of the test, out of 10, are also presented in Table 1, test score column (Y).



ID	*Flight Time (X)*	*Mean (μ)*	*Deviation (x =X − μ)*	*Squared Deviation (x^2)*	*Standard Deviation, σ = Σx^2/N*	*Standardized Flight Time, zx= x/σ*	*Test Score (Y)*	*Mean (μ)*	*Deviation (y =Y − μ)*	*Squared Deviation (y^2)*	*Standard Deviation, σ = Σy^2N*	*Standardized Test Score, zy = y − μσ*	*zx × zy*
1	9.0	6.0	3.0	9.0	2.6	1.2	1.0	6.0	−5.0	25.0	2.6	−1.9	−2.3
2	9.0	6.0	3.0	9.0	2.6	1.2	4.0	6.0	−2.0	4.0	2.6	−0.8	−1.0
3	8.0	6.0	2.0	4.0	2.6	0.8	4.0	6.0	−2.0	4.0	2.6	−0.8	−0.6
4	8.0	6.0	2.0	4.0	2.6	0.8	5.0	6.0	−1.0	1.0	2.6	−0.4	−0.3
5	6.0	6.0	0.0	0.0	2.6	0.0	5.0	6.0	−1.0	1.0	2.6	−0.4	0
6	6.0	6.0	0.0	0.0	2.6	0.0	6.0	6.0	0.0	0.0	2.6	0.0	0
7	4.0	6.0	−2.0	4.0	2.6	−0.8	7.0	6.0	1.0	1.0	2.6	0.4	−0.3
8	4.0	6.0	−2.0	4.0	2.6	−0.8	9.0	6.0	3.0	9.0	2.6	1.2	−1.0
9	2.0	6.0	−4.0	16.0	2.6	−1.5	9.0	6.0	3.0	9.0	2.6	1.2	−1.8
10	2.0	6.0	−4.0	16.0	2.6	−1.5	10.0	6.0	4.0	16.0	2.6	1.5	−2.3
Tot.	Σ*X* = 58			Σx^2 = 66			Σ*Y* = 60			Σy^2 = 70			Σ*zx*×*zy* = −9.6

Seven Steps to Calculate the Pearson Product-Moment Correlation Coefficient

A cursory look at variables X and Y suggests that there is an inverse relationship between the flight times and test scores. However, you want to determine statistically the size of the correlation, its direction, and its level of statistical significance.

Seven steps can be followed to generate the Pearson product-moment correlation coefficient. In Step 1, the mean flight time to New Zealand (μ) is generated. In Step 2, the extent to which each of the 10 observed flight times deviate from the mean (x) is then calculated. In Step 3, each of the 10 deviations are then squared (x^2). In this step, the four negative deviations of −2.0, −2.0, −4.0, and −4.0, when squared, result in positive values of 4.0, 4.0, 16.0, and 16.0, respectively. Because resultant deviation values will always be positive, a total measure of deviation for the variable can be determined by simple summation (Σx^2). In Step 4, this total measure of deviation is divided by the number of observations in the sample, namely 10. This procedure ($\Sigma x^2/N$) provides for an average squared deviation for variable, flight time. By finding the square root of “$\Sigma x^2/N$,” the average squared variation is reduced back to its original hour metric, thus determining the standard deviation (σ) of the flight time variable. In Step 5, standardized flight times (zx) are generated for each observation. This is done by simply dividing the degree to which each original flight time deviates from the mean by the standard deviation (x/σ). In Step 6, the standardized test results (zy) are ascertained by following the same logical procedures followed in Steps 1–5. Note that the ascending pattern of standardized test results (zx) mirrors that of the original ascending pattern of the test score (Y): while the test scores value range from 1 to 10, the standardized test score values now range from −1.9 to 1.5 (in instances when variables are normally distributed, 99.7% of the time, standardized values range between −3.0 and 3.0). In Step 7, standardized flight times (zx) are multiplied by standardized test scores (zy) to generate a product for each of the 10 observations. In this case, with the exception of the two products of zero (0), all products are negative, suggestive of an overall inverse relationship between the two variables of interest. By summing each of these products, and dividing that sum by the number of participants in the sample (N = 10), you arrive at a type of average relational value between the two variables, namely the Pearson product-moment correlational coefficient (r). In this case, r = −.96.

As the head of curriculum, now that you have determined what the Pearson correlation is for the sample, how might you interpret and make sense of it? Is the correlation small, medium, or large? Is it statistically significant? Let’s now take a look at some of the conventions for answering these questions.

The Size, Direction, and Statistical Significance of the Correlation Coefficient

For correlations derived from contexts in the behavioral sciences, Jacob Cohen identifies small, medium, and large correlations as *r* = |.20|, |.30|, and |.50|, respectively (note that the vertical bar "|"denotes an absolute value, possibly positive or negative). Therefore, the correlation of *r* = −.96 from the earlier example could be considered very large. Although Cohen's rules of thumb are often cited, other slightly higher thresholds have been proposed. For example, Mavuto Mukaka suggests low-, moderate-, and high-correlation coefficients of *r* = |.30|, |.50|, and |.70|. Nevertheless, whatever the yardstick, the example correlation between students' flight time and test score is most certainly large.

In addition to size, a correlation coefficient can also be interpreted in terms of its associated level of statistical significance. With reference to the correlational value derived in the example, statistical significance addresses the following question:

> Assuming that the 10 sampled participants came from a wider population in which no (exactly zero) correlation exists between flight times and test score, and given the sample correlation of *r* = .96 and *N* = 10, what is the calculated probability of that sample result?

We would expect the probability that the correlation (*r* = .96, *N* = 10) was an anomalous function of the sampled participants themselves—not reflective of the population in which no relationship exists—to be quite small. Without the use of statistical software programs, such as IBM's SPSS, we would make use of statistical rules and critical value charts to determine the level of statistical significance associated with the said finding. However, with computer software programs, we can quickly and easily determine a two-tailed degree of statistical significance associated with the said sample to be *p*(*probability*) = .0001 (the result is two-tailed, as opposed to one-tailed as the correlation was free to be either positive or negative in direction). This means that the probability that the sample result was derived from a population in which no correlation exists is less than 1 in 10,000, a very unlikely scenario indeed. Given that the threshold for statistical significance is often set at *p* = .05, we can assert that a strong negative and statistically significant correlation exists between the English as a Second or Other Language students' flight time and test scores.

Ultimately, the very strong negative statistically significant correlation has implications for the design and delivery of curriculum for cohorts enrolled in the course. For example, it might be a sound decision to provide students from farther abroad with background readings and other content to help them prepare for material in upcoming lessons.

Data Assumptions

The calculation of the Pearson product-moment correlation coefficient, and subsequent tests of statistical significance, requires the assumption that both variables are (a) linearly related, (b) either interval or ratio, and (c) bivariate normally distributed. Each of these assumptions is briefly discussed in turn.

If, for example, the relationship between the X and Y variable was a perfect quadratic relationship, where $Y = X^2$, the correlation coefficient would still be zero. Thus, the Pearson correlation is a measure of the linear relationship only and does not imply that no other relationship exists between the two variables of interest.

Research by Kenneth Bollen and Kenney Barb suggested that Pearson correlations attenuate the relationship between ordered categorical variables. Therefore, it is assumed that the variables are either interval or ratio, that is, exist on part of a continuous scale. As an alternative, Karl Pearson developed the polychoric correlation coefficient (sometimes denoted as ρ). The polychoric correlation rests on the assumption that the observed categories on an ordinal scale function as proxies for latent continuous normally distributed phenomena and have been shown to be more realistic estimates of relationships between ordered categorical variables.

It is also assumed that the two variables of interest are bivariate normally distributed, as the Pearson correlation is sensitive to skewed distributions and outliers. As an alternative to the Pearson correlation, Charles Spearman (1863–1945) proposed the Spearman’s rank correlation coefficient as a measure of association between ordinal and nonnormally distributed variables. However, where variables are ordinal in nature, Joakim Ekström argues that the polychoric correlation is better suited for statistical inference, especially when original values (such as salary bands and age brackets) have been grouped into categories to form ordered categories.

Limitations

The existence of a strong and statistically significant correlation should always be viewed alongside caution and careful consideration of the data and research context. A researcher should always be aware of the possibility of hidden or intervening variables. To use a classic example, correlational analysis might identify a strong positive and statistically significant correlation between children's foot length and reading ability (N = 90). However, if the children were split into three age-groups (aged 6–8 years, 9–10 years, and 10–12 years, and for each subsample, n = 30), and correlation coefficients were calculated for each group, the researcher might find that no correlation exists at all. This would suggest that the original finding was spurious.

Matthew Gordon Ray Courtney

See also Bootstrapping; Measures of Central Tendency; Measures of Variability; Median Test; Standard Deviation

Further Readings

Bollen, K. A., & Barb, K. H. (1981). Pearson's *r* and coarsely categorized measures. American Sociological Review, 46, 232–239.

Bronowski, J. (1978). The common sense of science. Cambridge, MA: Harvard University Press.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155–159. doi:10.1037/00332909.112.1.155

Ekström, J. (2011). On the relation between the polychoric correlation coefficient and Spearman's rank correlation coefficient. Los Angeles, CA: Department of Statistics, UCLA.

Magnusson, K. (2016). Interpreting correlations: An interactive visualization. Retrieved from http://rpsychologist.com/d3/correlation/

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24(3), 69–71.

Popham, W. J., & Sirotnik, K. A. (1973). Educational statistics: Use and interpretation. New York, NY: Harper & Row.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72–101.

S. Earl Irving S. Earl Irving Irving, S. Earl

Percentile Rank

Percentile rank

1233

1234

Percentile Rank

Percentiles are a cumulative measure to indicate the proportion of an ordered univariate data set that lies below a certain value. Francis Galton advised anthropological travelers that measuring the height of every African man in a study population to calculate the mean might not be well received by the chieftain, so suggested that the travelers line up 1,000 men in order of height and report the height of the middle and quarter points as well as the 20th, 90th, 910th, and 990th person to describe the height characteristics of that population. Thus, he was calling for the 2nd, 9th, 25th, 50th, 75th, 91st, and 99th percentiles to be recorded to describe the population.

Percentiles are most commonly used for descriptive and diagnostic uses such as reporting performance on norm-referenced assessments, by showing the position of a student relative to a group for which the assessment was normed (i.e., the *percentile rank* of the student). A percentile rank should not be confused with *percentage correct*, the number of test items that the student correctly answered. Norm-referenced assessments provide tables showing the percentile associated with a percentage-correct score or a scaled score derived from the percentage-correct score. If we know that a score of 29 equates to a percentile rank of 70, then we know that 70% of students in the reference/norming group obtained a score less than 29. The percentile rank provides a common metric for comparing performances when tests are of differing or unknown length or difficulty. Likewise, in a clinical trial, a clinician may be interested in the dosage of a drug below which 95% of a population would have no side effects. The clinician would estimate the 95th percentile of the trial distribution.

The formula for finding the percentile rank given *N* cases in a data set is

$$\left(\frac{f_b + \frac{1}{2} f_a}{N}\right) \times 100,$$

where f_b is the frequency of cases below the value of interest and f_a is the frequency at the same value as the value of interest.

Attributes

A data set or distribution has the advantage of including every case or score in a class of data. However, it becomes difficult to see the forest for the trees, so a variety of *point measures* have been devised to represent some aspect of the class through the use of a single number. Two attributes of the data are usually sought: location and spread.

Quantiles, which divide a set of ordered data into equal parts in terms of frequency, are examples of such point measures or estimates, and percentiles are one of those quantile measures. Although each quantile is a location attribute, they indicate the way in which the data are spread. In the earlier example, the height at the middle point of the 1,000 men is the middle quantile (50th percentile), also known as the *median*, and the heights at the quarter points (25th, 50th, and 75th percentiles) mark the lower, middle, and upper *quartiles*.

The median is most useful as a measure of central tendency for a data set when the underlying distribution is or skewed (e.g., individual income or housing sale prices), as the mean is sensitive to extreme values (e.g., unusually high incomes or high house prices).

Carl F. Gauss derived a set of quantiles when he tabulated the error function to seven decimal places in determining the accuracy of observations by the function for $\theta(t) = 0.5, 0.6, 0.7, 0.8, 0.8427008, 0.9, 0.99, 0.999, 0.9999$, where the fifth of these values was given when $t = 1$.

$$\theta(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-x^2} \; dx.$$

Here we can see the median (0.5) and a range of other percentiles, including the 90th and 99th percentiles.

For a normally distributed data set, a score that is 1 standard deviation above the mean is at the 84.13th percentile (or the 84th percentile for simplicity). Therefore, 84% of the data lies below this score. The symmetric nature of the standard normal curve means that a score that is 1 standard deviation below the mean is at the 15.87th percentile (or 16th percentile).

Applications

Apart from descriptive and diagnostic purposes, percentiles have had limited use apart from quantile regression and visually in the box-percentile plot. Traditional regression models, which assess how the mean of a distribution varies with changes in a variable of interest, are of limited use when researchers wish to examine the effect at different points of the distribution, as indicated by specific quantiles (e.g., for the 25% most underserved members of a population). Instead, quantile regression models estimate how specified quantiles (or percentiles) of the distribution of the outcome variable vary with the variable of interest.

Unlike the box plot, which has a box of uniform width, the box-percentile plot uses the width of the box to capture information about the distribution across the full range of values. The underlying principle is that the width of the box is proportional to the percentile of a given variate up to the 50th percentile and proportional to 100 minus the percentile for values above the 50th percentile.

S. Earl Irving

See also Box Plot; Descriptive Statistics; Median Test; Quartile

Further Readings

Hald, A. (1998). A history of mathematical statistics from 1750 to 1930. New York, NY: Wiley.

Linn, R. L., & Gronlund, N. E. (2000). Educational tests and measurements (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Rogers, T. B. (1995). The psychological testing enterprise: An introduction. Belmont, CA: Brooks/Cole.

Ella G. Banda Ella G. Banda Banda, Ella G.

April L. Zenisky April L. Zenisky Zenisky, April L.

Performance-Based Assessment Performance-based assessment

1234

1238

Performance-Based Assessment

Performance-based assessment describes an approach to testing using tasks, which differs considerably from traditional assessment formats. Although wide variation exists among the range of tasks that fall under this approach, performance-based testing tasks are characterized by at least one (or more typically, two or more) of the following three elements. First, performance-based assessment often involves high levels of interactivity or engagement with the testing task/item. Second, the result of performance-based assessment is usually the generation of a unique product or performance. The third and perhaps most defining element of performance-based assessment is the extent to which the testing task is contextualized in highly realistic scenarios. For tests composing one or more performance-based tasks, the measurement expectation is that the test takers will demonstrate their proficiency through tasks that are complex in nature and structured to mimic or replicate the real-life situations in which the skills of interest would be used or needed. Some examples for understanding the idea of performance-based assessment at a basic level include the driver’s license road test, the writing and defense of a doctoral dissertation, the carrying out of a science experiment, the creation of an art project, and a musical performance.

Performance-based assessment is predicated on not only *knowing* but also *doing*. Regarding this point, there is another attribute that can set performance-based assessment apart: In addition to the outcome (the product or the performance) that is evaluated, in some performance-based testing contexts, test takers may be evaluated on explaining or otherwise showing the process by which they created the outcome. For some performance-based testing situations, the *how* of the performance may be valued as much as or perhaps even more than the *what*.

A note about terminology: There are a number of terms that have been used over the years to reference the general idea of performance-based assessment as described earlier. Performance-based assessment itself poses a semantic challenge, in that all testing tasks require a performance of some kind in the broadest sense, but the term *performance-based assessment* in the measurement literature has come to mean a specific and narrow view of performance that is marked by tasks that are generally highly open ended and contextualized. Other terms, such as *direct assessment* and *authentic assessment*, have been coined to reinforce the notion of situating testing tasks in real-life situations and the value thereof, but implicit in such terminology is a contrast that has been viewed as unnecessarily divisive in some contexts. Thus, while the terms *direct assessment* and *authentic assessment* are used at times, performance-based assessment provides a more neutral way of referencing these kinds of testing tasks.

Purposes and Uses of Performance-Based Assessments

Any decision about the use of performance tasks requires careful consideration of the assessment purpose within a given testing context. In thinking about the kind of measurement information needed, selected response test item formats are generally most efficient and reliable for measuring factual knowledge and test takers' proficiency in solving well-structured problems. The difference with performance-based assessment, however, is that the measurement information of interest typically involves the application of knowledge or generation of a product that results in a "performance" that is considerably unique to the test taker. The information obtained could involve the test taker's proficiency in finding, evaluating, synthesizing, and using knowledge in real-life settings (e.g., discussing how to develop and test a drug for newly identified bacteria); framing and solving nonroutine problems; producing research findings and solutions in settings that mimic the real-life situations (e.g., designing a city that is close to a lake); recognizing what kind of information matters, why it matters, and how to combine it with other information (e.g., writing an academic thesis); expressing points of view, rationalizing evidence, and displaying originality (speech presentation, debugging a software program, or performing a musical piece); or manipulating objects (e.g., driving, typing, or conducting an experiment). In each of these "performances," the product for evaluation is typically extensive and highly individualized.

and highly individualized.

Performance-based assessment is used within varied contexts. In educational settings, applications have ranged from highly structured, full-scale standardized tests (e.g., the Maryland School Performance Assessment Program) to local assessment initiatives (including tasks such as group projects, written assessments, portfolios, demonstrations, and experiments), where “local” can be understood to function at the level of individual teachers, grades, schools, and/or districts. Many agencies involved in certification and licensure assessment have long incorporated performance tasks on their tests, for example, the Step 3 Examination from the National Board of Medical Examiners, the Uniform CPA Examination from the American Institute of Certified Public Accountants, and the Architectural Registration Examination from the National Council of Architectural Registration Boards. In these professional settings, the appeal of performance-based assessment is rooted in the idea that setting up simulated scenarios appropriate for entry-level professionals and asking candidates to demonstrate their qualification to be credentialed are preferable to isolated answers to abstract questions. A further emerging use of performance-based assessment is in postsecondary testing and admissions, where these kinds of tasks can play a vital role in measuring the higher order skills that are critical to college and career success.

Design and Development of Performance-Based Assessment

Broadly, there are two general formats of performance-based assessments: (1) performance assessments and (2) portfolios or exhibitions. A *performance assessment* is a collection of performance tasks that allows students to be evaluated on both the execution of the process and the final product. A *performance task* is a structured situation in which stimulus materials and a request for information or action are presented to an individual, who generates a response that can be rated for quality using explicit standards. The standards may apply to the final product or to the process of creating it.

Performance tasks can be further divided into two categories depending on the restriction of the performance. The *restricted response task format* comprises short, constructed response items and essays, with narrow and more focused instructions on how to respond (e.g., read aloud, draw a bar graph, and type a letter). The *extended response task format* may require students to use an integration of a variety of skills involving understanding, problem solving, communication, and seeking information beyond what is provided by the task itself. In this case, students may be evaluated on how they carry out the task (e.g., conducting and presenting results of experiments; writing and presenting a research paper).

A portfolio, on the other hand, is a systematic collection of work accumulated over an extended period. Portfolios (e.g., of art projects or computer programs/apps) can be useful in providing evidence for students' progress and an opportunity for them to reflect and assess themselves.

Several methods are used to administer performance-based assessments, although administration of the performance task is in some ways naturally dependent on the construct of interest. In many places, a driver's road test involves the candidate at the wheel of an automobile completing a series of driving tasks at an examiner's request, and it may take place on public roads with other vehicles around or on a specially designed course. Similarly, a performance-based assessment involving musical skills would most likely involve some kind of performance for one or more examiners, in either a private or public setting.

Other performance-based assessments can be developed to be administered using paper and pencil, and these kinds of activities can include short and extended free-response items that can be developed to parallel more time-intensive, hands-on investigations. Such proxies may be less costly to develop and administer, but they may represent a bit of trade-off in terms of fidelity to the real-world scenario of interest. An example of a paper-based proxy is an item in which a student may be asked to describe an experiment they would use to separate salt from a mixture of sand and salt. In their discussion, students would be required to list the materials they would use, the procedure they would use, and the results they would expect.

Hands-on experiments and long-term projects are among the task formats that have been widely used for performance-based assessment, and these can be developed for large-scale testing. However, they can also be costly, time-consuming to develop and administer (some cannot be administered in a single testing session), and difficult to standardize. This has given rise to the growing use of computer-based simulations (especially in large-scale assessment contexts) where the task and the administration conditions must be standardized and scored as objectively as possible. Computer-based simulation tasks have gained popularity as more large-scale assessments are administered on computers because the simulations can be constructed to be lifelike and provide a closer match to hands-on (real-life) performance tasks. Computer simulations also have the potential to be scored more reliably through the use of machine scoring. Computer-based performance tasks can include short quantitative and verbal constructed response items, sets of items structured within scenarios, case-based simulations, problem-solving vignettes, scientific experiments, and information search and analysis.

Psychometric Considerations in Performance-Based Assessment

There are a number of psychometric concerns to be addressed when developing, administering, and using scores from performance-based assessment. These include validity, reliability, the generalizability of test results, and the impact of the test results on classification consistency in cases where the tests are used for such purposes.

A central issue in creating performance-based assessments is their content representativeness, which is directly linked to validity evidence based on content. In this case, it is desirable that the assessments should to a large extent represent or sample adequately the important concepts within the subject matter. There are a number of validity issues that have arisen with regard to performance-based assessments in large-scale assessments. For example, such tasks are not easily standardized in terms of the actual administration itself, leading to limitations in the comparability of results. Also, performance tasks may sample a smaller portion of test takers' performance (due in part to their typically time-consuming nature, it might be possible to administer only a small number of performance tasks within a testing session), raising questions about the generalizability of results to the larger domain of interest.

Another performance-based assessment concern is *task* and *method* heterogeneity. Task heterogeneity is when there are variations in an individual's performance that are dependent on the specific task completed. The issue here is that the relationships between tasks may not hold as the tasks themselves become more different, and this may affect the reliability of the assessment overall. Method heterogeneity occurs when the assessment methods (hands-on, paper and pencil, and computer) affect the comparability of the tasks (and hence, the scores). Another source of variability may be differences among raters in scoring, which can be helped by the development of clear scoring rubrics and appropriate procedures for training and recalibrating scorers.

Ella G. Banda and April L. Zenisky

See also Paper-and-Pencil Assessment; Portfolio Assessment; Rubrics

Further Readings

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. American Psychologist, 48(12), 1210–1218. doi:10.1177/016235329401700403

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. Applied Psychological Measurement, 24, 339–353. doi:10.1177/01466210022031796

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. Applied Psychological Measurement, 24(4), 310–324. doi:10.1177/01466210022031778

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4(4), 289–304. doi:10.1207/s15324818ame0404_3

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18(2), 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., & Bell, R. M. (1998). Analytic versus holistic scoring of science performance tasks. Applied Measurement in Education, 11(2), 121–137. doi:10.1207/s15324818ame1102_1

Lane, S. (2010). Performance assessment: The state of the art. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), Educational measurement (pp. 387–432). Westport, CT: American Council on Education.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13(1), 5–8, 15. doi:10.1111/j.1745-3992.1994.tb00778.x

Marion, S. F., & Buckley, K. (2016). Design and implementation considerations of performance-based and authentic assessments for use in accountability systems. In H. Braun (Ed.), Meeting the challenges to measurement in an era of accountability: NCME applications of educational measurement and assessment book series (pp. 49–76). New York, NY: Routledge.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), Technical issues in large-scale performance assessment (Report No. NCES-96-802). Retrieved from https://nces.ed.gov/pubs/96802.pdf

Pecheone, R. L., & Kahl, S. (2010). Developing performance assessments: Lessons learned. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. Applied Measurement in Education, 4(4), 347–362.

Sireci, S. G., & Zenisky, A. L. (2015). Computerized innovative item formats: Achievement and credentialing. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed.), pp. 313–334). New York, NY: Routledge.

Harrison J. Kell Harrison J. Kell Kell, Harrison J.

Personality Assessment

Personality assessment

1238

1243

Personality Assessment

The *Oxford English Dictionary* offers eight definitions for “personality,” the first appearing in the early 15th century. The research literature offers many definitions, but nearly all of them emphasize two elements: *consistency* and *continuity*. Consistency concerns the regularity with which people think, feel, and act in the same situations. It does not imply that people respond to all situations the same way, but simply that a person typically responds in a similar way when the person is in a similar situation. An individual’s personality may be characterized, for example, by the fact that the individual is consistently outgoing when interacting with coworkers but consistently shy when interacting with strangers. Continuity concerns the enduring nature of these responses. Although personality grows and changes over time, these changes are gradual and most people exhibit stable patterns of behaviors, thoughts, and emotions over long periods of time. This entry describes how conceptualizing personality in terms of consistency and continuity exemplifies the trait-based approach to personality and explains that approach. It explains why the “Big Five” personality traits are a useful means of describing personality and provides a brief summary of each. Next, four major methods for assessing the Big Five are detailed. Finally, the entry outlines conceptions of personality that go beyond the Big Five traits and additional means of assessment that go beyond traditional methods.

Trait Approach to Personality

Traits refer to the consistencies in cognition, affect, and behavior that are the defining features of personality. They can be conceptualized as *tendencies* or *dispositions*. Traits can be defined at different levels of abstraction. For example, a trait can be narrowly defined as “how an individual consistently behaves when interacting with subordinates in the workplace” or broadly defined as “how an individual consistently behaves when interacting with others.” The breadth of a trait’s conceptualization is often dictated by the practical aims intended by the personality assessment designed to measure it. Assessments derived from narrowly defined traits tend to better predict specific outcomes (e.g., coworker ratings of interpersonal skill), whereas assessments derived from broadly defined traits tend to better predict general outcomes (e.g., marital satisfaction).

The theoretical perspective of the researcher also influences how traits are conceptualized. Some interpret them simply as useful ways to describe people. When used in this descriptive sense, traits are *surface* (or *phenotypic*). The surface trait perspective does not make inferences about why people are doing, what they are doing, or how they have come to be the way they are but simply observes them “on the surface.” The *source* (or *genotypic*) trait perspective differs in that personality traits are conceptualized as internal causes of individuals’ affect, cognition, and behavior. A researcher interpreting traits at the phenotypic level would say, “Jenny is kind because she consistently performs many actions that are considered kind, thinks about ways to help others, and feels sympathy when she observes others’ distress.” On the other hand, a researcher interpreting traits at the genotypic level would say, “Jenny consistently performs many actions that are considered kind, thinks about ways to help others, and feels sympathy when she observes others’ distress because she is kind.” In the first case, the trait of kindness is a summary of Jenny’s thoughts, behaviors, and emotions, whereas in the second case, kindness is the internal cause of Jenny’s thoughts, behaviors, and emotions. In many personality assessment situations, especially when questionnaires are used, it is not possible to distinguish whether traits are better construed as surface or source traits, but one’s theoretical orientation can powerfully influence how data are interpreted.

Big Five Personality Traits

Evidence over the past 80 years has converged in finding that individuals can be well described by five broad traits, collectively known as the *Big Five*. The Big Five traits are useful because they can viably be interpreted at the surface or source level, they can be conceptualized at differing levels of abstraction, and they serve as an organizing framework to interpret and develop personality assessment item content. These traits usually do not refer to what individuals do, think, or feel in certain situations but instead how they can be characterized, in general, relative to other people. Further, the Big Five traits are usually not delineated in terms of specific combinations of affect, cognition, or behavior but instead offer overall evaluations of people, because of which adjectives are considered an ideal means for assessing these traits. The following are brief summaries of each of the Big Five traits.

Agreeableness

Adjectives that describe highly agreeable people include considerate, generous, and kind, and highly disagreeable people include selfish, cold, and hostile. Agreeableness scores are positively associated with conflict resolution by negotiation (vs. retaliation), tendency to engage in helping behaviors, and personal popularity. Agreeableness is negatively associated with aggression, prejudice, and competitiveness. Highly agreeable people are motivated to maintain harmonious relations with others, can effectively regulate the frustration that sometimes arises during interpersonal interactions, and experience empathic concern when they observe people in distress.

Conscientiousness

Adjectives that describe highly conscientious people include organized, responsible, and hardworking, and highly unconscientious people include extravagant, careless, and impractical. Conscientiousness scores are positively associated with a wide variety of variables that are important on a practical level, such as longevity, educational attainment, job performance, and marital stability. Conscientiousness is negatively related to criminality, alcohol abuse, smoking, unemployment, and homelessness. Highly conscientious people are able to successfully delay gratification in the service of achieving long-term goals, are likely to follow social norms, and feel guilt and shame when failing to meet others' expectations.

Extroversion

Adjectives that describe highly extroverted people include talkative, assertive, and energetic, and highly introverted people include timid, unadventurous, and

and energetic, and highly introverted people include timid, unadventurous, and inactive. Extroversion scores are positively associated with positive emotionality, numerous mates over the lifetime, and lower mortality rate. Extroversion is negatively associated with feelings of insecurity, depression, and anxiety. Highly extroverted people are biased in favor of attending to positive stimuli, strive for interdependence and intimacy, and tend to create positive social environments in the course of their interactions with others.

Neuroticism

Adjectives that describe highly neurotic people include emotional, nervous, and tense, and highly emotionally stable people include calm, relaxed, and contented. Neuroticism scores are positively associated with cardiovascular disease, alcohol abuse, criminal arrest rates, and a wide variety of forms of psychopathology (e.g., depression, eating disorders, and schizophrenia). Neuroticism is negatively associated with self-efficacy, subjective well-being, job satisfaction, and relationship satisfaction. Highly neurotic people tend to feel self-conscious and insecure, are prone to experiencing minor frustrations as emotionally overwhelming, and can act impulsively when upset.

Openness to Experience

Adjectives that describe people scoring high on openness include imaginative, creative, and curious, and people scoring low on openness include unsophisticated, unreflective, and shallow. Openness is positively associated with higher scores on intelligence tests, preferences for art, divergent thinking, and political liberalism. Openness is negatively associated with authoritarianism, racism, prejudice, and religiosity. Highly open people are competent in recognizing others’ emotions, seek out novelty and originality, and are primarily romantically attracted to other highly open people.

Assessment Techniques

The variety of methods available to assess personality is enormous. The majority of these methods can be categorized into four groups according to the information gathered: self-report data, observer data, life history data, and test data.

Self-Report Data

Inventories consisting of self-report items are the most commonly used form of personality assessment. These inventories can be roughly divided into two types: *objective* and *projective*. When respondents take objective surveys, they focus on making attributions about themselves, usually by rating the extent to which adjectives or short statements (e.g., “I like going to parties”) accurately describe themselves on a Likert-type scale (e.g., 1 = *very inaccurately,* 7 = *very accurately*). Objective items are often decontextualized and ask people to think about themselves “in general,” but questionnaires can be developed that specify concrete situations (e.g., “I keep to a tight schedule when I am at work”). Projective tests are commonly found in clinical settings and usually require participants to interpret an ambiguous stimulus; the classic example of a projective test item is a Rorschach inkblot. A test user makes inferences about a test taker’s personality based on the test taker’s interpretations of the projective items.

There are many benefits to self-report personality assessments (especially objective ones), the majority having to do with efficiency: They are quick and easy to administer and inexpensive to score, and participants can complete a relatively large number of items without becoming fatigued. Self-report items are also useful in that they allow respondents to report on their own cognition and emotions, which ultimately only they have complete access to. There are also drawbacks to self-report surveys, the primary one being that participants can easily distort their responses. People may purposefully provide untruthful answers if an important outcome, such as being admitted to an educational institution or selected for a job, is contingent on those answers. Even in low-stakes settings, some participants may reply untruthfully if they believe their responses will make them appear in a negative light. A potential remedy to such practices is to use disguised items that are worded so that respondents have difficulty discerning the answer that will make them appear in the best light; items of this sort (e.g., “My head is often callous”) often appear in clinical questionnaires such as the Minnesota Multiphasic Personality Inventory. Unfortunately, outside clinical settings, disguised objective items (along with projective ones) are poor predictors of practical outcomes; the most predictive items tend to be phrased in a straightforward, unambiguous way.

Observer Data

Many self-report items can be adapted so they can be completed by observers to evaluate a target individual’s personality traits. Oftentimes, these observers are people who are acquainted with the target individual, such as spouses, friends, coworkers, or family members. These observers are often asked to evaluate the target’s personality traits in general and over an unspecified period of time. Observers can also consist of people unknown to the target whose personality is being evaluated; this “zero-acquaintance” approach is often confined to the controlled conditions of laboratories. In these cases, individuals whose personalities are being evaluated perform identical tasks (e.g., reading a weather report, interacting with a confederate), and observations and ratings are conducted through a two-way mirror or by viewing videotapes of the target individuals. In addition, observational inferences about personality can be made during or from interviews, whether those interviews are explicitly designed to assess personality or not (e.g., job interviews). Regardless of the specific observational technique (and perhaps surprisingly), even strangers often moderately agree in their assessments of target individuals’ personalities; agreement is even higher among individuals who know the target person well.

Observer ratings of personality can predict the real-world outcomes (e.g., academic achievement and job performance) better than self-reports. Unfortunately, observer ratings are time consuming and often expensive to gather, precluding their use in many circumstances. In addition, some traits are more amenable to external observation than others. When the observers do not know the target well, agreement about traits more strongly defined in terms of manifest behavior (e.g., extroversion and openness) is highest. Only target individuals have access to their own thoughts and emotions; observers can often make strong inferences about them through verbal or physical behavior, but ultimately only the individual actually experiencing internal psychological phenomena can report on them. For example, Jane and John may keep to themselves at a social gathering; Jane may do so because she is shy, and John because he is anxious. To a stranger, Jane and John’s manifest behavior would appear similar, but their cognition and affect—important aspects of their personalities—would be quite different.

Life History Data

Life history data are diverse and generally consist of records or reports of concrete activities or events. Relevant records might include school transcripts, résumés, and employee personnel files (e.g., absences, disciplinary actions, and resumes, and employee personnel files (e.g., absences, disciplinary actions, and promotion decisions). Biographical data such as community service experience are informally solicited in many educational admission procedures, but “biodata” inventories consist of standardized questions that elicit specific information about such activities (e.g., “How many times have you volunteered in your community over the past year?”). As biodata items are self-reported, they can be faked, but because the truthfulness of the answers can be verified, the presumption is that participants are less likely to distort their responses to biodata inventories than to traditional self-report questionnaires.

Test Data

Test data require individuals to actually demonstrate some aspect of a personality trait. This “performance” of the trait can be evaluated according to some external, consensually defined standard. Perhaps because of conceptual, practical, and even ethical difficulties, test-based evaluations of personality traits are currently relatively rare, although they enjoyed some prominence in the mid-20th century.

Because of the strong association between openness to experience and higher intelligence, some researchers consider measures of intelligence to also be tests of personality. A performance test of conscientiousness might consist of asking individuals to solve as many of a practically infinite (e.g., 10,000) number of simple math problems as they can, given no time limit. As conscientiousness is implicated in persistence and self-control, those who are more conscientious would be predicted to solve more problems before giving up than those less conscientious. A performance test of agreeableness might consist of asking individuals to behave in as friendly a manner as possible during a simulated job interview. A performance test of emotional stability might consist of exposing individuals to progressively more disturbing stimuli (e.g., abrasive noises and videos of traumatic events) until they refuse to continue.

Tests of personality traits are attractive in that they cannot be faked and they offer external criteria for the evaluation of their results. In some cases, however (as in the agreeableness test previously mentioned), this external criterion might consist merely of observers’ ratings, making it difficult to distinguish test data and observer data. In tasks where individuals are asked to act out trait-relevant behavior to the maximum degree they are capable of, verification that they are truly exerting maximal effort ultimately depends on self-report. Finally, if personality traits are defined in terms of *typical* cognition, affect, and behavior, but some tests ask individuals to perform at the upper limit of aspects of their personalities, it could be argued that such methods are not truly the measures of personality traits.

Beyond the Trait-Based Approach

The trait-based approach, exemplified by the Big Five, provides a strong foundation for conceptualizing personality, constructing personality assessments, and interpreting the results of those assessments. Nonetheless, and as would be expected given a topic as complex and intimate as human personality, ideas are constantly evolving, and there are many theoretical perspectives, some of which favor a different level of analysis than broad traits. Indeed, even some Big Five enthusiasts now advocate that a sixth trait—humility-honesty—be measured distinctly.

Prominent alternatives to the traditional trait-based approach are *social cognitive* theories of personality. This family of theories emphasizes "persons in context" (rather than "persons in general"). Social cognitive personality assessments focus on identifying the cognitive mechanisms (e.g., goals, knowledge, and self-efficacy beliefs) that cause people to behave differently in different situations. For example, one important social cognitive variable is *situation perception* (or *encoding*), which embodies the idea that different individuals can be in normatively identical situations but view those situations differently, in turn leading to different behaviors. For example, one person might view being unexpectedly given a large project at work as an opportunity, while another might view it as a burden, causing the first person to do his best, the second to exert minimal effort to finish the task as quickly as possible.

Other perspectives on personality focus on how individuals give meaning to their lives in terms of the personal stories they develop about themselves or personal projects they undertake. Yet, other theories emphasize how individuals relate to their cultural milieu and how they view the roles they play in their family, work, and social lives. Compared to typical self-report questionnaires, these theories demand different assessment approaches, such as extended responses to open-ended questions about important episodes in individuals' lives. The data these methods generate have a richness that responses to typical trait-based items cannot match, but they present challenges in terms of the subjectivity of interpretation in identifying consistency across responses and the sheer magnitude of the coding task itself. Care must be taken to ensure that the process of reducing these complex data into manageable elements does not rob them of their nuanced psychological insights, which constitute the very advantage they have over responses to trait-based questionnaires.

Human personality is complex and can validly support many different theories and perspectives. To some extent, the approach chosen by investigators will be dictated by the practical demands of the situation and their own preferences—which of course are reflections of personality itself.

Harrison J. Kell

See also Intelligence Quotient; Intelligence Tests; Likert Scaling; Minnesota Multiphasic Personality Inventory; Projective Tests; Rating Scales; Self-Report Inventories; Social Cognitive Theory

Further Readings

Bornstein, R. F. (2007). Toward a process-based framework for classifying personality tests: Comment on Meyer and Kurtz (2006). Journal of Personality Assessment, 89, 202–207. doi:10.1080/00223890701518776

Boyle, G. J., Matthews, G., & Saklofske, D. H. (Eds.). (2008). Personality theory and assessment (Vols. 1 & 2). London, UK: SAGE.

Buss, A. H., & Finn, S. E. (1987). Classification of personality traits. Journal of Personality and Social Psychology, 52, 432–444. doi:10.1037/0022-3514.52.2.432

Campbell, D. T. (1957). A typology of tests, projective and otherwise. Journal of Consulting Psychology, 21, 207–210.

Cervone, D., Shadel, W. G., & Jencius, S. (2001). Social-cognitive theory of personality assessment. Personality and Social Psychology Review, 5, 33–51. doi:10.1207/S15327957PSPR0501_3

Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). New York, NY: HarperCollins.

Fiske, D. W. (1978). Strategies for personality research. San Francisco, CA: Jossey-Bass.

Fiske, D. W., & Butler, J. M. (1963). The experimental conditions for measuring individual differences. Educational and Psychological Measurement, 23, 249–266. doi:10.1177/001316446302300203

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84–96. doi:10.1016/j.jrp.2005.08.007

Hogan, R., Johnson, J. A., & Briggs, S. R. (Eds.). (1997). Handbook of personality psychology. San Diego, CA: Academic Press.

John, O. P., Robins, R. W., & Pervin, L. A. (Eds.). (2008). Handbook of personality: Theory and research (3rd ed.). New York, NY: Guilford.

Leary, M. R., & Hoyle, R. H. (Eds.). (2009). Handbook of individual differences in social behavior. New York, NY: Guilford.

Mayer, J. D. (2004). A classification system for the data of personality psychology and adjoining fields. Review of General Psychology, 8, 208–219. doi:10.1037/1089-2680.8.3.208

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: Fundamental principles for an integrative science of personality. American Psychologist, 61, 204–217. doi:10.1037/0003-066X.61.3.204

Meehl, P. E. (1986). Trait language and behaviorese. In T. Thompson & M. D. Zeiler (Eds.), Analysis and integration of behavioral units (pp. 315–334). Hillsdale, NJ: Erlbaum.

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C., & Duncan, L. E. (1998). Traits and motives: Toward an integration of two traditions in personality research. Psychological Review, 105, 230–250. doi:10.1037/0033-295X.105.2.230

Ludmila N. Praslova Ludmila N. Praslova Praslova, Ludmila N.

Personnel Evaluation

Personnel evaluation

1243

1246

Personnel Evaluation

Personnel evaluation in an important component of the performance management process in organizations. Also referred to as a *performance appraisal*, it provides developmental feedback for increasing competence, enhancing performance, and making personnel decisions such as distributing rewards. These activities are expected to contribute to an overall improvement in organizational effectiveness. Because of its perceived importance, research on evaluating personnel performance has been a major focus of organizational scholars and practitioners since the first half of the 20th century. More recently, however, the validity of personnel evaluations and the overall effectiveness of traditional approaches to them have been challenged, and alternative approaches to performance management have been proposed. This entry describes the key approaches to personnel performance evaluation, challenges to ensuring the effectiveness of the evaluation process, and developments aimed at improving performance management in organizations.

Approaches to Evaluation Systems

Most organizations have formal evaluation systems to assess employee performance. Such practices are typically focused on the assessment of an employee’s contributions during a specified period of time, usually a year. Unfortunately, annual evaluations are often dreaded by both managers and employees, and their accuracy, usefulness, and fairness are often challenged within companies and in litigation. Both the technical aspects of ensuring rating accuracy and the larger issues of defining “performance” and designing systems that support fairness are seen as essential for effective personnel evaluation systems.

Measurement Accuracy

The traditional approach to performance evaluation focuses on measurement accuracy (the psychometric quality of evaluations resulting from performance ratings). Two key approaches to improving measurement accuracy are developing better rating instruments and rater training.

Rating Instruments

A good rating instrument should reliably measure performance and avoid biases such as halo, leniency, range restriction, and severity. The most notable instruments for increasing objectivity are behaviorally based, requiring raters to evaluate the frequency or the effectiveness of specific employee actions as related to job performance. Examples of behaviorally based instruments include *Behavioral Observation Scales*, which call for rating the frequency of specific employee actions, and *Behaviorally Anchored Rating Scales*, in which observed employee actions are matched to examples associated with meeting or not meeting specific criteria (behavioral anchors).

This focus on employee behaviors is thought to safeguard against subjectivity, which is more likely to be associated with trait-based instruments. Such instruments require raters to evaluate employees on characteristics such as leadership, initiative, or creativity. While trait-based ratings have long been popular and still appear on the evaluation forms of many organizations, most researchers see them as imprecise and as philosophically incompatible with the evaluation of personnel performance, defined as on-the-job actions. For example, an individual who may generally be characterized as high on the trait of shyness might repeatedly step into a leadership role in the workplace when necessary. A behaviorally based evaluation is more likely to reflect this individual’s leadership actions than a trait-based evaluation. The more visible link between behaviorally based instruments and job performance also makes approaches such as Behavioral Observation Scales and Behaviorally Anchored Rating Scales more legally justifiable. Trait-based scales require evaluators to make inferences about abstract traits, and such inferences are less defensible against legal challenges.

Comparative ratings, such as forced-distribution formats or ranking scales, require the rater to make relative evaluations. Some researchers have argued that it is easier for raters to evaluate performance in comparative formats because comparison is a natural process of social cognition, and employers may believe that comparisons are more effective for driving performance, especially when rewards and punishments are tied to them. The most extreme form of this approach is the “rank-and-yank” method, in which a predetermined percentage of the lowest ranked employees (e.g., the bottom 15%) is fired.

There are multiple concerns with comparative instruments. As with all systems based on distribution curves, they may exaggerate otherwise negligible differences in employee performance. Performance in organizations is almost never normally distributed, and “forcing” a curve can lead to unintended consequences. In a group of high performers, even the lowest ranked employees may meet absolute standards of performance; it is also possible that in a lower performing group, even the top-ranked employees do not meet absolute standards. Moreover, forced-distribution systems often hinder collaboration and create cutthroat environments, of which the culture at Enron before its 2001 bankruptcy might be the most infamous example. Such systems can also be difficult to justify in a legal challenge. Finally, comparative assessments do not provide the level of detail necessary for employee feedback and development.

At the same time, criterion-based instruments such as Behaviorally Anchored Rating Scales are not a guarantee of useful ratings. Developing appropriate criteria is difficult, as evidenced by the voluminous literature on the “criterion problem.” It is vital to ensure that the criteria used in performance evaluations are relevant to the requirements of a specific job; if an organization uses the same evaluation form for all or most jobs, it is difficult to for such an instrument to be valid, useful, and legally defensible.

Rater Training

As in many other areas, “user error” may negate the advantages of even the best rating instruments. Thus, rater training has been an important focus in personnel performance evaluation. The three most popular approaches are rater error training, behavioral observation training, and frame-of-reference training.

The goal of rater error training is to ameliorate the potential effects of common rating biases such as halo error (rating employees based on general impressions; in a positive halo, a generally liked employee will receive high evaluations overall, even if performance in some of the areas is lacking), leniency error (high ratings across employees), severity error (a tendency to assign low ratings across employees), and range restriction error (concentrating ratings on a narrow portion of the scale). Discussion of these topics is thought to reduce their effects on the rating process, and such training has indeed been shown to reduce rating errors. Ironically however, it also has the potential of lowering rating validity because of its assumption that performance is normally distributed—which, as mentioned previously, is rarely the case. If all employees perform well and deserve high evaluations, avoiding uniformly high ratings could in fact inadvertently reduce their validity.

The goal of behavioral observation training is to develop the rater's ability to observe and accurately evaluate employee performance by introducing metacognitive strategies to direct attention to the relevant aspects of performance and associated behaviors. Although research is limited, the available studies suggest that behavioral observation training is effective in improving rating accuracy.

Frame-of-reference training aims to ensure that raters formulate accurate impressions of employee performance. This is achieved by calibrating rater judgments to develop agreement on the relevance of specific behaviors to specific performance dimensions, the level that is most effective for specific behaviors, and the rules for combining separate observations and judgments into summary evaluations for specific performance dimensions. Research supports the effectiveness of frame-of-reference training for increasing rating accuracy.

Challenges to Personnel Evaluation

Although much of the traditional literature on personnel evaluation focuses on measurement accuracy, several alternative views focusing on the larger context of evaluation have been developed. One prominent alternative view states that the measurement-focused or "test" metaphor is based on flawed assumptions. In their seminal 1992 work, Robert Folger, Mary Konovsky, and Russell Cropanzano described these flawed assumptions as the belief that work arrangements allow for reliable and valid measurement, raters will assess performance accurately, and a rational, unitary criterion exists. In reality, the nature of employment is increasingly such that many work behaviors cannot be observed, raters lack not only the ability to accurately assess performance but also the motivation to do so, and individuals function in the world of elusive, often politically negotiated rather than objective, criteria. Organizational realities might be better described by a more “political” metaphor—performance evaluations are manipulated by managers to suit political agendas and rarely reflect “true” performance.

Folger, Konovsky, and Cropanzano proposed an alternative “due process” metaphor, which stresses procedural fairness as the way to improve accuracy and address the shortcomings of the test and political metaphors. The due process model includes adequate notice (criteria communicated clearly and in advance), fair hearing (including employee input in evaluation), and judgment based on evidence. It stresses employee rights and the need for proper channels of dispute resolution. Other broad conceptualizations of rating effectiveness also stress the need to consider the social context of personnel performance evaluations and the role of employee reactions. Thus, an effective rating might be one that employees perceive as fair and that motivates them in intended ways.

One popular way to improve the effectiveness of performance evaluations is the multisource or 360° system, in which evaluations are performed by managers, the employees themselves, and peers or customers, as relevant. Such systems are perceived as fairer than single-rater evaluations, although they are not a panacea. While managerial ratings have the potential for error, so do self-evaluations and customer evaluations. Nevertheless, multisource systems, at the very minimum, are likely to result in the discussion of discrepancies between ratings, which may lead to helpful insights about employee performance. More detailed and multidimensional feedback is also more likely to be useful for employee development.

Another way to increase the effectiveness of evaluations is to increase their frequency. If evaluation is closer in time to behavior, individuals are more likely to remember the behavior; more frequent evaluation is also more likely to lead to timely improvement. A variation of this approach is the system of semiannual evaluation of managers at Google by their employees, which has resulted in improved employee ratings of managerial performance.

The most radical approach to changing performance evaluation is replacing annual performance evaluations altogether in favor of regular developmental feedback or simple and brief monthly or quarterly check-ins focused on goal setting and development. Microsoft, Deloitte, Accenture, Adobe, Gap, and other companies are experimenting with doing away with traditional systems, which are seen as expensive, insufficiently accurate, and even harmful to morale, teamwork, and creativity. Instead, companies are introducing more frequent and less formal multisource feedback and coaching, and some are testing the idea of not using any performance ratings.

Future Directions

Personnel performance evaluation is a well-established and well-researched field. Yet, it is currently undergoing a significant shift in focus from approaches rooted in early 20th century "command-and-control" management styles involving annual, high-stakes, manager-determined performance ratings to real-time, developmentally focused, and multisource feedback. The latter is seen as more effective in modern organizations, which likely involve performance that is team based, highly cognitive in nature, and in need of constant adaptation to rapidly changing contexts. Many traditional systems have the potential to hurt this type of performance by pitting employees against each other and directing employee attention from solving organizational problems to "gaming the system" to achieve desired ratings. Because the transition to more flexible and developmentally focused approaches is relatively new and ongoing, it will be important to evaluate the success of these newer models in the coming years.

Ludmila N. Praslova

See also Formative Assessment; Formative Evaluation; Paradigm Shift; Reliability; Summative Assessment; Validity

Further Readings

Bock, L. (2015). Work rules! Insights from inside Google that will transform how you live and lead. New York, NY: Hachette Book Group.

Burkus, D. (2016). Under new management: How leading organizations are upending business as usual. New York, NY: Houghton Mifflin Harcourt.

Folger, R., Konovsky, M. A., & Cropanzano, R. (1992). A due process metaphor for performance appraisal. In B. M. Staw & L. L. Cummings (Eds.), Research in organizational behavior (Vol. 13, pp. 129–177). Greenwhich, CT: JAI

Press.

Harris, M. M., Ispas, D., & Schmidt, G. F. (2008). Inaccurate performance ratings are a reflection of larger organizational issues. Industrial and Organizational Psychology, 1, 190–193. doi:10.1111/j.1754-9434.2008.00037.x

Hekman, D. R., Aquino, K., Owens, B. P., Mitchell, T. R., Schilpzand, P., & Leavitt, K. (2010). An examination of whether and how racial and gender biases influence customer satisfaction. Academy of Management Journal, 53, 238–264. doi:10.5465/AMJ.2010.49388763

Kline, T. B., & Sulsky, L. M. (2009). Measurement and assessment issues in performance appraisal. Canadian Psychology/Psychologie Canadienne, 50(3), 161–171. doi:10.1037/a0015668

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. Journal of Management, 6, 881–905. doi:10.1016/j.jm.2004.06.005

London, M. (2003). Job feedback: Giving, seeking, and using feedback for performance improvement (2nd ed.). Mahwah, NJ: Erlbaum.

Murphy, K. R. (2008a). Explaining the weak relationship between performance and ratings of job performance. Industrial and Organizational Psychology, 1, 148–160. doi:10.1111/j.1754-9434.2008.00030.x

Murphy, K. R. (2008b). Perspectives on the relationship between job performance and ratings of job performance. Industrial and Organizational Psychology, 1, 197–205. doi:10.1111/j.1754-9434.2008.00039.x

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. Industrial and

Organizational Psychology, 1, 174–179. doi:10.1111/j.1754-9434.2008.00033.x

Reb, J., & Greguras, G. J. (2008). Dynamic performance and the performance-performance rating relation. Industrial and Organizational Psychology, 1, 194–196. doi:10.1111/j.1754-9434.2008.00038.x

Tubre, T., Arthur, W., Jr., & Bennett, W., Jr. (2006). General models of job performance: Theory and practice. In W. Bennett, Jr., C. Lance, & D. Woehr (Eds.), Performance measurement: Current perspectives and future challenges. Mahwah, NJ: Erlbaum.

Nicholas J. Shudak Nicholas J. Shudak Shudak, Nicholas J.

Phenomenology

Phenomenology

1246

1249

Phenomenology

This entry provides a general overview of phenomenology as a methodological approach for conducting educational research. Etymologically speaking, phenomenology is a compound of the Greek *phainomenon* (the thing that appears or is seen) with *logos* (study). Phenomenology can be roughly translated to mean the science or study of things as they appear to us. Phenomenology, broadly defined for the purpose of encompassing the varied phenomenologies, is the study of or inquiry into how a person’s conscious experience with things—with phenomena—provides deeper and more truthful understandings of those things and ultimately of the self and of the world. The basic unit of analysis in phenomenology is phenomena, not people.

As a research methodology, phenomenology is deeply indebted and closely connected to the disciplinary field of philosophical inquiry that goes by the same name and that was largely developed during the 20th century. As the century came to a close, and in the early decades of the 21st century, there was interest in distinguishing phenomenology as a way of doing philosophy from phenomenology as a research methodology by referring to the latter as phenomenography. For the purposes of this entry, the philosophy and methodology are inextricably bound, and phenomenology is the preferred term in reference to each.

What follows is a tracing of the philosophical roots of phenomenology, descriptions of key terms, the identification of overarching assumptions guiding phenomenological research, and criticism of phenomenology as a research method.

Philosophical Roots of Phenomenology

Philosophically speaking, *phenomenology* emerged from Germany during the late 19th and early 20th centuries. And though the term is found in the works of German philosophers such as Immanuel Kant (1724–1804), Johann Gottlieb Fichte (1762–1814), and G. W. F. Hegel (1770–1831), it is through the work of Franz Clemens Brentano (1838–1917) and his pioneering work developing a science of descriptive psychology that phenomenology finds its beginnings. Brentano had an interest in creating and ushering in a new scientific form of psychology predicated on empirical and accurate descriptions of those things—phenomena—as they appear to and through consciousness. Considering that phenomenology begins in the descriptive psychology of Brentano, its development and refinement as a new and unique philosophical discipline is attributed to two 20th-century German philosophers, Edmund Husserl (1859–1938) and Martin Heidegger (1889–1976).

It has been said that all of Western philosophy is a series of footnotes leading back to Plato. The same could be said of phenomenology and Husserl. Husserl is largely characterized as the founder of phenomenology, the progenitor from which many intellectual lineages are traced. As the founder, Husserl saw himself creating an entirely new and even radical way of doing philosophy, a novel approach to knowing the essential nature of things—phenomena. What made this approach so radical was that through Husserl’s new philosophy, objective knowledge of things—objectivity—can be gained through an investigation of the qualitative characteristics of our conscious experience of things—subjectivity. That objectivity was to be found through and even dependent upon subjectivity was a distinct departure from the norms of the more traditional philosophical approaches to knowing. And though phenomenology is many and varied, such that there are multiple phenomenologies, the notion of finding objectivity through subjective experience is familiar to all.

To know a thing, a phenomenon, Husserl suggested that we return to the things themselves through a phenomenological investigation. Conducting such an investigation is no easy task. On Husserl’s terms, a phenomenological investigation requires a person to suspend, bracket, or reduce the person’s preconceived notions that might lead to a distortion of the thing itself; it requires a decontextualizing and emptying of oneself in order to return to the thing being investigated. A phenomenological investigation, then, is a return to a thing by returning to one’s intimate, conscious, sensory, and lived experiences with that thing. It is a way of knowing by analyzing the composite “first-person” lived experiences with regard to how the thing itself appears to the person.

Knowing through this kind of investigation requires one to be open to how a phenomenon emerges or appears or captures our consciousness and not how our consciousness captures the thing. The difference is found in the role that one’s worldly context, or lifeworld, plays in coming to know. In regard to the former, and as Husserl believed, the “essence” or the pure characteristic of a thing is “seen” and thus known in a nonprejudiced and “presuppositionless” way that in turn structures our consciousness. It is the thing itself as it appears through one’s conscious experience of it that informs an understanding of one’s larger and encapsulating lifeworld. Whereas in the latter, one’s context is necessary for a deep and full understanding of the thing being investigated. Essentially the difference between the former and the latter is the difference between the importance placed on consciousness and context. It is also the difference between Husserl and Heidegger.

For a brief time, Heidegger was a former student and assistant to Husserl while they were both at the University of Freiburg in Germany. For Heidegger, phenomenological investigations require that one’s context be taken into consideration when attempting to study phenomena. To this extent, there is no need to bracket out the world, as a presuppositionless stance is impossible when investigating phenomena. For Heidegger, phenomenology is less an investigation into our conscious experiences with things for the sake of knowing the essence of that thing, as it is an investigation of and an analysis into our conscious and yet highly contextualized and interactive experiences that bring forth things for the sake of knowing the human being. Heidegger’s is a philosophical approach to studying what it means to “be” or to exist as a human being through an interpretation of our experiences and of our relationships with those things with which we consciously interact.

Key Phenomenological Terms

Phenomenology as a disciplinary field and as a research methodology relies on a few key and distinguishing terms. *Phenomena* are those things that make themselves known to us through the conscious experience of those things. *Intentionality* is a reference to being conscious of the connectedness between things, especially to humans as subjects and the objects that are experienced. The *essence* of a thing is the immutable quality of that thing that makes it one thing and not something else.

Intuition is a form of knowing a thing when that thing appears to consciousness in a way that its essence is fully known through that experience. A person's *lifeworld* is that worldly context in which that person lives and experiences things, the context in which meaning is made or given. The *phenomenological reduction* is a step taken by phenomenologists to reduce their own preconceived notions of a particular phenomenon when investigating that phenomenon. By reducing, bracketing, or bridling oneself, a person is really opening themselves to the possibility that the phenomenon appears unencumbered.

Guiding Assumptions of Phenomenology as Research Methodology

As a research methodology, phenomenology is of the qualitative type, and as such, it places primary importance on the systematic study of subjective experience. In this regard, phenomenological research is an effort at making sense of the world and ourselves as we experience the world and its objects. Phenomenological research proceeds through an objective and methodological analysis of how people experience things as those things appear, manifest, or make themselves known, whether those things are other people, animals, objects, events, or ideas.

And though phenomenologists seek to conduct research objectively, phenomenology cannot be reduced to one approach or a singular method for analyzing and interpreting experience. What makes phenomenology unique is that it resists methodological codification, while at the same time adherents rely on common approaches to procure and analyze data. Some of those approaches include crafting research questions rooted in experience, the use of interviews to mine experience and provide data, efforts at bracketing, identifying units of meaning through transcription, clustering the units in relation to the research question, and the identification of recurring themes to help make sense of the experience of a thing.

For the phenomenologist, there is always something going on, there is always some phenomenon showing itself in our daily lives. Phenomenology, then, is an approach to researching and understanding more deeply our everyday lived experiences, a method of taking what is commonplace and ordinary and looking at it so that it becomes new, unique, and extraordinary. Phenomenological research that focuses on mining lived experiences results in a more truthful way of being in and with the world. In that regard, phenomenology as both a form of philosophical inquiry and a research methodology is largely a careful reflective and descriptive endeavor.

Criticisms

As a research methodology, there are a few criticisms levied against phenomenology. An enduring criticism of phenomenology is that because it does not proceed from an experimental base (for instance, it lacks a hypothesis, variables, and replicability), it cannot be considered scientific regardless of the phenomenologists' claims of objectivity. Another concern is that because phenomenology is rooted in lived experience, and because the data are usually procured through interviews, the number of subject participants studied will invariably be limited due to the inordinate amount of time it takes to analyze the data. Because of this, another criticism is that the research results can hardly be considered generalizable.

Nicholas J. Shudak

See also Educational Research, History of; Epistemologies, Teacher and Student; Narrative Research; Nonexperimental Design; Objectivity; Postpositivism; Transcription

Further Readings

Giorgi, A. (1997). The theory, practice, and evaluation of the phenomenological method as a qualitative research procedure. Journal of Phenomenological Psychology, 28(2), 235–260.

Heidegger, M. (2008). Being and time. New York, NY: HarperPerennial Modern Classics. (Original work published in 1927) Husserl, E. (2001). Logical investigations (Vols. 1 & 2). New York, NY: Routledge.

Hycner, R. H. (1985). Some guidelines for the phenomenological analysis of interview data. Human Studies, 8, 279–303.

Moran, D. (2000). Introduction to phenomenology. New York, NY: Routledge.

Moran, D., & Mooney, T. (2002). The phenomenology reader. New York, NY: Routledge.

Moustakas, C. (1994). Phenomenological research methods. Thousand Oaks, CA: SAGE.

Sokolowski, R. (2000). Introduction to phenomenology. New York, NY: Cambridge University Press.

Vagle, M. D. (2014). Crafting phenomenological research. Walnut Creek, CA: Left Coast Press.

van Manen, M. (2001). Researching lived experience: Human science for an action sensitive pedagogy. London, Canada: Althouse Press.

van Manen, M. (2014). Phenomenology of practice: Meaning-giving methods in phenomenological research and writing. Thousand Oaks, CA: Left Coast Press.

James Dean Brown James Dean Brown Brown, James Dean

Phi Coefficient (in Generalizability Theory) Phi coefficient (in generalizability theory)

1249

1251

Phi Coefficient (in Generalizability Theory)

The ϕ coefficient in generalizability theory—not to be confused with the ϕ correlation coefficient used to estimate the degree of association between dichotomous categorical variables or the ϕ (λ) statistic used to estimate the dependability of test scores at various cut points—is one of two coefficients used in generalizability theory (or G theory) to estimate score dependability (which is analogous to score reliability in classical test theory). The first coefficient, which is often called the *generalizability coefficient* (or G coefficient), is used to estimate the dependability of scores for tests designed for relative decisions (also known as norm-referenced decisions like those typically made with standardized tests). The second, which is often called the ϕ *coefficient*, is used to estimate the dependability of scores on a test for absolute decisions (also known as criterion-referenced decisions like those typically made with classroom tests). This entry explains how the ϕ coefficient is calculated and how it can be used to improve the dependability of a test.

Both the G and ϕ coefficients provide an estimate of the overall dependability of the scores or as it is expressed in G theory: the proportion of universe score variance. Such dependability estimates are calculated using the following general equation:

$$\text{Dependability} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2}.$$

In this case, is the estimated persons variance component, and is the estimated error variance (all variance components discussed in this entry are derived from specially adapted analysis of variance procedures—steps that are beyond the scope of this entry). Then, the dependability estimate is the ratio of estimated persons variance to the estimated persons variance plus estimated error variance .

The ϕ coefficient (also known as Φ, or the dependability coefficient for absolute decisions) in particular is used to estimate the overall dependability, or proportion of universe score variance, of a set of scores used for absolute (or criterion referenced) decisions. ϕ is interpreted on a .00 to 1.00 scale, where .00 indicates zero dependability (or zero universe score variance) and 1.00 represents 100% dependability (or 100% universe score variance).

ϕ coefficient is calculated using the G theory equation that follows:

$$\Phi(\Delta) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\Delta)}.$$

Here, Φ(Δ) is the ϕ dependability estimate for absolute error is the estimated persons variance component, and is the estimated error variance for absolute (or criterion referenced) decisions. Then, the dependability estimate is the ratio of estimated persons variance to the estimated persons variance plus absolute error variance .

Consider a situation in which a tester wants to study the relative effects of three potential sources of error (called facets)—including persons (p), raters (r), and rating categories (c)—and the four possible interactions of those facets pr, pc, rc, and prc. Such a study could have included other facets like composition topics (e.g., two different topics), rating occasions (i.e., raters doing the scoring two different times), rater types (e.g., teachers vs. naive raters), and so forth. Based on variance components for each facet and their interactions, it is assumed in G theory that all facets (except persons) and their interactions with each other and persons can contribute to error in absolute decisions. Thus, absolute error for this example is defined as

$$\hat{\sigma}_e^2(\Delta) = \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}.$$

Note that the various *n* values in the denominators of the components making up the error are used to account for what happens to dependability when there are various numbers of raters and categories.

Placing in the general equation in lieu of , the equation for the ϕ coefficient for absolute decisions becomes

$$\Phi(\Delta) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2(\Delta)}$$

$$= \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_{pc}^2}{n_c} + \frac{\hat{\sigma}_{rc}^2}{n_r n_c} + \frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}}.$$

For example, let’s say that the estimated variance components for and turn out to be 2.95, .26, .70, .30, .50, .23, and 1.91, respectively. Then, the ϕ coefficient for absolute error for the test using two raters (n_r = 2) and five categories (n_c = 5) would be .80076 (or about .80) as follows:

$$\Phi(\Delta)=\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2+\frac{\hat{\sigma}_r^2}{n_r}+\frac{\hat{\sigma}_c^2}{n_c}+\frac{\hat{\sigma}_{pr}^2}{n_r}+\frac{\hat{\sigma}_{pc}^2}{n_c}+\frac{\hat{\sigma}_{rc}^2}{n_r n_c}+\frac{\hat{\sigma}_{prc,e}^2}{n_r n_c}}$$

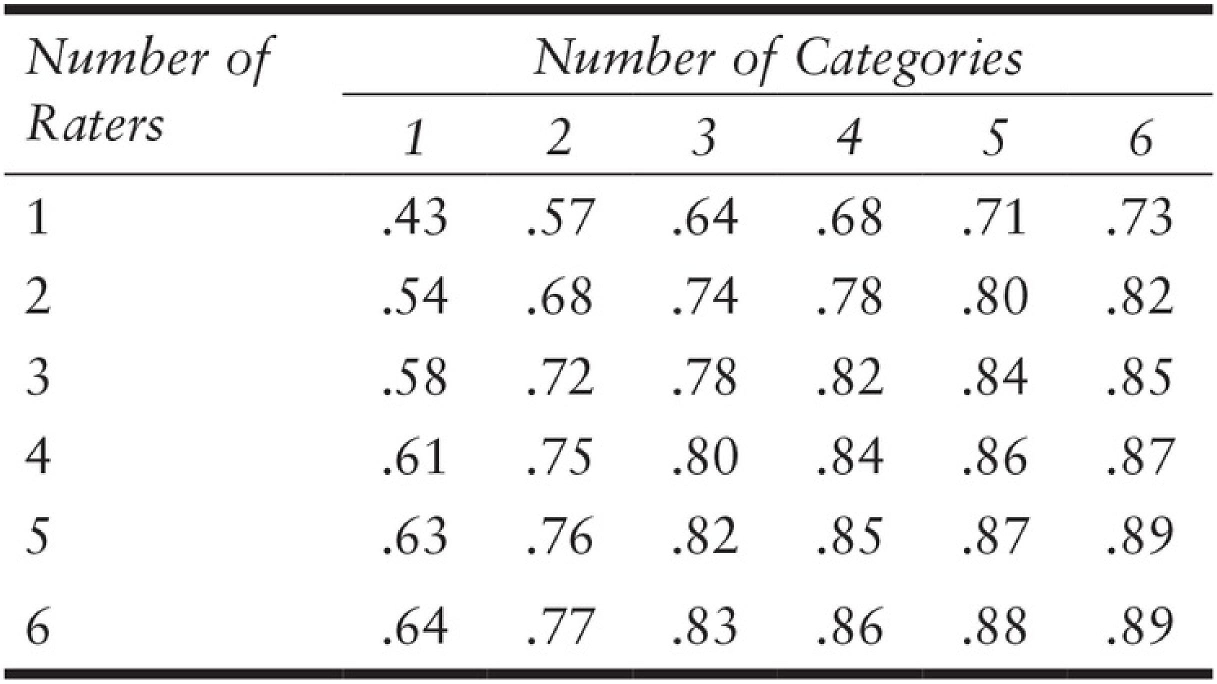
$$=\frac{2.95}{2.95+\frac{.26}{2}+\frac{.70}{5}+\frac{.30}{2}+\frac{.50}{5}+\frac{.23}{2(5)}+\frac{1.91}{2(5)}}$$

$$=\frac{2.95}{2.95+.13+.14+.15+.10+.023+.191}$$

$$=\frac{2.95}{2.95+.734}=\frac{2.95}{3.684}=.80076\approx .80.$$

This of .80 indicates that the scores based on two raters and five categories are about 80% dependable or represent about 80% universe score variance.

In a second stage, called a decision study or D study, a tester can change the values of n_r and n_c in the aforementioned formula and estimate ϕ coefficients for other potential numbers of raters and categories as shown in Table 1. Tables such as these are useful for studying potential changes in test design because they show the ϕ coefficients that are likely to occur if the numbers of levels in facets like categories and raters are changed. For example, Table 1 verifies that the scores from two raters using five categories are dependable at .80 as found about (see bold italics). However, the table also allows asking what-if questions such as (a) what if we wanted to achieve a ϕ coefficient of .84 (that can be achieved by using five categories and three raters, or four categories and four raters) or (b) what if we decided to only use one rater with our five categories (that would likely produce a ϕ coefficient of .71). Naturally, decisions about any test design changes will need to take into account the priorities, conditions, stakes, and resources of the specific people and institution involved. However, G theory and the ϕ coefficient allow testers to make such decisions on a rational basis.



Number of Raters	Number of Categories					
	1	2	3	4	5	6
1	.43	.57	.64	.68	.71	.73
2	.54	.68	.74	.78	.80	.82
3	.58	.72	.78	.82	.84	.85
4	.61	.75	.80	.84	.86	.87
5	.63	.76	.82	.85	.87	.89
6	.64	.77	.83	.86	.88	.89

James Dean Brown

See also Analysis of Variance; Classical Test Theory; Decision Consistency; Generalizability Theory; Phi Correlation Coefficient; Reliability; Scales

Further Readings

Brennan, R. L. (2001). Generalizability theory. New York, NY: Springer.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York, NY: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137–163. doi:10.1111/j.2044-8317.1963.tb00206.x

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: SAGE.

Mary L. McHugh Mary L. McHugh McHugh, Mary L.

Phi Correlation Coefficient Phi correlation coefficient

1251

1253

Phi Correlation Coefficient

The phi correlation coefficient (phi) is one of a number of correlation statistics developed to measure the strength of association between two variables. The phi is a nonparametric statistic used in cross-tabulated table data where both variables are dichotomous. *Dichotomous* means that there are only two possible values for a variable. As an example, the variable addressing life has only two levels, “alive” and “not alive” (or dead). So if a public health department was researching the proportion of newborns born alive versus born dead, each baby could be born alive or born dead; there are no other possibilities. Typically, such data are coded numerically for the computer. One level of the variable can be assigned the number 0 and the other level is assigned a number 1. To use the phi, both variables must be measured with only two levels. The symbol for the statistic is the lower – case Greek letter phi: ϕ.

The phi is the effect size statistic of choice for 2 × 2 (read two-by-two) table statistics such as the Fisher’s exact or a 2 × 2 chi-square. The data in columns and rows should be nominal, although it is frequently used with two-level variables measured at the ordinal level and for collapsed interval/ratio data. After providing background on the phi correlation coefficient, this entry reviews its assumptions and explains how to calculate and interpret and then concludes with a worked example.

Background

The phi was developed by Karl Pearson, who was one of the mathematicians involved in the development of the theory of general linear models. Pearson had a particular interest in correlation and developed a variety of measures, including the phi and the Pearson product moment correlation coefficient, better known today as the Pearson *r*. The phi is also a product moment correlation and provides correlation coefficient and significance results similar to results of the Pearson *r*. Many statistical computer programs (e.g., STATA, SPSS, SAS) compute the phi statistic and provide a significance level for the result.

As a correlation statistic, the phi measures the *strength* of an association between two variables. Correlation statistics provide four items of information:

1. They answer the question, “Do these two variables covary?” That is, does one variable change when the other changes?
2. When two variables do covary, these statistics describe the direction of the association, which can be positive or negative. A positive correlation means as one variable increases, the other also increases. A negative correlation means that as one variable increases, the other decreases.
3. Correlations describe the strength of the association. Strength in this context means how closely do the two variables change together? In a perfect correlation, for every one level of rise in one variable, the other variable would change exactly one level; it would either rise (positive correlation) or fall (negative correlation) that one level. The phi value can range from 0 to +1.0. (Given that the calculation requires the square root of a number, the result cannot be negative with the standard formula. Some other methods of calculation can return a negative number.)
4. The significance of the obtained phi value can be determined if hand calculated, and the statistical programs that produce the phi will provide a significance level.

Assumptions

The phi coefficient, like virtually all inferential statistics not specifically designed to test matched pairs or other related measures, assumes that the sample was randomly selected from a defined population. It assumes subjects were independently sampled from the population. That is, selection of one subject is unrelated to selection of any other subject. Like the chi-square, there must be an adequate sample size for the computed phi statistic to be useful. The chi-square demands that 80% or more of the cell expected values must be at least 5, and if this assumption is violated, neither the chi-square nor a phi calculated on the basis of that chi-square can be relied upon. It should be noted that samples smaller than 30 are considered to be very small samples, and small samples are less likely to be representative of the population of interest than larger samples. A sample size of 30 will, in most studies, provide a minimum of five for the expected values in all four cells.

Calculation

A great advantage of the phi coefficient is that it is so easily calculated from the chi-square result. The calculation is as follows:

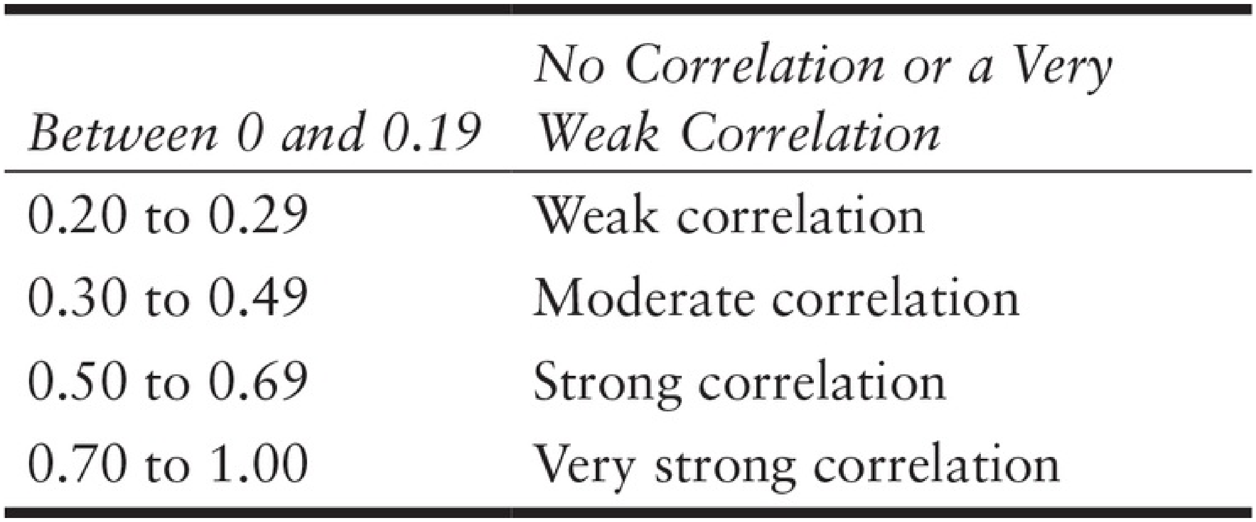
$$\varphi = \sqrt{\frac{\chi^2}{n}}.$$

An equivalent formula is phi^2 = chi-square ÷ sample size. The phi will be needed only if the significance of the contingency table is $p < .05$. It is a mistake to conduct further analysis, such as effect size testing, if the original test of independence on the table fails to produce a significant result. When the chi-square (or Fisher's exact) on a 2 × 2 table is nonsignificant, the range containing the correlation value will contain the value of 0. Thus, calculation of the phi coefficient is unnecessary because it is, by definition, not significantly different from 0.

Interpretation

Values for the phi coefficient can range from 0 to +1.0. A value of 1.0 means there is a perfect 1-to-1 correlation between the two variables. Like the Pearson *r*, the phi can be squared to obtain a measure of the amount of variance in the dependent variable that is explained by the independent variable. A phi coefficient of 0.68 means that the independent variable accounts for 46% of the variance in the dependent variable.

While different authors may use different values for weak, moderate, and strong correlation measures, the following table can be used as a general guide to the interpretation of the strength of effect size represented by various values of the phi correlation coefficient:

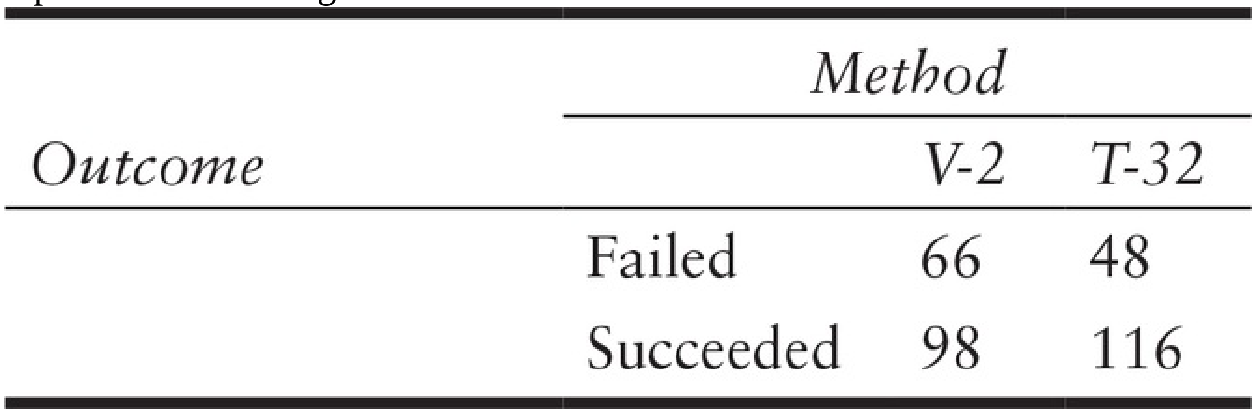


Between 0 and 0.19	*No Correlation or a Very Weak Correlation*
0.20 to 0.29	Weak correlation
0.30 to 0.49	Moderate correlation
0.50 to 0.69	Strong correlation
0.70 to 1.00	Very strong correlation

These interpretations are based on the amount of variance in the dependent variable explained by the independent variable. A correlation of +0.29 means that even if statistically significant, only about 8% of the variance in the dependent variable is explained by the independent variable.

Example of Phi Coefficient Use

Assume that a sample of 328 third-grade students in four elementary schools were prospectively studied for the effectiveness of a new method of teaching the multiplication tables. Method V-2 is the experimental method versus the existing method called T-32. It is known that there is a 60% success rate for students learning with the current (V-2) method. The purpose of the study is to discover if the new method can improve success rates by 10% or more. The following table represents the findings:



Outcome		*Method*	
		V-2	*T-32*
	Failed	66	48
	Succeeded	98	116

The results are as follows: chi-square = 4.36 (rounded), *p* < .05.

The phi coefficient is calculated as follows

The phi coefficient is calculated as follows:

$$\varphi = \sqrt{\frac{4.36}{328}} = \sqrt{0.0133} = 0.12.$$

Interpretation: The T-32 method resulted in a significantly higher success rate than the V-2 method (chi-square = 4.36, *df* = 1, *p* < .05). The effect size was very small (phi = .12). However, the success rate for the new method was 71%, and so the new method did achieve the desired increase of 10% in success rates over the old method.

Mary L. McHugh

See also Chi-Square Test; Correlation

Further Readings

Agresti, A. (1996). Introduction to categorical data analysis. New York, NY: Wiley.

Dattalo, P. (2016). Nominal association: Phi and Cramer’s V. Retrieved June 10, 2016, from http://www.people.vcu.edu/~pdattalo/702SuppRead/MeasAssoc/NominalAsso

Norton, B. T. (1978, February). Karl Pearson and statistics: The social origins of scientific innovation. Social Studies of Science, 8(1), 3–34.

Tonya Rutherford-Hemming Tonya Rutherford-Hemming Rutherford-Hemming, Tonya

Pilot Studies

Pilot studies

1253

1254

Pilot Studies

A pilot study is a research study that tests the feasibility of an approach that will later be used in a larger study. Pilot studies are conducted in quantitative and qualitative research. They can be extremely useful in providing justification or testing procedures for a larger future study. Pilot studies can benefit researchers by providing a “dress rehearsal” that saves time and avoids problems in the later study.

Pilot studies are not intended to test the hypothesis or research question for the larger study. Rather, pilot studies are meant to assess feasibility. Assessing the pragmatics of recruitment efforts, research instruments, randomization and data collection procedures, training sessions for staff, collaborative efforts, and intervention implementations are reasons why a pilot study is conducted. For example, a researcher may have concerns as to whether students would consent to being videoed during an actual class examination—a requirement of participants who enroll in a larger study investigating test anxiety. Because the researchers are worried about recruitment efforts, they may complete a pilot study initially to assess these efforts. Or the researcher may want to determine whether a 10-item multiple choice instrument is consistently read and understood by students who take it. In this instance, the researcher may opt to complete a pilot study to assess how the measurement instrument performs prior to using it in a larger study.

Because pilot studies are not focused on hypothesis testing, the sample size for a pilot study is often small. The sample size of the pilot study only needs to be large enough to provide meaningful information about the aspects that are being assessed for feasibility.

Completed studies with a sample size that did not meet the power analysis established a priori should not be referenced as a pilot study. This is a common error seen in the literature. Calling a study a pilot simply because it has a small sample size is incorrect. As mentioned previously, a pilot study focuses on feasibility, which will have different questions and objectives from the larger main study.

Some researchers use pilot testing as a means to determine effect sizes and sample size determination for the main study. This is controversial in the literature with some authors arguing that this process exceeds the limits of what a pilot study can do because of the small sample size. If pilot studies are used in this manner, it should be done cautiously, especially with treatment effects, as the estimates from the pilot study may be unrealistic or biased due to the small sample size.

Tonya Rutherford-Hemming

See also Effect Size; Power Analysis; Qualitative Research Methods; Quantitative Research Methods

Further Readings

Connelly, L. M. (2008). Pilot studies. MEDSURG Nursing, 17(6), 411–412.

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. Journal of Psychiatric Research, 45(5), 626–629. doi:10.1016/j.jpsychires.2010.10.008

Mazurek Melnyk, B., & Morrison-Beedy, D. (2012). Intervention research: Designing, conducting, analyzing, and funding. New York, NY: Springer.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., & Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. BMC Medical Research Methodology, 10(1). doi:10.1186/1471-2288-10-1

PIRLS

See Progress in International Reading Literacy Study

PISA

See Programme for International Student Assessment

Phillip K. Martin Phillip K. Martin Martin, Phillip K.

Ryan W. Schroeder Ryan W. Schroeder Schroeder, Ryan W.

Placebo Effect

Placebo effect

1254

1255

Placebo Effect

The placebo effect is a positive change in a person’s symptoms or condition after the administration of an inert substance (e.g., a sugar pill) or an ineffectual procedure (e.g., a teaching method that is not expected to achieve results) administered under the guise of being an effective treatment. Placebo effects have been demonstrated with a variety of medical conditions, including depression, irritable bowel syndrome, pain, asthma, and Parkinson’s disease, as well as occasionally in educational research. The presence and magnitude of the effect is not constant across situations; rather, it varies as a function of disease condition, type of outcome, and method of administration. Although symptoms under conscious awareness are often susceptible to a placebo effect, objective disease markers, such as blood sugar levels, are typically less amenable. Additionally, individual characteristics, such as age, personality, level of optimism, previous treatment experience, and presence of psychopathology, have all been found to influence the likelihood of a placebo response.

Mechanism

The placebo effect can be explained by two psychological theories: *expectancy* and *classical conditioning*. Research indicates that expectancy strongly influences the effect of a placebo response, and the response can be enhanced with verbal messages emphasizing that a positive change in symptoms should be anticipated. Classical conditioning theory has also been applied, and related research has found that placebos can effectively alleviate pain even after subjects are alerted to the fact that they are not receiving actual treatment, so long as the association between the placebo and analgesia has been well conditioned. The placebo effect, while psychologically mediated, has been found to affect neurobiological as well as cognitive processes. For example, increases in striatal dopamine and endogenous opioids in Parkinson's disease and pain patients, respectively, have been demonstrated following administration of a placebo.

Implications

Medical practitioners have been aware of the placebo effect for centuries, and knowledge of the phenomenon has become well incorporated in medical research. Because of the placebo effect, as well as the impacts of natural recovery and regression toward the mean, randomized controlled trials have become the gold standard for researching new drugs. In randomized controlled trials, participants are randomly assigned to either a treatment or a control group, with control participants often receiving a placebo. Ideally, assignment is double-blind, meaning that neither the participants nor the researchers are aware of who is receiving the placebo versus actual treatment. The treatment under investigation must then evidence incremental benefit above and beyond the placebo to demonstrate that any measured improvement is not an artifact of the placebo effect. Such a requirement poses a challenge for researchers investigating treatments for certain conditions such as depression, where the placebo effect is often quite high. Given the significance of the phenomenon, recent research has begun to investigate how lessons learned from the placebo effect can be harnessed by medical providers and clinically applied.

Phillip K. Martin and Ryan W. Schroeder

See also Classical Conditioning; Double-Blind Design; Hawthorne Effect; Random Assignment; Regression Toward the Mean

Further Readings

Benedetti, F., Enck, P., Frisaldi, E., & Schedlowski, M. (Eds.). (2014). Placebo. New York, NY: Springer.

Sanderson, C., Hardy, J., Spruyt, O., & Currow, D. C. (2013). Placebo and nocebo effects in randomized controlled trials: The implications for research and practice. Journal of Pain and Symptom Management, 46(5), 722–730. doi:10.1016/j.jpainsymman.2012.12.005

David Kahle David Kahle Kahle, David

Poisson Distribution

Poisson distribution

1255

1257

Poisson Distribution

The Poisson distribution is a family of discrete probability distributions on the counting numbers 0, 1, 2, 3, … , typically representing the number of occurrences of some event over a given unit of time, length, area, or other continuous unit. The distributions are parameterized by their expected number of observations over a given interval, a quantity referred to as its *rate* or *intensity* and denoted as λ. One of the oldest and most commonly used families of probability distributions, the Poisson distribution, has deep mathematical connections to many other important distributions, including the binomial distribution and the normal distribution.

Historical Context and Assumptions

The Poisson distribution is attributed to the prolific 19th – century French mathematician Siméon Denis Poisson, although it was known by the probabilist Abraham de Moivre well over a century earlier. Poisson's discovery was chiefly motivated as an approximation to the binomial distribution, described in the next section.

Technical details aside, the number of occurrences of an event over a given period of time can be said to follow a Poisson distribution with rate λ if four basic assumptions are met. Suppose an experimenter is going to observe some phenomenon over time (or other continuous unit). The basic Poisson assumptions are as follows:

1. The probability of observing one event in a short period of time is

approximately equal to the rate λ times the duration of the period.
2. The likelihood of observing two or more events in a short interval is approximately zero.
3. The probability of observing j events in one time period and k events in a separate time period is equal to the product of those probabilities individually.
4. The rate referred to in (1) does not change over time.

These assumptions have been used in the past to justify many applications, including (famously) the number of soldiers killed per year by horse kicks, eye movements of various types per minute while reading, and the defects in manufactured magnetic tape per yard. These ideas are generalized by the notion of a *Poisson process*, an important stochastic process studied in probability and statistics.

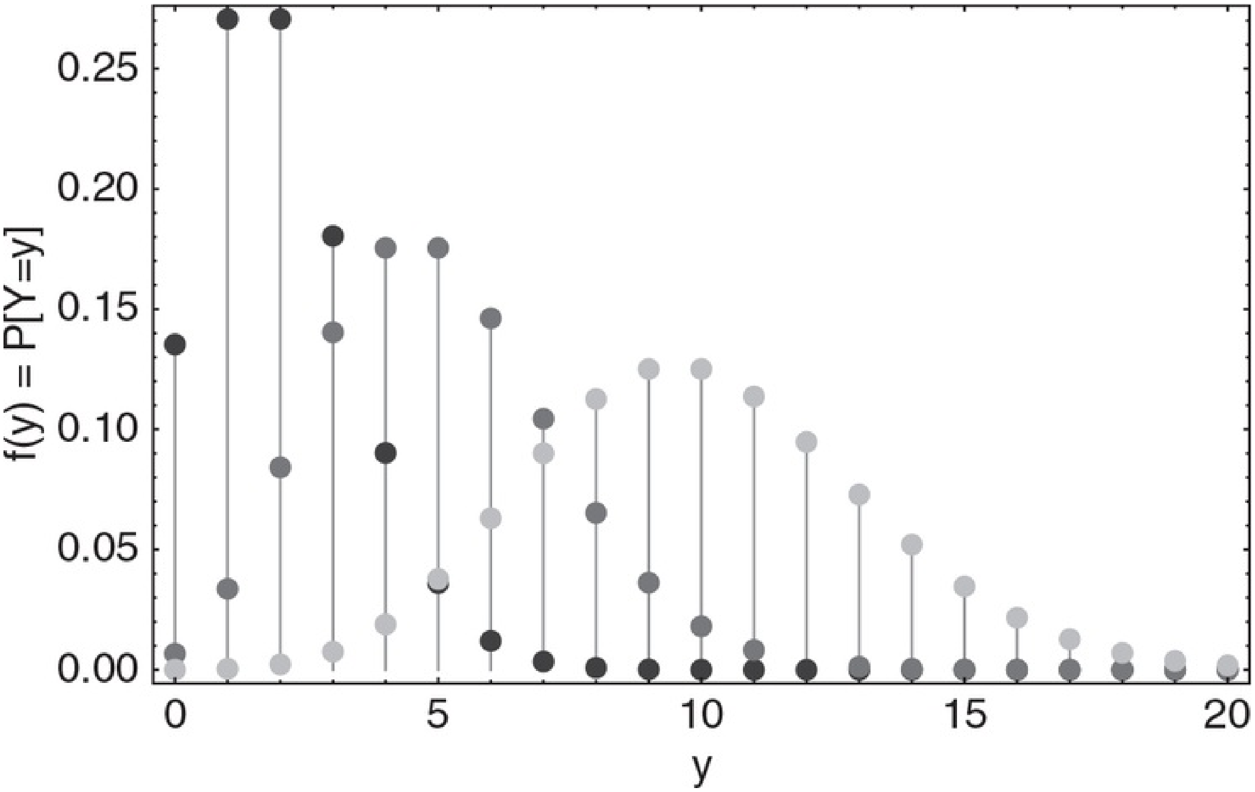
Mathematical Properties

A random variable Y with the Poisson (λ) distribution, denoted $Y \sim \mathrm{Pois}(\lambda)$, assigns a potential outcome $y = 0, 1, 2, \ldots$, probability following the Poisson formula:

$$f(y) = P[Y = y] = \frac{e^{-\lambda}\lambda^{y}}{y!},$$

where $\lambda > 0$ is the rate constant, e is Euler’s exponential constant (roughly equal to 2.72), and $y! = y(y-1)(\mathrm{y}-2) \ldots (2)(1)$ is the factorial of the positive integer y and 0!, defined to be 1. Its mean (expected value) and variance are both equal to the rate parameter λ, and consequently, the larger the expected number of counts, the larger those counts are expected to vary across repeated sampling. The distribution is characteristically right skewed, as shown in Figure 1; however, the skewness diminishes as λ grows.

Figure 1 Probabilities assigned by Poisson distributions with λ = 2, 5, and 10



The Poisson distribution is closely related to the binomial distribution through a mechanism referred to as the *law of small numbers* or rare events. Specifically, the probabilities assigned to the numbers 0, 1, 2, … , *n* by a binomial distribution with parameters *n* and *p*, , converge to those assigned by the Poisson distribution with λ = *np* as *n*→∞ and *p*→0, with *n* and *p* "balancing out" to the fixed positive constant λ. As a consequence, the Poisson probability formula can be used to approximate that of the binomial; it is typically recommended that *n* be at least 20 and *p* be less than 5%.

The Poisson distribution shares other similarities with the normal distribution as well. Like the binomial distribution, the Poisson distribution exhibits the reproductive property: if Y_1 and Y_2 are independent Poisson random variables with parameters λ_1 and λ_2, then their sum $Y = Y_1 + Y_2$ also has a Poisson distribution with parameter $\lambda_1 + \lambda_2$. Moreover, the same is true of any finite collection of Poisson variates. However, because the central limit theorem demands that the sum of independent random variables follow a normal distribution, the Poisson distribution can be well approximated by a normal distribution for large λ; at least 20 is often recommended.

Common Applications of the Poisson Distribution

Common Applications of the Poisson Distribution

Perhaps the most common use of the Poisson distribution is in Poisson regression, a generalized linear model that seeks to understand the association between a count response variable *Y* and one or more predictors $X_1, X_2, \dots, X_p$. In Poisson regression, the count variable is assumed to follow a Poisson distribution, the logarithm of whose rate is modeled as a linear function of predictors. Mathematically, if λ represents the rate of the Poisson variate *Y*, then λ is modeled as

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

In practice, the regression parameters typically fit with maximum likelihood, and they are interpreted as having multiplicative effects on the rate. For example, if β_k is the regression parameter corresponding to the predictor X_k, the interpretation of β_k is that a one-unit increase in X_k affects an $\exp(\beta_k)$ multiplicative increase in the rate of *Y*. If $\beta_k < 0$, $\exp(\beta_k) < 1$, so that the “increase” is in fact a decrease.

The Poisson distribution is also used extensively in the analysis of contingency tables. In the context of discrete multivariate analysis, the Poisson distribution is used as a component of loglinear models, which are closely related to logistic regression. In such a setting, the cells of a contingency table are assumed to follow Poisson distributions whose rates decompose into overall baseline effects, first-order effects, and so on, similar to the analysis of variance.

One of the chief limitations of the Poisson model is that the variance is “tied” to the mean: As the expected count increases, so too does the variability about that count. In practice, this assumption is often easily assessed. A scenario where variability is greater (or less) than what would be expected under the Poisson model is termed *overdisperse* (or *underdisperse*). A common remedy of this is to transition from a Poisson model to a negative binomial model where two parameters are available and the variance is not tied to the mean. From a Bayesian perspective, the negative binomial distribution characterizes both the prior and posterior predictive distributions of a Poisson data model with a γ prior on the rate λ, which is a conjugate prior. From the frequentist perspective, the negative binomial can be seen as a continuous mixture of Poisson distributions, weighted by the γ distribution.

Another common limitation of the Poisson model is the probability it places on observing zero counts. In practice, it is common to encounter data that have more zero counts than would be expected under a Poisson model. A common remedy for this is to use a modified Poisson model called a *zero-inflated Poisson* model.

David Kahle

See also Binomial Test; Normal Distribution

Further Readings

Agresti, A. (2007). An introduction to categorical data analysis (2nd ed.). New York, NY: Wiley.

Agresti, A. (2012). Categorical data analysis (3rd ed.). New York, NY: Wiley.

Bishop, Y., Fienberg, S. E., & Holland, P. (2007). Discrete multivariate analysis: Theory and applications. New York, NY: Springer-Verlag.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). Univariate discrete distributions (3rd ed.). New York, NY: Wiley.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). Discrete multivariate distributions. New York, NY: Wiley.

Poisson, S. D. (1838). Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, Précédées des Règles Générales du Calcul des Probabilités. Paris, France: Bachelier, Imprimeur-Libraire.

Sheskin, D. J. (2011). Handbook of parametric and nonparametric statistical procedures (5th ed.). Boca Raton, FL: Chapman and Hall/CRC.

Stigler, S. M. (1982). Poisson on the Poisson distribution. Statistics &

Probability Letters, 1(1), 33–35. doi:10.1016/0167-7152(82)90010-4

Thompson, J. (2000). Simulation: A modeler’s approach. New York, NY: Wiley.

Chi Yan Lam Chi Yan Lam Lam, Chi Yan

Policy Evaluation

Policy evaluation

1257

1261

Policy Evaluation

Policy evaluation refers to the systematic investigation and determination of value of a policy—and can take place in various sectors, including education. Policy evaluators apply evaluation methodologies and employ social scientific research methodologies to answer evaluation questions in support of policy making, policy development, and policy decision making. Policy evaluation is rarely undertaken for its own sake but mostly conducted in connection to the policy cycle. Because of the inherent political nature of policies, policy evaluation is increasingly undertaken by policy actors—both governmental and nongovernmental parties—with interests in shaping the political agenda. Hence, awareness into both the technical and political dimensions of policy evaluation is important to its understanding and execution.

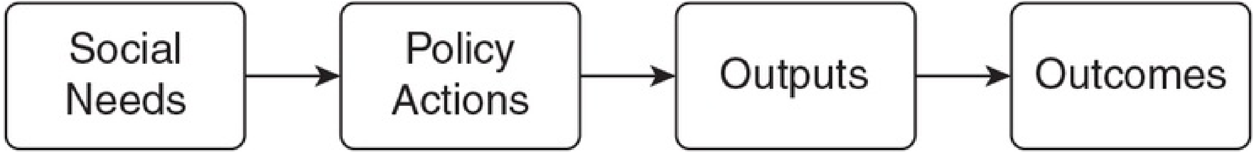
Policy as the Evaluand

The focus on policy as the evaluand—the object to be evaluated—in policy evaluation demarcates the unique domain within which policy evaluation operates. The specific nature of policy—its constitution, function, and purpose—carries implication for its evaluation.

A policy can be broadly understood in terms of the system of ideas and actions that are intended to either promote or constrain certain actions and behaviors for the purpose of achieving certain intended valued outcomes. Policies are almost always introduced in response to a perceived discrepancy between what is desired and what is the current state, otherwise known as a need. The system of ideas and actions, which constitute a policy, ought to offer a logical remedy to the perceived need; policy actions usually take on the form of policy instruments, such as legislation, agreements, and programs and services. Through implementation of the policy, intended valued outcomes are to be achieved. A core goal of policy evaluation is to determine the extent to which intended outcomes have been realized or not. Put simply: Is the policy making a difference in ways policy makers had expected?

Yet, as rational as a policy may sound in principle, the design of policies is never as rational or immutable as one would like. A helpful way to think about policy is through framing it in a problem-solving frame (Figure 1). A policy comprises policy actions (solution) intended to accomplish some valued goals (outcomes) in response to remedy a perceived social need (problem). The interrelationships between that of policy actions as a response to some perceived social need and that of policy actions and instruments and its ability in effecting intended valued outcomes is an important one to observe: They are tentative at best until a policy becomes implemented and its effects become knowable. Policy evaluation, therefore, serves an important function in bringing empirically generated evidence to bear on understanding the substance and consequences of a policy.

Figure 1 A model of the logical relationships underpinning components of a policy. A policy may be analyzed and understood in terms of the relationships which underpin various components of a policy. These relationships, once made explicit, can be examined in a policy evaluation.



Types of Policy Evaluation

In practice, policy evaluation can be conducted in a number of ways differing in the manner in which they are conducted, the policy actors who engage in them, and their intended effects. Michael Howlett, M. Ramesh, and Anthony Perl observe three broad categories: administrative evaluation, judicial evaluation, and political evaluation.

Administrative Evaluation

Administrative evaluation is concerned with the logical consistency and implementation of the policy concerning how resources are used to organize activities, and the extent to which organized activities are able to accomplish intended outcomes. Administrative evaluations are typically undertaken by governmental actors, usually as a part of policy implementation, to understand how a policy is implemented in context and its effects on populations. Conducting an administrative evaluation of a policy requires the systematic collection of data concerning the policy’s implementation: demographics and other characteristics of interest concerning the targeted population, detailed description about organized activities, as well as careful, defensible measurement concerning outcomes of interest. The scope of administrative evaluation is wide and the efforts needed to conduct one can be extensive. For that reason, administrative evaluations are typically focused on more manageable aspects of the policy.

Process evaluation typically focuses the evaluation on the efficient delivery and operation of the activities. The relationship of interest is that of inputs and activities. The objective is usually to identify processes for which more resources are used than is typically needed (efficiency), the extent to which the processes make a meaningful contribution toward the goals of the policy (effectiveness), or whether the activities are conducted in a responsible way or as agreed upon between the funding party and the implementation party (accountability).

Performance evaluation typically focuses the evaluation on the immediate outputs stemming from administrating the activities. Examples include enrollment, attendance, retention (the number of students or clients who remain with the intervention until the end), and “success” (the number of students or clients who remain with the intervention until the end and have achieved the predetermined standard).

Efficiency evaluation focuses the evaluation on understanding whether a policy can achieve a comparable level of performance in outputs using less resources. Such a determination requires a careful examination into the inputs–activities–outputs relationship. Duplication or redundancy in services and organizational processes are typical foci in efficiency evaluation. Such evaluation may also make a determination as to whether a comparable level of performance can be achieved by substituting a less costly option over a more costly one.

Effectiveness evaluation focuses the evaluation on determining the ways in which and the extent to which a policy achieves its intended outcomes—that is, changes in state—given the inputs invested. In practical terms, effectiveness evaluation is sometimes referred to as value-for-money evaluation. It is helpful to distinguish policy outcomes by logical expectancy in terms of near-term outcomes and far-term outcomes (impact). This is because policies are often articulated in far-term outcomes, which tend to be broad and ambiguous. Focusing efforts on realizing the near-term outcomes helps to ensure that far-term outcomes are achievable in the long run.

Much of administrative evaluation compares what gets implemented of a policy against what was planned or articulated. This mode of comparison is sometimes referred to as goal-attainment evaluation. Focusing an evaluation on goals can ignore wider dynamics at play. First, a policy evaluation is wise to account for unintended consequences arising from the implementation of the policy. Doing so would require the evaluation to be open to unintended changes in the context or systems that the policy is intended to shape; adopting correspondingly a methodological approach that would facilitate examination into consequences beyond what had been anticipated, expected, and intended is important. Second, it may be important to approach an evaluation without reference to a policy's action or targets in order to describe and "reflect back" to policy stakeholders what is truly happening "on the ground." On that basis, a determination is made. This approach is known as the goal-free approach to evaluation. The value of this approach is derived from the observation that the premise upon which a policy was formulated or its policy actions may in fact be flawed or incomplete in addressing the underlying social need.

Judicial Evaluation

Judicial evaluation is concerned with the legal issues surrounding the substance and implementation of the policy. Judicial evaluation applies legal principles and standards to determine the soundness of policies. For that reason, judicial evaluation is often undertaken by the judiciary and by professionals with legal training. In education, examples include testing the constitutionality surrounding particular policies and examining policy with respect to providing equitable access to programs and fair treatment of groups in their participation. A policy may also be subjected to an ethical evaluation. An ethical evaluation is concerned with measuring a policy against preexisting value systems and established ethical standards.

Political Evaluation

Political evaluation refers to evaluation undertaken by political actors who desire to shape the political discourse around a policy. This type of evaluation distinguishes itself from the other two by placing less emphasis on the rigor with which evaluative claims are generated and advanced and more on the messages that the evaluation sends. Hence, the goal is rarely aimed at improving policy but to advance a political agenda. In 2009, Howlett, Ramesh, and Perl observed that "unlike administrative and judicial evaluations, political evaluations are usually neither systematic nor technically sophisticated. Indeed, many are inherently partisan, one-sided, and biased" (p. 174). In some cases, a politically motivated evaluation can result in the evaluation being judged as gratuitous (i.e., lacking in any genuine intention or effort in using the evaluation) or as symbolic (i.e., using evaluation as a mechanism to justify preexisting positions). In other cases, a political evaluation can lead to genuine policy improvement by enabling participation from political actors.

Logic of Evaluation

Despite the many types of policy evaluation, at the core of any policy evaluation is the requirement to determine the value—merit, worth, and significance—of the policy. This determination generally involves selecting criteria for determining merit, setting standards of performance, measuring performance, and synthesizing the results into a value statement. What constitutes a policy of value is itself a value-laden determination. Whose values ought to be incorporated and by what means could values be surfaced have been the subject of many debates, particularly in hotly contested political issues, and the topic of many scholarly articles. Any policy evaluation that shies away from making an evaluative claim risks falling short of today's professional standards.

Approaches to Evaluation

Policy evaluations may be further classified by their intended purpose. A summative evaluation renders a summative judgment about the value of a policy. Summative evaluations typically inform decisions surrounding continuation or cessation of funding or to allow for comparison between alternative policy strategies. Formative evaluation renders a judgment about the value of a policy but well in advance of the summative evaluation. A formative evaluation is out well in advance of the summative evaluation. A formative evaluation is typically narrower in scope than a summative evaluation, allowing for a more focused examination around processes, structures, and other components of the program. Formative evaluation typically supports improvement processes. Developmental evaluation, a recent methodological advancement in evaluation, supports the development—refinement (doing something better), innovation (doing something differently), and evolution (ongoing change)—in the overall policy and/or its components. The evaluator works collaboratively as an embedded member of the development team to support evaluation inquiry. Together, these three approaches afford the contemporary policy evaluator evaluation approaches that are compatible with all phases of a policy lifecycle—from its inception to its cessation.

Challenges

Policy evaluation can suffer from a multitude of challenges:

- Difficulty in identifying and observing target population, particularly with difficult-to-serve populations
- Reaching agreement among stakeholders concerning values systems and performance standards
- Lack of strong evidence base to support or justify policy content
- Lack of appropriate or sufficiently robust measures
- Difficulty in accessing or generating appropriate data
- Difficulty in establishing a basis for comparison
- Undue pressure to conduct an evaluation at a hastened pace or to produce favorable results.

To guard against these challenges, the field of evaluation has established professional standards. Examples include the *American Evaluation Association Guiding Principles for Evaluators* and Canadian Evaluation Society’s *Competencies for Canadian Evaluation Practice*. The field has also made a concerted effort to delineate what constitutes quality evaluation processes and products. The *Program Evaluation Standards*, published by the Joint Committee for Standards for Educational Evaluation, is one example. The Canadian Evaluation Society has established a credentialing program to assist evaluation consumers in identifying and procuring services from competent evaluators. These developments reflect the ongoing professionalization of the field to better meet the needs of evaluation users and enhance the quality and utilization of evaluation.

Chi Yan Lam

See also American Evaluation Association; Developmental Evaluation; Formative Evaluation; Logic Models; Process Evaluation; Program Evaluation; Stakeholders; Summative Evaluation

Further Readings

Centers for Disease Control and Prevention. (2014). Using evaluation to inform CDC’s policy process. Atlanta, GA: Centers for Disease Control and Prevention & U.S. Department of Health and Human Services.

Howlett, M., Ramesh, M., & Perl, A. (2009). Studying public policy: Policy cycles and policy subsystems. Toronto, Canada: Oxford University Press.

Patton, M. Q. (2011). Developmental evaluation: Applying complexity concepts to enhance innovation and use. New York, NY: Guilford Press.

Hamish Coates Hamish Coates Coates, Hamish

Policy Research

Policy research

1261

1264

Policy Research

In general, policy research involves developing, implementing, and evaluating the legislative and allied instruments that, at a minimum, set conditions in key areas such as funding, regulation, and quality assurance. Although policy plays out in diverse venues, this entry focuses on approaches related to education systems and institutions. Such policy involves the public, political advisers, consultants, leaders, academics, unions, and interest/lobby groups, among others. International agencies such as the World Bank, the Organization for Economic Cooperation and Development, the Asian Development Bank, and the United Nations Educational, Scientific and Cultural Organization play a prominent role in education policy. Specialists may be called on to examine particular problems; although as sectoral and jurisdictional boundaries are blurred by general sociopolitical changes, there is increasing emergence of generalist education researchers and leaders. A useful distinction can be made between intrinsic and extrinsic policy research, whereby intrinsic research is interested in policy itself as the object of study and extrinsic research looks at policy in context. This entry examines intrinsic and extrinsic perspectives in relation to education policy research, describes developments in extrinsic policy research, and provides two case studies.

Intrinsic Research

Research on policy itself may be conducted by an eclectic array of people from a range of disciplinary and institutional backgrounds. Political scientists, ethicists, sociologists, and economists may bring differing theories and methods to similar policy issues. Policy research of this kind may focus on the nature of policy as a phenomenon or in a large-scale sector, rather than education concerns per se. Such researchers may invoke approaches such as *agenda setting*. By this account, an agenda gains traction through problematization (with problem recognition and definition playing out in a range of ways, e.g., research, conflict, crises, collaboration, or lobbying). Proposals are the next step for advancing an agenda. Such proposals take many forms. The third force is politics, which goes to matters such as social mood, voting patterns, and partisan preferences. These three agendas—problems, proposals, and politics—are seen to operate relatively independently, with at least two being required, either by serendipity or design, to spawn a policy window. As an example, agenda setting might begin by using the media to evoke problems with current education funding policy, sending proposals to opinion leaders to garner engagement in revised approaches, then seizing an opening provided by an election or budget to advance the agenda.

Another approach is *policy entrepreneurship*, in which people or agencies take risks to advance new policy ventures. Such entrepreneurship involves social acuity, defining problems, building teams, and leading by example. By way of example, a policy entrepreneur might work in a think tank or a university, build networks and resources to articulate an initiative, and then draw on their tenacity and authority to advance a proposal.

In practice, of course, the evolution of policy, particularly radical policy, is likely to be far messier than this theory portrays. Economists and lawyers may examine policy through regulatory or market perspectives or by invoking pan-sectoral theories of power or pricing.

Extrinsic Research

Most extrinsic research treats education policy as an indirect or instrumental phenomenon related to substantive educational matters. Such research deploys a range of contextually appropriate theories and methods, which may be played out in either conventional or contemporary ways.

Conventional Approaches

The conventional education policy cycle frames various kinds of research. In articulating this cycle, it is important to keep in mind that it is an abstraction;

policy is invariably formed and researched in very cultured ways, and espoused approaches to policy leadership often vary quite considerably in practice.

Essentially, during elections or time in power, political or organizational leaders adopt positions spanning broad ideology and views on system dynamics, particular practices, and systems settings. Supportive policy research at this stage is more likely to involve market research, opinion polls, or syntheses of prior studies. Particular energy may be invested in foresight work and international reconnaissance to develop policies that resonate with electorates by responding to known issues and advancing core agendas. More reactive research might involve ways to inform lobbying or the shaping of public discourse.

Once policies have reached a relevant level of maturity, governments seek to enact strategies in their attempt to steer systems. These range from private negotiations with system actors to open public consultations. New policy resources are prepared along with legislation, as required, for approval by an appropriate political forum. Approved policy is then implemented via a range of instruments that deliver information, compliance, and incentive measures. Research may well be involved in this process, taking the form of formative evaluation, focus groups, and stakeholder research or feasibility assessment. A diverse suite of researchers may be involved in this work, including sociologists, political scientists, education researchers, assessment specialists, ethnologists, and environmental scientists.

Following policy implementation, monitoring mechanisms are put in place, with reviews conducted on an ad hoc or scheduled basis. Conventionally, it is this stage of the cycle in which policy research might be expected to play a particularly prominent and diverse role. General critical analyses may draw together an eclectic range of theories to analyze the shortcomings of existing policies or to suggest alternative solutions. Economic or financial analyses focus on the costs and returns of education for systems, institutions, or individuals. Assessment studies test members of a target population and use these to produce summary results regarding various context or demographic groups. Audit methods involving document reviews and stakeholder consultations may be used. Sometimes, policy analysts may research the failure of policy implementation or outcomes. Increasingly, more structured and deliberative approaches are being developed to guide integrated research into the social, economic, and environmental facets of education policy. Although experience is always more complex, in principle, the outcomes of research into policy implementation feed back into political deliberations, and the cycle iterates.

implementation feed back into political deliberations, and the cycle iterates.

This conventional approach offers a frame for understanding education policy research, which is parsimonious and persuasive but not without its limitations. It assumes a top-down, cyclical, abstract, and functional approach that may work in ideal cases but seems unlikely to play out most of the time. Education involves people, and education policy is almost invariably going to be more complex and difficult than a simplified rational model can convey. The need for a more sophisticated and powerful understanding of policy research has spurred development of a range of contemporary approaches.

Contemporary Approaches

A contemporary and well-tested suite of policy methodologies is starting to infiltrate education policy, shaping new embedded kinds of research. These methodologies are referred to variously and in overlapping ways as *cocreation*, *coproduction*, and *codesign*. This suite of approaches involves the design and delivery of education in ways governed equally by stakeholders such as governments, institutions, industries, and communities. In practice, such shared creation plays out via workshops, meetings, and one-to-one liaison between individuals. Such work does not attempt to simplify an invariably complex ecosystem. It operates with a leveled dialogue by transgressing the dialectic between the governors and the governed. Technology plays an important role in facilitating broader communication and collaboration, spawning what may be referred as “digital-era governance.” Research in this setting can involve “big data” (very large data sets that defy conventional forms of capture, storage, and modeling) as well as expansive forms of qualitative and behavioral insight. The agile and dynamic interplays support and stimulate a more entrepreneurial and consultative approach to generating policy ideas and practices. Research plays a more ongoing, dynamic, and intrinsic role within these emerging policy environments.

These digitally devolved forms of policy governance and research have implications for conventional forms of regulation involving governments, institutions, and quality agencies. Rather than regulatory and funding agencies working with institutions to determine the distribution and level of supply (a supply-driven model), stakeholders, who include potential students, play a greater role in shaping participation (a demand-driven model). The role of regulatory agencies declines or is repositioned given the added voice of a wider stakeholder group. More marketized systems diminish the role and power of stakeholder group. More marketized systems diminish the role and power of political forms of regulation. This shift echoes changes being discussed or legislated in many systems.

Case Studies

Reviewing two case studies of research on a frontier education policy gives life to these ideas. The assessment of students’ learning outcomes is a pressing matter for higher education, and those involved contend that this area could be improved, although there is marked divergence of opinion regarding the nature and extent of change.

A case study of conventional policy research can be seen in the promulgation and decline of the Assessment of Higher Education Learning Outcomes Feasibility Study conducted by the Organization for Economic Cooperation and Development. The initiative was launched in 2006 at a ministerial meeting that advocated the need for more reliable and valid information about the student-learning outcomes of higher education. Expert and stakeholder lobbying led the launch of this 17-country study run between 2009 and 2013. Organization for Economic Cooperation and Development contracted a range of organizations to deliver top-down policy and technical research. These organizations worked with experts and policy makers to design tests and questionnaires; deliver them to students, faculty, and policy makers; and analyze and report results to inform international and national initial policy concerns. A series of evaluation mechanisms were put in place during the feasibility assessment with a view to establishing Assessment of Higher Education Learning Outcomes itself as a trusted benchmark. Ultimately, however, the lack of stakeholder and political engagement rendered the study unable to contend with entrenched sectoral and institutional power interests.

An example of a more contemporary approach to education policy is provided in the large-scale collaboration by the European Commission titled Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe. The Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe project has advanced a more consultative and collaborative approach involving cocreation among key stakeholders. Rather than convene expert groups that design resources and approaches, Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe involves working with practicing academics and institutional representatives to specify what higher education students should know and be able to do, to develop instruments from existing materials, and to implement assessments on the back of existing collegial vehicles. This program of policy research involves the same kinds of study of Assessment of Higher Education Learning Outcomes, using structured and consultative bottom-up as opposed to structured but closed top-down approaches.

Policy researchers are articulating this kind of progression from conventional to contemporary approaches to policy across the education spectrum, with immediate implications for policy research. Drawing on the advances afforded by big data, the future of education policy research promises to be more nimble, consultative, and influential. Rather than help with policy design, development, and implementation, researchers will be actively engaged in more seamlessly shaping how people are educated from early childhood through to old age.

Hamish Coates

See also Policy Evaluation; Program Evaluation

Further Readings

Asian Development Bank. (2017). Education [web page]. Retrieved from http://www.adb.org/sectors/education/main

Coates, H. (2017). The market for learning: Leading transparent higher education. Singapore: Springer.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2005). New public management is dead—Long live digital-era governance. Journal of Public Administration Research and Theory, 16, 467–494. doi:10.1093/jopart/mui057

Izard, J. (1997). SACMEQ’s approach to educational policy research. In K. Ross (Ed.), Quantitative research methods for planning the quality of education. Paris, France: IIEP.

Organization for Economic Cooperation and Development. (2017). Directorate for education and skills. Retrieved from http://www.oecd.org/edu

Pestoff, V., & Brandsen, T. (2007). Coproduction: The third sector and the delivery of public services. London, UK: Routledge.

United Nations Educational, Scientific and Cultural Organization. (2017). Education for the 21st century. Retrieved from http://en.unesco.org/themes/education-21st-century

Daniela Jiménez Daniela Jiménez Jiménez, Daniela

Portfolio Assessment

Portfolio assessment

1264

1267

Portfolio Assessment

A portfolio is a collection of items that document the professional trajectory or performance of a person in a particular field. Although portfolios originally emerged in artistic contexts as a way of documenting an artist’s work with actual samples, their use has extended to multiple fields. In the field of education, portfolio assessments for teachers have been popular since the 2000s. Portfolios are also commonly used to assess students in the classroom. While this entry focuses on portfolio assessment within an education context, the principles and components apply equally to any holistic assessment procedure that includes a collection of products used as evidence of an individual’s performance.

Portfolio Features

The design of a portfolio is tailored to the purpose of the assessment. It is therefore important for teachers or test developers to clarify this purpose at the beginning of the design process in order to make decisions that ensure coherence in its features. Portfolio attributes can be organized in terms of four categories: its structure, its format, the types of evidence collected, and the scoring procedure.

Structure

The level of flexibility in the kinds of evidence and the formats to be submitted during a portfolio assessment can vary from highly structured and standardized to fully open-ended. For instance, when portfolio assessment is used to compare performance among many people, as is the case during large-scale evaluation performance among many people, as is the case during large-scale evaluation programs, portfolio specifications can include detailed instructions indicating not only the products to be submitted and the goal of each but also carefully specifying the order and format in which each is to be presented. Additionally, portfolio assessments may include forms to be attached to the pieces of evidence, so that the respondent has the responsibility of correctly presenting the requested work. Highly structured portfolio assessments are useful to minimize complexity when scoring and to ensure consistency in evaluation; however, it is worth noting that such standardization implies a loss of authenticity in how the evidence is presented and can increase the burden on respondents who may need to invest additional time and work in order to adhere to the portfolio specifications.

In contrast, when the purpose of the portfolio is to assess learning outcomes or progression, as is the case within a small-scale context such as a college course, the instructions for the respondents can be considerably more general, only indicating criteria to choose some pieces of work, for example, by chronological order or specified content. If the assignment is aimed at capturing thoughts, opinions, or reflections about practices, the respondents are usually asked to add labels to clearly link their expressed views to specific portfolio works. It is important to keep in mind that the fewer specifications the respondents are given, the wider the variety of collected evidence will be, and the greater the time and effort needed from the reviewers in interpreting and scoring the portfolios. Thus, when respondents are given fewer instructions, which could be desirable when teachers have few students or to enhance inspired and creative work, then more guidelines will be needed during the scoring procedure to ensure that the same criteria are applied to interpret similar evidence.

Format

The portfolio format addresses the kinds of evidence that are admissible within it. For instance, a portfolio in arts education may not allow the inclusion of a student’s actual pieces of work, instead allowing photographs showing the work or clips showing how the student is applying a particular learned technique. An important distinction regarding format is whether the portfolio to be assessed is paper or computer based. This choice of format involves not only the way in which the respondents obtain the instructions, complete, and submit their portfolio evidence but also the way the evidence is assessed and feedback or results are provided.

An e-portfolio can simplify the collection of evidence and reduce the amount of time spent in this procedure. In teacher evaluation contexts, for example, student work, photographs of materials, and video clips can all be created or saved using cell phones or tablets and uploaded quickly and easily, saving paperwork and time. Appropriately labeled computer files can simplify the classification of evidence and help with tracking. Finally, computer-based portfolios are easier to transfer from one place to another, and they allow for simple archiving and back-up.

In contrast, paper-based portfolios are less flexible in terms of the evidence they can encompass. They usually look like folders and allow the inclusion of flat pieces of evidence. Many portfolios for teacher evaluation purposes include recorded lessons or photographs, but they require a clear structure to classify and organize the evidence. Still, they are a good choice when the target population is not accustomed to using computer-based technology.

Types of Evidence

As a collection of work, a portfolio should include a variety of evidence that covers a set of content areas or competencies. The types of evidence depend on the purpose of the assessment and the instrument it is intended to measure. The inclusion of a reflection assignment based on the pieces of evidence is common as is a request for justification of the selection of work submitted. For example, a student portfolio can include a sequence of work with a reflection about how the sequence indicates improvement over the semester or a collection of the best work during the semester and a justification of why each is a good example.

Portfolios that are used to evaluate teachers include artifacts that account for different aspects of teacher performance. Even when the work to be assessed depends on the underlying teaching model or current teaching standards, they usually consist of documents related to lesson planning, partly or fully videotaped lessons, and examples of student assignments. They can also be complemented by comments, questions, and reflection tasks based on the submitted work, allowing the teachers to explain and justify their pedagogical choices and decisions.

Scoring Procedure

Once submitted, a portfolio is reviewed to interpret the evidence and give the respondent a score or rating. The scoring procedure typically involves the use of a rubric with specific criteria, which ensures coherence with the purpose of the assessment and what it is intended to measure. The scoring scheme can include the following criteria:

- Quality of the evidence submitted either individually or as a set. For instance, an essay is scored in terms of specific quality components, a set of written assignments is scored according to quality criteria indicated in the curriculum or syllabus, and a recorded lesson is scored using an observing protocol based on literature on instructional quality or current teaching standards.
- Quality of a task based on the evidence submitted. Such a task is usually reflective in nature, with the evidence serving as illustration or justification. This aspect considers, among other things, whether the selection of evidence is complete, adequate, coherent, or robust enough to support the reflective argument. For example, a portfolio for teacher evaluation can include a written reflection based on a videotaped lesson (also included in the portfolio), in which the teacher assesses the appropriateness of the teacher’s pedagogical choices or identifies strengths and weaknesses regarding a specific part of the lesson or teaching competency.

Because different types of evidence are collected together in a portfolio, it is also meaningful to consider the interplay between tasks and pieces of evidence in the scoring scheme. If every piece of evidence is scored independently and there is no single overall judgment, the sense of choosing a portfolio over other assessment formats could get lost, and the application of such an instrument should be reassessed.

An important issue regarding the scoring of portfolios is rater reliability, especially when the instrument is used in large-scale and/or high-stakes programs. Even when the scoring scheme is similar to other open-ended question procedures, there is an added complexity involved that is associated with the quantity of pieces of evidence and tasks as well as the connection between them. In order to guarantee that portfolio scores will not vary substantially depending on the person who rated them, the raters need to be specially trained in the use of the scoring rubric and should undergo a strict practice phase before starting the definitive scoring phase. Furthermore, the scoring procedure needs constant monitoring, usually involving double-scoring and/or expert rating; thus, scores are compared among raters, and interrater reliability is computed. Depending on the results, improvement measures could be implemented, such as the retraining of raters and rescoring of portfolios.

Disadvantages and Advantages

One of the disadvantages of portfolio assessments is that they are time consuming for the respondents and expensive to apply on a large scale due to the complexity of implementing a reliable scoring procedure. In order to deal with both these issues, it is recommended for test developers to carefully design the instrument to solicit the minimum evidence necessary. In addition, in the field of teacher evaluation, even when portfolios are supposed to account for everyday practices, they are highly structured and usually require documentation or production of material that go beyond the teacher’s usual work. Furthermore, it has been argued that the ability to follow instructions as well as writing ability could be implicitly measured in such an instrument, resulting in a contamination of the construct of interest that the portfolio is intended to assess.

On the other hand, a portfolio is useful as a single instrument that accounts for both *knowing what* and *knowing how*—that is, they can combine evidence stemming from both declarative statements along with authentic work samples. It is especially suitable for promoting individuals’ reflections on their own work, firstly because the purposeful selection of evidence implies a reflective process and secondly because the inclusion of authentic pieces of work is often complemented with related written tasks or documentation. In addition, as the collection of work can occur over a long period of time, a portfolio can represent a progression in learning processes or professional accomplishments. For these reasons, a portfolio is a versatile assessment instrument, one that can be used for summative or formative purposes; it can also be used in contexts where assessment is not the objective, for example, as a tool for professional development.

Daniela Jiménez

See also Authentic Assessment; Formative Assessment; InterRater Reliability; Performance-Based Assessment; Rubrics; Teacher Evaluation

Further Readings

Burner, T. (2014). The potential formative benefits of portfolio assessment in second and foreign language writing contexts: A review of the literature. Studies in Educational Evaluation, 43, 139–149. doi:10.1016/j.stueduc.2014.03.002

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. Teaching and Teacher Education, 16(5–6), 523–545. doi:10.1016/S0742-051X(00)00015-9

Karlin, M., Ozogul, G., Miles, S., & Heide, S. (2016). The practical application of e-portfolios in K–12 classrooms: An exploration of three Web 2.0 tools by three teachers. TechTrends, 60(4), 374–380. doi:10.1007/s11528-016-0071-2

Mansvelder-Longayroux, D. D., Beijaard, D., & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. Teaching and teacher education, 23(1), 47–62. doi:10.1016/j.tate.2006.04.033

Qvortrup, A., & Keiding, T. B. (2015). Portfolio assessment: Production and reduction of complexity. Assessment & Evaluation in Higher Education, 40(3), 407–419. doi:10.1080/02602938.2014.918087

Struyven, K., Blieck, Y., & De Roeck, V. (2014). The electronic portfolio as a tool to develop and assess pre-service student teaching competences: Challenges for quality. Studies in Educational Evaluation, 43, 40–54. doi:10.1016/j.stueduc.2014.06.001

Tyler Hicks Tyler Hicks Hicks, Tyler

Positivism

Positivism

1267

1270

Positivism

Positivism denotes a vague picture rather than precise thesis. In this picture, scientists accept the content of observable phenomena as *posits* in need of no further explication. Religion (i.e., theology) and metaphysics (i.e., traditional philosophy) try to describe the world “as it really is” rather than “as it appears to be,” with contested success. Yet, science yields agreed-upon knowledge, because it abandoned the quest to peek behind the veil of phenomena. In the early 20th century, educational researchers were trained to think about research like a positivist. They valued research that did not infer beyond observable phenomena. Since the 1960s, however, positivism has largely fallen out of favor among researchers (except among radical behaviorists) after a paradigm shift toward constructivist approaches occurred in education research. This entry provides an overview of positivism and discusses its legacy on educational research.

Empiricism

The roots of positivism stretch back as far as the British empiricists of the late 17th and early 18th centuries. The founder of empiricism, John Locke (1632–1704), claimed that rational agents had an ethical duty to arrange their beliefs like a skyscraper. As skyscrapers require a foundation on which their floors stand, beliefs should be stacked up so those on the “upper floor” could be inferable from those on the “lower floor.” Probable reasoning was claimed to suffice for justifying beliefs so long as the chain of justification eventuated in “foundational” beliefs. A belief needed to be intrinsically warranted (i.e., self-justifying) to pass as foundational.

Locke asserted that only beliefs known with certainty should qualify as foundational. He restricted this class of beliefs to those stemming from reflection on self-evident truths or the empirical senses. Self-evident truths included logical relations, such as, “It is either raining or it is not raining.” The empirical senses included taste, touch, smell, sight, and hearing. He considered these two sources of belief to be incorrigible (i.e., not correctable). For instance, a woman can mistake a tree for a person from afar. Yet, close analysis reveals the woman was actually led astray by an inference rather than observation. If the woman had not inferred (i.e., reported observing what *appeared* to be a person), she would not have needed correction.

Empiricists after Locke developed a psychology to match this bold epistemology. They hypothesized that any conceivable idea, regardless of its complexity, was built up from bundles of sensation. The idea of an apple, for example, was a composite of a variety of taste, visual, and other tactile sensations. The mind constructed the apple-idea by mental operations (e.g., habits of associating sensations). All entertained ideas, except perhaps logical and mathematical ones, originated in the rearrangement of sensations. Thus, an empiricist might say a person born blind could never conceive shades of color.

David Hume (1711–1776) proposed an empiricist test, typically known as “Hume’s fork,” to determine whether a claim was knowable *in principle*. To implement it, one evaluated whether the claim’s content consisted of ideas that had their ultimate source in pure sensation or reflection. If the claim did, then its truth or falsity was found to be knowable with inquiry. Otherwise, the claim referred to things beyond the limits of human knowledge. The test, of course, assumed an empiricist psychology. The purpose of this test, which later positivists liked, was to distinguish unknowable metaphysical speculation from plausible scientific hypotheses.

Classical Positivism

The French philosopher and social reformer Auguste Comte (1798–1857) coined the term *positivism* in the late 19th century. He believed that branches of science, in development, will pass through three successive stages: the theological, the metaphysical, and the positive. In other words, scientists progressed from the darkness of theorizing deities to the light of positive science, wherein they described nothing but regularities in phenomena. Metaphysics was only a transitory step between the theological and positive stages. In this in-between stage, impersonal principles, such as the Aristotelian essence or the Platonic form, did the explanatory work rather than a personal deity.

The classical approach to gravity illustrated Comte’s conception of scientific maturity. Recall that Isaac Newton (1642–1726) discovered that gravitational attraction conformed to an inverse square law. He gave no pretense to know the cause of gravity (e.g., he did not posit that matter had occult properties that produced gravity). Instead, he argued that his elegant law, which predicted all the relevant phenomena, sufficed. For Comte, then, scientists will, like Newton, renounce metaphysical explanation and find the fewest number of natural laws needed to predict the largest number of phenomena.

Comte claimed that the branches of science could be classified into six fundamental sciences: mathematics, astronomy, physics, chemistry, biology, and sociology. This order gives a locked sequence of maturation (i.e., sociology requires biology, biology requires chemistry). Comte thus adduced that sociology (another term he coined) would be science’s crowning achievement, as it inquired into the most complex of all phenomena—human society. Sociology’s arrival would constitute a major turning point in history. As the uppermost science, sociology would amplify the insights of all lower sciences and thus provide tools to remake the world.

Logical Positivism

Logical positivism is associated with the Vienna Circle, a selective group of scholars primarily trained in the natural sciences, philosophy, and mathematics. The group gathered in the early 1920s at the University of Vienna. While founded by Moritz Schlick (1882–1936), the group’s most prolific advocate was Rudolf Carnap (1891–1970). Taking their cue from Comte, they announced that metaphysics had no place in a scientific era. They also admired the work of logicians such as Ludwig Wittgenstein (1889–1951), which suggested to them that in a scientific era, the main task of philosophers would be to come alongside scientists and clarify the murky concepts embedded in science (e.g., unseen atoms, curved space time, fields).

Logical positivists considered *verifiability*, which had a strong and a weak sense, to be the dividing line between meaningful and meaningless claims. A claim was verifiable in the strong sense if scientists could physically verify it. A claim was verifiable in the weak sense if it was verifiable in principle. For instance, a scientist cannot verify a dinosaur’s actual color from the spotty fossil record. Yet, scientists can verify it in principle (e.g., if the requisite data had been preserved in the fossil record). A specific claim about a dinosaur’s color is thus verifiable in the weak sense but not the strong sense.

Verification can further be subdivided into *definitive proof* and *probable evidence*. A black swan constitutes definitive proof that not all swans are white. A sample of white swans (provided no black swans were yet seen) only counts as probable evidence for the claim that all swans are white. The weight of evidence can also vary for a claim. Some evidence supports a claim more strongly than others. Logical positivists never satisfactorily parceled out the full implications of the plurality of forms that verification can acquire in science. Disagreements on the technical details haunted the logical positivists.

However, there was general agreement among them that the meaning of a statement was its method of verification. If a claim was not verifiable, then it was worse than false—its content was meaningless. The verification criterion went beyond Hume’s fork. Hume had only declared metaphysics to be speculative. In contrast, the logical positivists charged metaphysicians with embarrassingly mistaking meaningless claims for meaningful ones.

They buttressed their verification criterion in the analytic–synthetic distinction. This distinction stated that the content of intelligible statements had logical and factual components. The synergy between these components constituted the meaning of a claim. When a sentence’s truth depended only on its logical content, it was analytic; if its truth also depended on its factual content, it was synthetic.

A sentence verifiable under all empirical circumstances, these positivists argued, must be analytic in nature. For instance, a bachelor can be known to be unmarried no matter what (i.e., it seems true by definition). Such a claim then seems to be analytic. Conversely, if empirical inquiry was required to know that a claim was true or false, positivists said it must be synthetic. For instance, the claim “all bachelors are bald” seems synthetic. Its truth depends on conditions in the world. If no empirical or logical method could verify a claim, it was hence meaningless.

The analytic–synthetic distinction, supplemented with the verification criterion, was designed to eliminate pseudoscience (i.e., metaphysics, theology). Logical positivists predicted that metaphysics would soon become extinct. Perhaps ethics could survive in a scientific era because such talk may be valued as an emotional outlet like instrumental music. For example, the claim "murder is wrong" could be used to express emotional disgust rather than state a moral fact. However, for logical positivists, most of the bits of metaphysics and theology, which had no similar emotional value, would have to go.

The Positivist Legacy on Educational Research

Positivism has had an important influence on the history of educational research. In the early 20th century, positivism supplanted pragmatism in the United States as the prevailing ideology among progressive educational thinkers. It provided them with a definition of valuable research, standards for posing their questions, and agreed-upon methods for answering them. For instance, the writings of Edward Thorndike (1874–1949) actively discouraged researchers from inferring a cause beyond what was found in measurable data. Positivist thinking was codified in the behaviorist maxim that a scientist observes rather than infers.

However, the backlash against logical positivism in the second half of the 20th century is also an important piece of the story. Most leading positivists by the 1960s had recognized that the verification criterion could not be implemented without wreaking havoc on science as actually practiced. They had hoped that it would enclose all science but no metaphysics. Yet, drawing such a tight circle proved an intractable problem. If they calibrated their criterion too narrowly, it excluded bits of science as well as metaphysics. An unseen atom, for example, might have to be rejected as pure nonsense. If the logical positivists relaxed the criterion to permit unseen atoms (e.g., they were pragmatically useful computational devices or a shorthand for observable phenomena), then some metaphysics might creep into science too.

In the early 1950s, W. V. O. Quine (1908–2000) attacked the verification criterion. He showed that the distinction between analytic and synthetic claims was actually conventional rather than principled. For instance, scientists can make any claim behave like an "analytic" truth with the assistance of ad hoc hypotheses (e.g., by positing measurement error to save it from negative findings) or behave like a "synthetic" truth (e.g., by changing definitions of key terms to falsify it). Quine's widely conceded epiphany constituted a logical defeat for logical positivism.

Given its past prominence, many critics of positivism have used positivism—or a caricature of it—as their main foil for ideological contrasting. Despite what is sometimes claimed, however, actual positivists were rarely committed to the idea that science gave a “true description of reality” (i.e., absolute truth) or that only quantitative research was acceptable. What really defined positivism was a sweeping grand narrative about the apocalyptic arrival of a scientific era, wherein the scientific method would be the only source of public knowledge and enlightened researchers would no longer infer beyond observable phenomena. Since the 1960s, a critical mass of education researchers has rejected positivism in favor of more constructivist approaches. Consequently, researchers have subsequently had to operate in an intellectual environment, wherein many philosophies, besides positivism, compete for their allegiance.

Tyler Hicks

See also Behaviorism; Constructivist Approach; Epistemologies, Teacher and Student; Paradigm Shift; Postpositivism

Further Readings

Ayer, A. J. (1952). Language, truth, and logic. New York, NY: Dover.

Comte, A. (1970). Introduction to positive philosophy (F. Ferre, Trans.). Indianapolis, IN: Bobbs-Merrill. (Original work published 1830) Friedman, M. (1999). Reconsidering logical positivism. New York, NY: Cambridge University Press.

Lagemann, E. C. (2002). An elusive science: The troubling history of education research. Chicago, IL: University of Chicago Press.

Paul, J. L. (2005). Historical and philosophical influences shaping perspective on knowledge. In J. L. Paul (Ed.), Introduction to the philosophies of research and criticism in education and the social sciences (pp. 1–20). Upper Saddle River, NJ: Pearson.

Phillips, D. C. (1992). Positivism. In D. C. Phillips (Ed.), The social scientist's bestiary (pp. 95–105). New York, NY: Oxford: University Press.

Quine, W. V. O. (1951). Two dogmas of empiricism. The Philosophical Review, 60, 20–43. doi:10.2307/2266637

Rebecca Tipton Rebecca Tipton Tipton, Rebecca

Grant B. Morgan Grant B. Morgan Morgan, Grant B.

Post Hoc Analysis

Post hoc analysis

1270

1273

Post Hoc Analysis

Post hoc analysis, or *a posteriori* analysis, generally refers to a type of statistical analysis that is conducted following the rejection of an omnibus null hypothesis. Post hoc analysis can be conducted for a variety of statistics including proportions and frequencies, but post hoc analysis is most commonly used for testing mean differences. This entry focuses primarily on post hoc analysis for investigating mean differences. Following are brief discussions of the omnibus test under a one-way analysis of variance (ANOVA), Tukey honestly significant difference test for pairwise contrasts, Scheffé’s procedure for pairwise or complex contrasts, other available post hoc procedures, and considerations for using post hoc analysis.

Omnibus Test

An omnibus test, from root *omnis*, meaning “for all,” is a kind of statistical test that simultaneously tests multiple null hypotheses. An omnibus test should be conducted when the researcher does not have specific planned, or a priori, comparisons of theoretical or applied interest. The one-way ANOVA is perhaps the most common and simplest example of an omnibus test. The omnibus test for the one-way ANOVA states that (or tests if) the group means are simultaneously equal to one another. Consider an experimental design with three groups, where the groups represent control, Treatment A, or Treatment B group. Under the null hypothesis (i.e., H_0), the omnibus test states that all group means are equal (see

Equation 1):

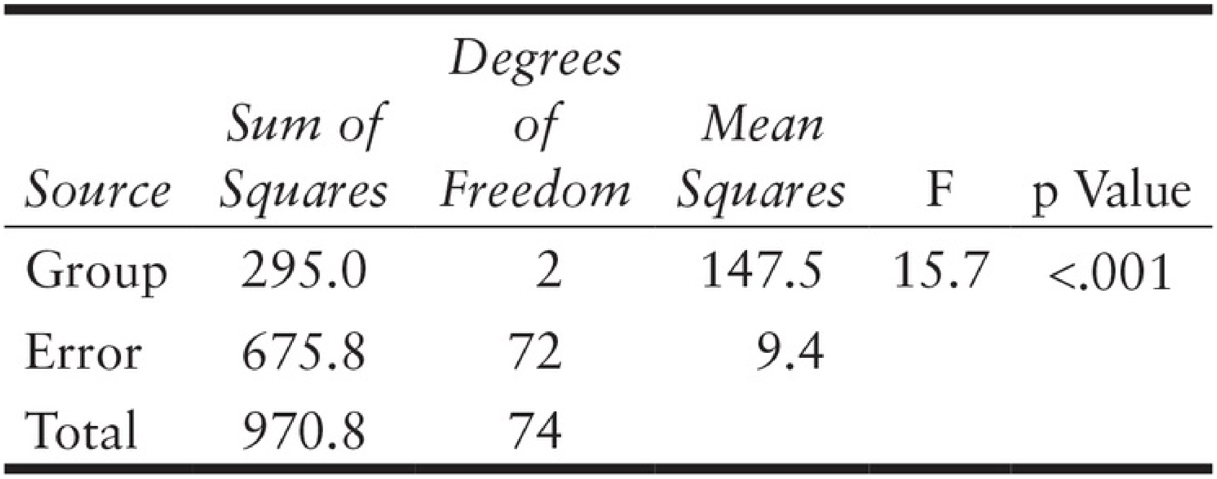
$$H_0 : \mu_{\text{Control}} = \mu_{\text{Treatment A}} = \mu_{\text{Treatment B}} \cdot$$

This hypothesis is tested using a preestablished Type I error rate (α), which is commonly set at .05 in the social sciences. If the null hypothesis is rejected, it is not evident how the null hypothesis is false. That is, the omnibus test does not indicate which means are different from each other. It is possible, for example, that the control group mean is different from both the Treatment A group mean and the Treatment B group mean, but the means of the treatment groups are not different from one another. Post hoc analysis is conducted to explore which means are different. In this sense, post hoc analysis can be considered an exploratory procedure.

Tukey Test for Pairwise Mean Comparisons

Following a rejected omnibus test, a researcher may opt to compare each group mean with every other group mean. These types of comparisons are called pairwise comparisons. The most common procedures for conducting all pairwise comparisons are the Tukey honest significant difference (HSD) test or Tukey–Kramer method. Tukey HSD test allows a researcher to make all pairwise comparisons among group means when each group has the same number of participants. The Tukey–Kramer method allows a researcher to make all pairwise comparisons among group means when the groups do not have the same number of participants. In practice, the groups frequently are not made up of the same number of participants.

From the earlier hypothetical example, the results of the omnibus ANOVA null hypothesis test are shown in Table 1. The group means and standard deviations were made up for this example. The mean and standard deviation, respectively, were 9.8 and 3.46 for the control group, 13.9 and 3.69 for Treatment A group, and 14.1 and 1.6 for Treatment B group.



Source	Sum of Squares	Degrees of Freedom	Mean Squares	F	p Value
Group	295.0	2	147.5	15.7	<.001
Error	675.8	72	9.4		
Total	970.8	74			

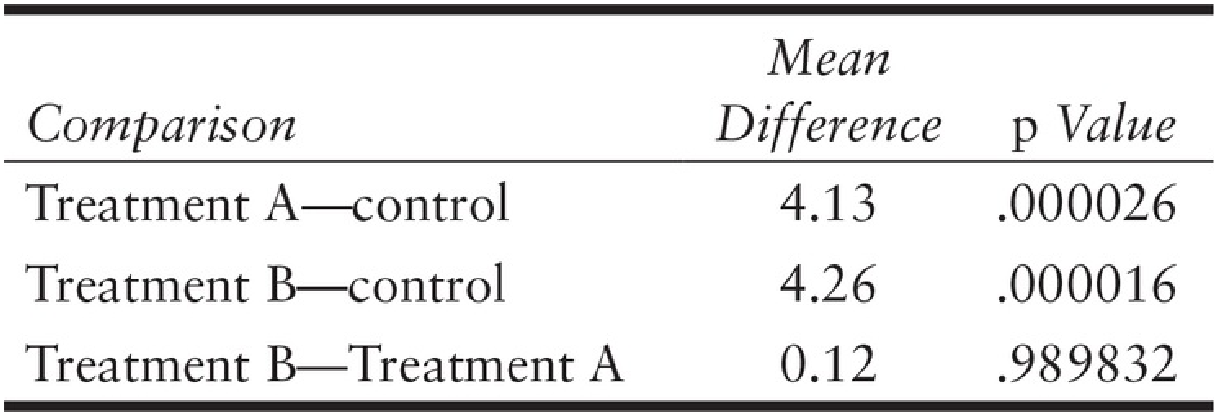
Assuming the researcher used a preestablished maximum allowable Type I error rate (α) of .05, the null hypothesis can be rejected because the *p* value ($p < .001$) is less than the Type I error rate. As noted earlier, it is not clear where the differences exist between group means. The hypothetical data used in the omnibus test were made to have 25 participants in each group; therefore, the Tukey HSD procedure can be used to conduct the post hoc analysis. Post hoc analysis is typically conducted using statistical analysis software, but it can be done by hand as well. The critical value used for the Tukey HSD procedure is based on the studentized range distribution, *Q*. Many *Q* distribution value calculators are available online, and *Q* distribution value tables can also be found in many statistics textbooks. The necessary pieces of information for finding the relevant *Q* value are the number of groups (i.e., three in this example) and the Error degrees of freedom (i.e., $df = 72$) from the ANOVA table. The associated *Q* value from a statistics textbook is 3.39 when the Type I error rate is .05. The critical value of *Q* can then be computed using Equation 2:

$$Q_{cv} = Q \times \sqrt{MS_{\text{Error}}/n},$$

where *Q* is the relevant value of the *Q* distribution, MS_{Error} is the mean square estimate of the Error term in the ANOVA table, and *n* is the sample size of each group. In this example, the critical value is

$$Q_{cv} = Q \times \sqrt{MS_{\text{Error}}/n} = 3.39 \times \sqrt{9.4/25} = 2.08.$$

This is a value against which each pairwise difference can be compared in order to determine whether it is statistically significant. The difference between the Treatment A group mean (13.9) and the control group mean (9.8) is 4.1 (13.9 − 9.8 = 4.1). This difference exceeds 2.08 so the researcher can conclude with 95% that the two means are statistically different. This process would be repeated for comparing Treatment B group mean against the control group mean as well as for comparing the group means of the two treatment groups. The results of the Tukey HSD from a statistical analysis software are shown in Table 2. The pairwise comparison shown above corresponds to the first comparison in Table 2. Reviewing the *p* value column of Table 2 indicates that two pairwise comparisons are statistically significant because two are less than .05. These mean differences are the reasons why the omnibus test from the one-way ANOVA was rejected.



Comparison	*Mean Difference*	p *Value*
Treatment A—control	4.13	.000026
Treatment B—control	4.26	.000016
Treatment B—Treatment A	0.12	.989832

Note: HSD = honest significant difference.

Scheffé’s Test for Pairwise or Complex Comparisons

The Scheffé’s test can be used for making any post hoc comparison—pairwise or complex. As noted earlier, pairwise comparisons are comparisons of means between intact groups. Complex comparisons involve two or groups being aggregated into one mean and compared against the mean of one group or the mean of other aggregated groups. Following are examples of a pairwise contrast and a complex contrast. The process is the same for both types of contrast, and the critical value for the Scheffé’s test will be the same for both because it is based on the critical value and degrees of freedom from the omnibus test. The critical value of the Scheffé test is shown in Equation 3.

$$CV_{\text{Scheffé}} = \sqrt{df_{\text{Group}} \times F_{(df_{\text{Group}}, df_{\text{Error}}, 1-\alpha)}},$$

where df_{Group} is the degrees of freedom for the Group variable in the omnibus ANOVA table and $F_{(df\text{Group},\ df\text{Error},\ \backslash\ -\ \alpha)}$ is the critical value from the omnibus test. The F critical value can be found in an online calculator, statistical software package, or statistics textbook. For this example, the critical value of the omnibus test is the value of the F distribution with 2 degrees of freedom for Group, 72 degrees of freedom for the Error, and Type I error rate of .05; this value is 3.12. Thus, the critical value for the Scheffé can be computed using Equation 3:

$$CV_{\text{Scheffé}} = \sqrt{df_{\text{Group}} \times F_{(df_{\text{Group}}, df_{\text{Error}}, 1-\alpha)}}$$

$$= \sqrt{2 \times 3.12} = 2.50.$$

Pairwise Comparisons via Scheffé

To illustrate a pairwise contrast using the Scheffé test, the mean of the control group and the mean of the Treatment A group are compared. The difference between the Treatment A group mean (13.9) and the control group mean (9.8) is 4.1 (13.9 − 9.8 = 4.1). Like all test statistics, this mean difference must be divided by the standard error, which for a pairwise comparison is equal to

$$\hat{\sigma}_{\psi} = \sqrt{MS_{\text{Error}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

where MS_{Error} is the mean square of the Error term in the omnibus ANOVA table, $n_1$is the sample size of the first group being compared, and n_2 is the sample size of the second group being compared. In the current example, the estimated standard error is

$$\hat{\sigma}_{\psi} = \sqrt{MS_{\text{Error}}\left(\frac{1}{n_{\text{Treatment A}}} + \frac{1}{n_{\text{Control}}}\right)}$$

$$= \sqrt{9.4\left(\frac{1}{25} + \frac{1}{25}\right)} = 0.87.$$

The mean difference divided by the standard error is

$$\frac{4.1}{0.87} = 4.73.$$

This value, 4.73, exceeds the Scheffé critical value of 2.50 so the researcher can conclude with 95% that the two means are statistically different.

Complex Contrast via Scheffé

To illustrate a complex contrast, the two treatment groups will be aggregated and compared against the control group. This type of comparison is done using weights or contrast coefficients. A contrast is defined as in Equation 4

$$\psi = \sum_{k=1}^{K} w_k \mu_k ,$$

where k represents each group, w_k represents the weight given to group k, and μ_k is the mean of group k. The sum of the weights must total zero in a contrast. To aggregate the two treatment group means, one could add the means together and divide by two. Alternatively, one could simply multiply each of the group means by 0.5 (i.e., contrast weights) and then add the halves together. To find the difference between this aggregated mean and the mean of the control group, the control group mean would need to be multiplied by −1 (i.e., contrast weight). Applying these weights forms the contrast because the sum of the weights equals zero (0.5 + 0.5 + −1 = 0). The contrast can then be estimated as

$$\hat{\psi} = \sum_{k=1}^{K} w_k \hat{\mu}_k = (0.5 \times 13.9) + (0.5 \times 14.1) + (-1 \times 9.8$$

$$= 14.0 - 9.8 = 4.2.$$

The difference between the aggregated treatment group mean (14.0) and the control group mean (9.8) is 4.2 (14.0 − 9.8 = 4.2). This difference must be divided by the standard error, which for a complex comparison is shown in Equation 5:

$$\hat{\sigma}_{\psi} = \sqrt{MS_{\text{Error}} \sum_{k=1}^{K} \frac{w_k^2}{n_k}},$$

where MS_{Error} is the mean square of the Error term in the omnibus ANOVA table, n_k is the sample size of the kth group. In the current example, the estimated standard error is

$$\hat{\sigma}_{\psi} = \sqrt{MS_{\text{Error}} \left(\frac{w_1^2}{n_{\text{Treatment A}}} + \frac{w_2^2}{n_{\text{Treatment B}}} + \frac{w_3^2}{n_{\text{Control}}} \right)}$$

$$= \sqrt{9.4 \left(\frac{.5^2}{25} + \frac{.5^2}{25} + \frac{-1^2}{25} \right)} = \sqrt{9.4 \left(\frac{.25}{25} + \frac{.25}{25} + \frac{1}{25} \right)}$$

$$= 0.75.$$

The mean difference divided by the standard error is

$$\frac{4.2}{0.75} = 5.6.$$

This difference exceeds the Scheffé’s critical value of 2.50 so the researcher can conclude with 95% that the aggregated treatment mean is statistically different from the control group mean. *Any* post hoc contrast can be performed using the same process.

Other Common Post Hoc Procedures

There are many post hoc analyses for comparing mean differences between groups. As noted earlier, the Tukey–Kramer procedure can be used when the groups being compared have different numbers of participants. The Brown–Forsythe procedure can be used when the variances of the groups being compared are not assumed to be equal (discussed more in the following section). Newman–Keuls is a procedure very closely related to the Tukey HSD but uses a slightly different critical value. Each of these procedures is readily available in most statistical software packages.

Considerations for Using Post Hoc Analyses

With so many options, researchers may feel overwhelmed by technical detail when trying to make an informed decision. Following are some considerations. If all pairwise comparisons are of interest, one might consider opting immediately for the Tukey HSD or Tukey–Kramer test instead of conducting an omnibus test. Either can be conducted post hoc as well. The equality of the sample sizes will determine which should be conducted. The assumptions for the Tukey procedures are that the errors are independent within and between groups, normally distributed, and have equal variance across groups. When the normality assumption is not tenable, Brown–Forsythe can be considered. To conduct any post hoc comparison, the Scheffé method can be used. It makes the same assumptions about the error distributions as Tukey.

Ultimately, these procedures provide different mechanisms for controlling Type I error. If making all pairwise comparisons, Tukey provides the most statistical power. Scheffé allows any type of comparison to be made. Neither is able to provide a one-sided test of statistical significance. One-sided tests should be handled with a priori contrasts with Bonferroni adjustment (not discussed here).

All post hoc analyses discussed here assume normally distributed errors. Although these procedures are somewhat robust against violations of nonnormality, if this assumption is not tenable, a researcher should consider using a nonparametric alternative, which conceptually mirrors what has been presented here. If the assumption of independently distributed errors is not tenable, a different statistical approach must be taken to account for the tenable, a different statistical approach must be taken to account for the correlated errors (e.g., hierarchical linear model and random effects ANOVA).

Rebecca Tipton and Grant B. Morgan

See also Analysis of Variance; Bonferroni Procedure; Experimental Designs; *F* Distribution; Power; Type I Error

Further Readings

Kirk, R. E. (2013). Experimental design: Procedures for the behavioral sciences (4th ed.). Thousand Oaks, CA: SAGE.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experimental and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

Meghan K. Cain Meghan K. Cain Cain, Meghan K.

Zhiyong Zhang Zhiyong Zhang Zhang, Zhiyong

Posterior Distribution Posterior distribution

1273

1275

Posterior Distribution

In Bayesian analysis, the *posterior distribution*, or *posterior*, is the distribution of a set of unknown parameters, latent variables, or otherwise missing variables of interest, conditional on the current data. The posterior distribution uses the current data to update previous knowledge, called a *prior*, about that parameter. A posterior distribution, $p(\theta|\boldsymbol{x})$, is derived using Bayes's theorem

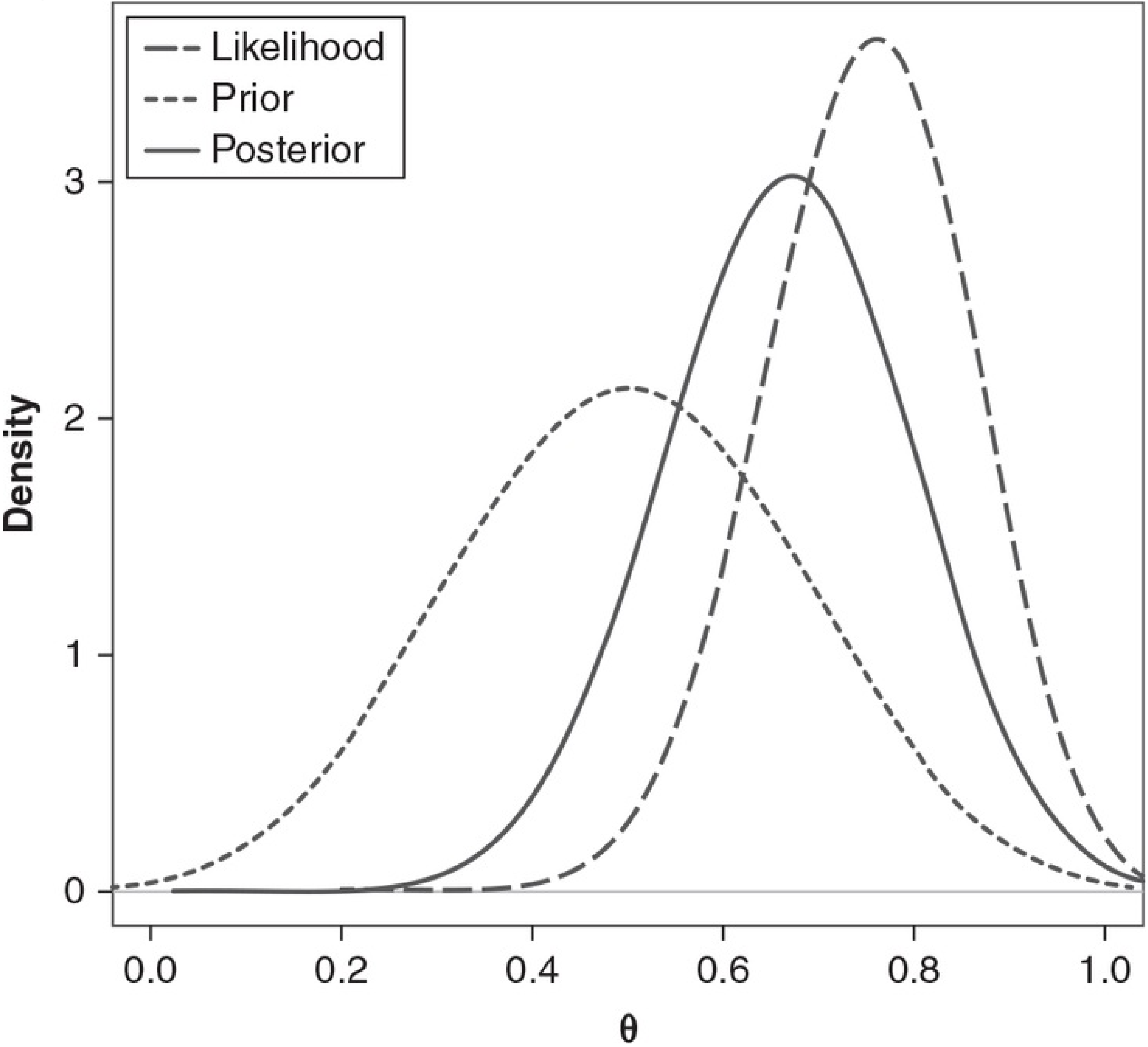
$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{p(x)}=\frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta},$$

where θ is the unknown parameter(s) and $\boldsymbol{x}$ is the current data. The probability of the data given the parameter $p(\boldsymbol{x}|\theta)$ is the likelihood $L(\theta|\boldsymbol{x})$. The prior distribution, $p(\theta)$, is user specified to represent prior knowledge about the unknown parameter(s). The last piece of Bayes's theorem, the marginal distribution of data, $p(\boldsymbol{x})$, is computed using the likelihood and the prior. The distribution of the posterior is determined by the distributions of the likelihood and the prior and scaled by the marginal distribution of the data. Therefore, the posterior can be represented as

$$\text{Posterior distribution} \propto \text{Likelihood} \\ \times \text{Prior distribution},$$

where ∝ means "proportional to." The relationship between the posterior, the prior, and the likelihood is shown in Figure 1.

Figure 1 The likelihood and the prior determine the posterior distribution



The prior distribution is *conjugate* to the likelihood if the resulting posterior distribution has the same form as the prior distribution. The mean and variance of the posterior distribution are also determined by these two distributions. In certain situations, the posterior mean is a weighted average of the mean of the data and the prior, using the precision of each as weight. The precision, the reciprocity of variance, of the posterior is a function of the precision of the data and the prior. Thus, when a researcher is more confident in a prior, it is given more weight by specifying a smaller variance for the prior distribution.

The posterior distribution can be analytically computed by integration or it can be approximated using a Markov chain Monte Carlo algorithm. With increases in computational power, the latter is often the easier option, and the Markov chain Monte Carlo method is what is used in software such as WinBUGS and Mplus. A commonly used Markov chain Monte Carlo method is *Gibbs sampling*, which recursively generates random numbers from the conditional posterior distribution for each parameter in turn, conditional on the current values of all other parameters.

The resulting posterior distribution is what is used to make inferences about the model. The mean, median, or mode of the posterior distribution can be used as a point estimate, much like a maximum likelihood estimate (MLE) can be used within the frequentist framework. If the prior $p(\theta)$ is a constant, the mode of the posterior, if it exists, is equivalent to the MLE. *Credible intervals* can also be constructed using the posterior distribution. These are analogous to confidence intervals in the frequentist framework but differ in theory and interpretation. Credible intervals provide the $(1 - \alpha)\%$ probability that a parameter lies between a lower and upper bound. Thus, credible intervals assume the parameter is random and the lower and upper bounds are fixed, whereas confidence intervals assume the opposite.

Example

To illustrate, let's say Researcher F finds a coin in his attic. He wants to know whether the coin is fair, so he flips it 20 times and records 15 heads landings. He is a frequentist, so he would like to find an MLE of the probability of the coin landing on heads. First, he computes the likelihood using a binomial distribution,

$$L(\theta \mid x) \sim Bin(n,p) = \theta^{k} (1-\theta)^{n-k} ,$$

where *n* is the number of tosses, *p* is the probability of landing on heads, and *k* is the number of heads landings ($n \times p$). This results in an MLE of . Using a binomial test, he concludes that the coin is not fair, $z = 2.236$, $p < .05$.

The next day, Researcher F tells his colleague, Researcher B, about the coin he found that lands on heads significantly more often than it lands on tails. In response, Researcher B says "Wait a minute, shouldn't we take into account that most coins are fair?" She would like to use Bayesian analysis to investigate further by choosing a prior that is centered at 50% with small variance because she knows that most coins land on each side with equal probability. Because the β distribution is conjugate to the binomial likelihood function, she would like to use a prior with the following form:

$$p(\theta) \sim \beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

She chooses $p(\theta) \sim \beta(5,5)$, which is a symmetrical distribution with mean = 0.5 and variance = 0.007. Using this prior with the binomial likelihood, she calculates the posterior distribution for the proportion of heads,

$$p(\theta|x) = \frac{L(\theta|x)p(\theta)}{\int_0^1 L(\theta|x)p(\theta)d\theta}$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a+k)\Gamma(b+n-k)} \theta^{a+k-1} (1-\theta)^{b+n-k+1}.$$

Because she used conjugate distributions, she gets a β distribution with parameters $a + k$ and $b + n - k$. Plugging in a and b from the prior and n and k from the data, she obtains β(20,10) as the posterior. Using the mean of the posterior, the new point estimate of the fairness of the coin is , slightly less extreme than it was for Researcher F. The credible interval is [0.49,0.82], and so we are not quite sure whether the coin lands on heads more than tails.

If Researcher F had instead flipped the coin 100 times and gotten 75 heads, using the same prior would have yielded the posterior β(80,30) with point estimate . This is because getting 75 heads out of 100 tosses is much stronger evidence against the coin being fair than getting 15 heads out of 20 tosses. In this case, the data are given more weight in calculating the posterior, so our point estimate is closer to that of the data. The credible interval in this case would be [0.64,0.81], indicating that there is a 95% probability that the coin is not fair.

In sum, the posterior distribution is proportional to the product of the likelihood and the prior in Bayesian analysis, and it can be used to make inferences about model parameters. Any inference made with the posterior is based on prior information about a model that has been updated after collecting new data. Inferences can be made using the mean, median, or mode of the posterior as inferences can be made using the mean, median, or mode of the posterior as point estimates or through credible intervals, among other methods.

Meghan K. Cain and Zhiyong Zhang

See also Bayes’s Theorem; Bayesian Statistics; Binomial Test; Markov Chain Monte Carlo Methods; Prior Distribution

Further Readings

Gelman, A., Carlin, J. B., & Stern, H. S. (2003). Bayesian data analysis. London, UK: Chapman Hall.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.). Cambridge, MA: Academic Press.

Lee, P. M. (2012). Bayesian statistics: An introduction (4th ed.). Hoboken, NJ: Wiley.

Tyler Hicks Tyler Hicks Hicks, Tyler

Postpositivism

Postpositivism

1275

1278

Postpositivism

Postpositivism subsumes a plurality of epistemological stances intended to supersede positivism without requiring objective knowledge to succumb to epistemological anarchy (i.e., “anything-goes” relativism). Postpositivists participate in two levels of debate: The first pits them against positivists, the second against relativists (e.g., postmodernists and certain constructivist approaches). In contrast to positivists, they are not classical foundationalists, who claim that knowledge needs a secure foundation. In contrast to relativists, they acknowledge that scientists put forth claims to truth that are warranted despite being fallible. Postpositivism thus offers researchers another choice besides positivism or relativism. This entry gives an overview of postpositivism and provides examples of postpositivist epistemologies.

Fallibility and Warrants

Positivists sought to ground science in an incorrigible (uncorrectable) source of knowledge (e.g., sense data and logical truths). They held that there was a duty to build up the knowledge base on a foundation of self-justifying beliefs. They also believed scientists should never go beyond observable phenomena. However, not even the best-entrenched claims in the natural sciences were found to fully meet these conditions. For example, physicists invoke atoms in their theories even though such entities must be inferred from observable phenomena. The content of most theories in science (except perhaps in pure mathematics) would thus need to be substantially altered to conform to strict positivist standards.

Postpositivists, in contrast, champion fallible knowledge (i.e., warranted truth claims can originate from a purely fallible source). For example, people seem to know what they ate for breakfast even though they have faulty memories. A passable epistemology of science as actually practiced, including education research, needs to recognize that scientists rely on fallible sources of information. Natural and social science have historically flourished despite the repeated failure of philosophers to ground it in only incorrigible truths.

Postpositivists further claim that knowledge can be objective without the need for absolute certainty. Objective knowledge has two senses: In the strong sense, it implies that humans can know about objects that exist independently of individual or social conceptions of reality. Such objective knowledge thus contrasts with subjective knowledge, which is limited to the content of one’s own mind (e.g., whether vanilla ice cream tastes good). In the weak sense, objective knowledge is based on intersubjective agreement, in which agents who engage in competent inquiry should converge in their beliefs. Such inquiry can yield warranted assertions. In this weaker sense, “objective” contrasts with being biased rather than subjective. Bias (e.g., racism, sexism, and classism) is a tendency to arrive at conclusions that support a vested interest, regardless of what the obtained evidence suggests. Scientists always embody a subjective outlook, but they need not be biased.

Postpositivists disagree about whether science can recover strong objective knowledge. However, they typically agree that there are objective criteria for deciding what constitutes a warranted truth claim. They also concur that science comes from a fallible source. To reconcile the concepts of fallibility and warrant in a single position, postpositivists turn to one of three approaches: evolutionary epistemology, realism, and pragmatism.

Evolutionary Epistemology

Karl Popper (1902–1994) provides a vivid example of an evolutionary epistemologist. He likened theory selection in science to natural selection. In the Darwinian model, natural selection adapts species to environments without foresight. Some random mutations confer advantages. Animals lucky enough to inherent these survive longer than their evolutionary competitors, and thus, they have more offspring. This process naturally adapts species to their environments.

Popper proposed that science advances toward truth in a process akin to evolution. There is no logic of proposing true theories; rather, scientists simply propose whatever theory appeals to them. For instance, Dmitri Mendeleev (1834–1907) acquired an idea for the periodic table from a dream. Although only a few of the theories ever proposed in science will likely be true, scientists have a way to identify those lucky winners.

There is a logic of theory selection. Popper noticed a disparity between the process of verification and falsification: It is difficult to verify, easy to refute. Consider a claim such as “All swans are white.” To verify it, one would need to examine every swan. But to falsify it, one would only need to find a single black swan—a relatively easier feat. Consequently, a scientist’s best strategy is to seek to falsify theories rather than verify them.

Popper used falsification to reconcile fallibility and warrant. He argued that scientists could achieve weak objective knowledge (i.e., intersubjective agreement) without an infallible source of information. Popper used the notion of strong objective knowledge (i.e., true description of reality) as only a regulative ideal in science. In his theory of increasing verisimilitude, Popper hoped a Darwinian process of competition among theories would ensure that proposed theories that were better approximations of reality would continuously replace their predecessors. The result would be that science could sustain an upward spiral toward the truth.

Realism

Realism is premised on the supposition that the aim of science—or at least of theorizing—is to describe the world “as it really is.” In this picture, the world is external of the human mind and the purpose of science is to grasp its ontology (theory of things) in its own terms. To borrow a metaphor from Plato’s writings, realists long for theories that carve up the world at the joints. They think a successful theory is likely true, but they further think that its ontology likely maps reality’s structure.

The epistemology underpinning realism is murky, yet perhaps a revised foundationalist’s approach can illuminate it. Classical foundationalists restricted foundational beliefs to only those originating from incorrigible sources (e.g., sense data), but this restriction may have been a fatal mistake. Realists may presume that other types of beliefs should be included. For instance, people seem to be within their rights to believe that other people have minds, even though it cannot be successfully justified with arguments from incorrigible beliefs.

The question of which beliefs count as foundational may depend on external factors (i.e., those beyond the mere content of entertained beliefs). Developmental psychologists have shown that young children acquire certain beliefs about the world (e.g., object permanence) at predictable stages of maturation. Philosophers such as Alvin Plantinga (1932–) propose that many instances of true belief are being automatically outputted by human cognitive faculties when they are (a) properly functioning, (b) situated in a suitable environment, and (c) competently designed to output true beliefs. Scientists may thus approach inquiry about the world with a richer stock of foundational beliefs than most positivists were prepared to recognize. Realists are optimistic that scientists, including education researchers, can succeed in the task of recovering strong objective knowledge. Realists thus may interpret scientific theories as literal attempts to objectively describe reality.

Pragmatism

Pragmatism offers a more modest alternative to realism. It is premised on the supposition that truth is not really a viable criterion for belief. Pragmatists argue that there is no way to implement a criterion for truth, as scientists cannot recognize when beliefs match up with reality. It is a modern truism that evidence underdetermines theory selection. Just as an infinite number of lines can always be drawn to connect a finite number of dots, absurd theories fitting all of the available evidence can easily be produced. For example, the conjecture that the world was created 5 minutes ago with memories intact is compatible with all evidence.

Thankfully, other criteria besides a theory’s empirical adequacy exist (e.g., simplicity, better predictions, and consistency with other accepted theories). To illustrate this point, the ancient Greeks posited gods and goddesses to explain weather patterns, and although this theory has since fallen out of favor, it has never been disproven. Scientists could continue to invoke a bad-tempered god to explain a particularly nasty storm, if they wished. Given enough ad hoc hypotheses, they could even “fit” all available evidence to it. But modern meteorologists forgo undertaking such heroic efforts because they have formulated alternate theories that have more pragmatic advantages (e.g., better predictions).

Scientists select theories after balancing their pragmatic virtues, such as simplicity or increased predictability. All else being equal, scientists prefer simple theories over convoluted ones, even though there is no guarantee the truth will be simple. In the broadest sense, these types of pragmatic considerations ultimately prevail in science. Scientists have no choice but to select theories by evaluating their consequences. If two theories yielded the same consequences, scientists would have no reason to prefer one over the other. This sentiment is best captured by an apt statistical maxim: All models are false but some are useful.

Pragmatism—at least when in the hands of postpositivists—can reconcile the two concepts of fallibility and warrant. To pragmatists, science, given its self-correcting method, is a democratic form of inquiry. It attracts adherents not by brute power but by solving practical problems. Thus, science may be the best way to publicly “fix” beliefs in a free society. It can yield intersubjective agreement by working backward (i.e., evaluating terminal consequences rather than examining starting principles).

Legacy of Postpositivism in Education Research

The term *postpositivism* is a bit unfortunate. It can wrongly suggest that postpositivism is just an evolved form of positivism. Postpositivism is better interpreted as a “catch-all” phrase encompassing positions that coherently affirm fallibility and warrant. It can thus subsume evolutionary epistemologists, realists, and pragmatists even though these positions otherwise have little else in common; they all equally undermine positivism and relativism. The importance of postpositivism, then, is in its provision of a third epistemological option for education researchers.

The postpositivist sales pitch as a “moderate position” between the two extremes of positivism and relativism gives it broad appeal. In 2002, a version of postpositivism was even codified in the National Research Council’s report, *Scientific Research in Education*, which was intended to inform politicians and policy makers in the United States about how to support and develop a science of education. The conception of science embodied in that report meets the presumed definition of postpositivism in that it is neither positivist nor relativist. The report, which took a “least common denominator” position, subsequently guided many federal funding decisions and exerted an influence on education research.

As expected, the report has been criticized by those not subscribing to a postpositivist outlook. However, it can hardly be labeled “the postpositivist approach” without some further qualification because its outlook so strongly gravitates toward a Popperian account of science. For instance, it valued qualitative research only to the extent it contributes to proposing hypotheses that can be tested with more rigorous designs. Postpositivists such as realists and pragmatists, who may not consider the logic of falsification to be so definitional of good science, may hence fault this report.

Postpositivism, regardless of its particular formulation, provides researchers with an intriguing vision of a science of education. Three things seem to follow from this conception of science: (1) there is a methodological continuity between natural science and social science (including education research); (2) the debate among positivists, relativists, and postpositivists is not really about method so much as it is about what constitutes valuable knowledge; and (3) it is imperative to recognize, according to postpositivists, that science can yield warranted truth claims without a solid foundation. Knowledge and certainty need not overlap with one another. One can have knowledge without certainty. In other words, scientists can objectively know something even though they cannot prove that they know it on the basis of absolutely certain truths.

Tyler Hicks

See also Constructivist Approach; Epistemologies, Teacher and Student; Objectivity; Positivism; Pragmatic Paradigm

Further Readings

Biesta, G. J. J. (2003). Pragmatism and educational research. Lanham, MD: Rowman & Littlefield.

Hacking, I. (1983). Representing and intervening: Introductory topics. New York, NY: Cambridge University Press.

House, E. R. (1993). Realism in research. Educational Researcher, 20, 2–9.

National Research Council. (2002). Scientific research in education. In R. J. Shavelson & L. Towne (Eds.), Committee on scientific principles for education research. Washington, DC: National Academies Press.

Phillips, D. C., & Burbules, N. C. (2000). Postpositivism and education research. Lanham, MD: Rowman & Littlefield.

Plantinga, A. (1993). Warrant and proper function. New York, NY: Oxford University Press.

Popper, K. (1963). Conjectures and refutations: The growth of scientific knowledge. New York, NY: Routledge.

Grant B. Morgan Grant B. Morgan Morgan, Grant B.

Rachel L. Renbarger Rachel L. Renbarger Renbarger, Rachel L.

Posttest-Only Control Group Design Posttest-only control group design

1278

1281

Posttest-Only Control Group Design

The posttest-only control group design is a research design in which there are at least two groups, one of which does not receive a treatment or intervention, and data are collected on the outcome measure after the treatment or intervention. The group that does not receive the treatment or intervention of interest is the control group. The general process for this design is that (a) two or more groups are formed; (b) the treatment or intervention is administered; (c) data are collected after the treatment or intervention has been administered, commonly using a behavioral, cognitive, or psychological assessment; and (d) the data are compared between groups to determine whether the treatment or intervention was effective. The goal of this design is often to make causal inferences; that is, to draw conclusions about whether or not a difference between groups (i.e., the effect) is observed as the result receiving the intervention (i.e., the cause). This entry presents considerations for using the posttest-only control group design for causal inference, discusses the advantages and limitations of this method, and provides an example.

Considerations for Causal Inference

The three commonly referenced conditions that must be met in order to infer causality are (1) temporal precedence of proposed cause and effect, (2) covariation of proposed cause and effect, and (3) elimination of alternative explanations for the effect. For the first condition to be met, the treatment or intervention must occur before differences between groups on the outcome variable (i.e., the effect) are observed. The posttest-only control group design is able to meet this condition because the treatment or intervention is administered before the potential group differences are observed. For the second condition to be met, it must be possible for the researcher to determine what happens both when the treatment or intervention is present and when it is absent. Again, the posttest-only control group design is able to meet this condition because it includes at least one treatment or intervention group and one control group. For the third condition to be met, potential alternative explanations for differences in group outcomes must be accounted for or ruled out. The posttest-only control group design may, but does not necessarily, meet this condition.

To minimize such alternative explanations, groups are formed using *random assignment*, so that any differences between groups that are observed before the administration of treatment or intervention will be due to randomness. Ideally, there will be no preexisting group differences due to random assignment and/or matching of some sort. The determining factor is how the groups are formed; if they are formed using random assignment with *matching* (in which the researcher creates pairs of participants, one from the treatment group and one from the control group, who have comparable important characteristics beyond group membership) or without, then the posttest-only control group design is able to meet this condition. If the groups are formed using a nonrandom mechanism (e.g., convenience, self-selection of participants into groups, criterion-based inclusion, or already-existing groups), then the posttest-only control group design cannot meet this condition. The way groups are formed for comparison is of crucial importance for inferring a causal relationship between the treatment or intervention and the observed differences between groups. For the types of research questions that lend themselves to posttest-only control group design, causal inference is commonly the desired outcome, so random assignment with or without matching is recommended.

Advantages

The posttest-only control group design is commonly compared against the pretest–posttest control group design. Because no data are collected before the administration of the treatment or intervention (i.e., no pretest), the posttest-only control group design requires fewer resources (e.g., time, money, energy) for data collection. In fact, collecting data prior to a treatment or intervention may not be possible or feasible. Furthermore, the process of collecting pretest data may prepare participants or give them clues into the intended effect of treatment or intervention (referred to as a *testing threat* to internal validity). This may affect the way participants interact with the actual treatment or intervention compared with how they would have acted without prior knowledge in a posttest-only design.

Limitations

There are a number of limitations of the posttest-only control group design. These limitations are primarily related to not being able to rule out alternative explanations. First, regardless of how groups are formed, it is possible that the treatment group(s) differ from the control group *before* any treatment or intervention is ever administered. If this is the case, there is no mechanism in place to provide evidence of these differences because no data were collected ahead of the treatment or intervention. Unknown preexisting differences may well affect the effectiveness of the treatment or intervention in some way. Such preexisting differences may result in the researcher reaching an incorrect conclusion about observed group differences being attributable to the treatment or intervention.

A second limitation associated with the posttest-only control group design is *maturation threat*, which is a more common threat to causal inferences in longer studies. Maturation refers to how people change over the course of time, whether it be emotionally, physically, or mentally. Although the length of the study is usually not long enough for people to naturally change in a meaningful way, the absence of a pretest prevents the researcher from being able to collect data on the natural change of participants over the course of the study. The expected growth of participants over the course of the study can only be assessed by examining the change of the control group. As a result, the posttest-only control group may be subject to the maturation threat on causal inference.

As noted earlier, the posttest-only control group design involves two or more groups, one of which is the control group. As a result, the design is subject to certain limitations related to multiple groups. These limitations have been referred to as *social interaction threats* to the inferring causal relationship(s), and many, if not all, research designs involving multiple groups are subject to these same limitations.

The first of the multiple-group limitations is *diffusion of treatment*. This occurs when participants in the control group receive the treatment or intervention from participants in the treatment or intervention group directly or through some other means, and the control group adopts or tries to replicate the treatment or intervention in their own group. The second of the multiple group limitations is *rivalry*. If the participants in each group know which group(s) their scores are being compared against, the group members may begin to compete with each other, thus affecting the results of the group comparisons. The third of the multiple-group limitations is *equalization of treatment*, which occurs when there is some perceived desirability, value, or inequity associated with being in the treatment or intervention group. If this is the case, there may be internal or external pressure to reassign participants to the treatment or intervention or to administer it to everyone to promote fairness. Obviously, this would interfere with the design's ability to demonstrate the effectiveness of the treatment or intervention. The last of the multiple-group limitations is *resentful demoralization*, which occurs when participants in the control group become aware that they are not receiving the treatment or intervention and, being demoralized by this discovery, attempt to sabotage the research. For example, they may perform worse on the assessments than they would ordinarily, artificially increasing the observed differences between groups.

Example

Two scenarios within the same context are presented here to demonstrate the posttest-only control group design and its implications on how the treatment and control groups are formed. Suppose a fifth-grade teacher hypothesizes that learning and practicing chess will increase math achievement in students who participate in an afterschool program. She decides to collect data that will allow her to test her hypothesis. In the first scenario, she randomly chooses half of the students in the afterschool program to learn and practice chess, and she administers the program with fidelity for 4 weeks. At the end of the 4-week period, she administers a math assessment with demonstrated reliability and validity evidence to all students in the afterschool program—those who participated in the chess program and those who did not. The students who participated in the chess program received higher math scores than those who did not. This scenario is an example of a posttest-only control group design. In a second scenario, the teacher allows any student who is interested to participate in the chess program. After students self-select into the program, she administers the program with fidelity for 4 weeks. At the conclusion of the 4-week period, she administers a math assessment to all students in the afterschool program—

those who participated in the chess program and those who did not. Those who participated in the chess program had higher math scores than those who did not participate in the chess program. This is also an example of a posttest-only control group design.

With which scenario would you feel more comfortable concluding that the observed differences in math achievement was attributable to the chess program? In the first instance, the groups were formed through a random process, and any differences that existed between the group members were therefore due to chance. In the second scenario, the groups self-selected into or out of the chess program, and as a result, there are known differences between the group members, namely, interest in chess. Therefore, the program may have a different effect for those students than it would have had for “typical” students. As a result, it would be impossible to separate the interaction between interest in chess and effect of chess on the observed difference in posttest math achievement.

For the purposes of this entry, the posttest-only control group design is exemplified through these two scenarios. As demonstrated, it is important to consider the threats to internal validity in the use and decision making associated with this research design.

Grant B. Morgan and Rachel L. Renbarger

See also Analysis of Variance; Causal Inference; Experimental Designs; Nonexperimental Designs; Pretest–Posttest Designs; Random Assignment; *t* Tests

Further Readings

Kirk, R. E. (2013). Experimental design: Procedures for the behavioral sciences (4th ed.). Thousand Oaks, CA: SAGE.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experimental and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

Trochim, W. M. K., & Donnelly, J. P. (2006). The research methods knowledge base (3rd ed.). Mason, OH: Atomic Dog.

Xiaofeng Steven Liu Xiaofeng Steven Liu Liu, Xiaofeng Steven

Power

Power

1281

1282

Power

Statistical power, or power for short, is the probability of rejecting a false *null hypothesis* in a hypothesis test, which involves the null hypothesis and the *alternative hypothesis*. The latter usually presents the researcher's theory—the *research hypothesis*. As the null hypothesis and the alternative hypothesis are opposite to each other, either one or the other is plausible but not both. Rejecting the null hypothesis establishes the plausibility of the alternative hypothesis (i.e., the researcher's theory). For example, a researcher hypothesizes that soothing music improves students' ability to solve puzzles. This becomes the alternative hypothesis. On the contrary, the null hypothesis states that soothing music does not help improve students' ability to solve puzzles. The probability of rejecting the null hypothesis represents the researcher's chance of confirming the hypothesized effect of soothing music on solving puzzles. In this regard, statistical power is always sought after because it affects the odds of affirming the research hypothesis.

Statistical power is determined by the *significance level*, the *effect size, error variance,* and *sample size* in a statistical test. The smaller the significance level (or lower the α value), the less power the statistical test will produce, other things being equal, and the larger the effect size, the higher the statistical power. Larger variability generally works against statistical power because error variance can be viewed as the background noise in detecting an effect. Increased error variance makes it more difficult to substantiate the possible effect stated in the alternative hypothesis. Although error variance inherent to a research setting may be beyond human control, a researcher can always increase the sample size to raise the statistical power to the desired level (e.g., .80, or 80% power).

Sample size determination is therefore synonymous with statistical power analysis.

The general recommendation is that a power analysis be conducted to calculate statistical power prior to the study (a priori) rather than after the study (post hoc or a posteriori). A researcher should determine power a priori so as to ensure an adequate chance of affirming the research hypothesis in planning a study. Upon completion of the study, a researcher should refrain from using the sample estimates to determine power a posteriori. The post hoc power, thus computed, merely validates what has already been observed in the study: An insignificant result yields low power and a significant result returns high power. In other words, the post hoc power based on the sample estimates does not add any new information to the study.

Xiaofeng Steven Liu

See also Effect size; *p* Value; Power Analysis; Sample Size; Significance

Further Readings

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, 55(1), 19–24. doi:10.1198/000313001300339897

Liu, X. (2013). Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques. New York, NY: Routledge.

Xiaofeng Steven Liu Xiaofeng Steven Liu Liu, Xiaofeng Steven

Power Analysis

Power analysis

1282

1285

Power Analysis

Statistical power analysis, or power analysis, is important to social science research because researchers and funding agencies usually wish to know whether a planned study has an adequate chance of detecting an effect in hypothesis testing. In other words, statistical power shows how likely it is that a scientific study can affirm a researcher's theory. Modern hypothesis testing features empirical data and a test statistic in examining two opposing suppositions: the *null hypothesis* (H_0) and the *alternative hypothesis* (H_a); the latter is often the research hypothesis or the researcher's theory. Power analysis involves the *test statistic*, the *significance level*, the *effect size, error variance*, and *sample size* in hypothesis testing.

Test Statistic

The empirical data for a research study are first collected from a representative sample and then summarized by a test statistic. The type of statistic used to test the research hypothesis is determined by the research design. Depending on random assignment, a study can be classified as either an *observational study* or a *true experiment*. An observational study does not involve any randomization. In this type of study, a *t* statistic can be used to test the correlation between two continuous variables. However, a true experiment involves random assignment of subjects into treatment and control conditions. A two-group randomized experiment (one treatment and one control group) can use a *Z* test to compare the two population means if the population standard deviation is known; when the standard deviation is unknown, a *t* statistic is used to compare the two population means. If there are two treatment groups and one control group in the randomized experiment, an *F* statistic will be computed to detect any mean differences among the three groups.

Hypothesis testing uses the probability behavior of the test statistic to assess the plausibility of the null and alternative hypotheses. The test statistic can deviate from its most expected value assuming the null hypothesis is true, but it can do so with predictably decreasing probability. The more the test statistic differs from the most expected value, the less likely such a test statistic can occur by chance. If the test statistic is highly discrepant from the value expected under the null hypothesis, it is construed as significantly contradicting the null hypothesis. In this case, the null hypothesis is deemed implausible and shall be rejected.

How deviant must the test statistic be from its expected value under the null hypothesis before it can be rejected? It depends on the *p* value of the test statistic or the probability of obtaining a statistic at least as deviant as the observed one. A small *p* value suggests that the statistic is atypical of its probabilistic occurrence when the null hypothesis is true. The null hypothesis therefore is rejected, and the alternative hypothesis (i.e., the research hypothesis) is plausible. On the contrary, a large *p* value indicates that such a statistic occurs often and does not contradict the null hypothesis. In this case, the null hypothesis shall be retained—the research hypothesis is not supported.

Significance Level

A *p* value less than or equal to 5% (i.e., $p \leq .05$) is conventionally considered small enough to reject the null hypothesis, and the 5% is the significance level (α). In practice, the exact *p* value of the computed test statistic does not have to be calculated. The computed test statistic is often compared with a critical value, whose *p* value is known to exactly equal the significance level of 5%. If the computed test statistic is further away from its expected value than the critical value, it can be inferred that the actual *p* value of the computed test statistic is smaller than the significance level of 5%, and the null hypothesis shall be rejected.

The probability of rejecting a false null hypothesis is the statistical power, and the significance level affects statistical power by dictating how aberrant a statistic needs to be before it is declared incompatible with the null hypothesis. Five out of 100 times (or 5%) is the normal standard for being rare or significant.

If the 5% significance level is changed to 1%, the standard for being significant becomes more stringent, and it will be much rarer to obtain an aberrant statistic that meets this more stringent standard. It is therefore more difficult to declare significance and reject the null hypothesis at the significance level of 1% compared with 5%. Once the significance level is set, calculating statistical power requires knowledge of the probability distribution of the test statistic under the alternative hypothesis.

Noncentrality Parameter

Statistical power is computed as the probability of the test statistic exceeding the critical value (i.e., rejecting the null hypothesis) under a *noncentral probability distribution*. A test statistic typically assumes a central probability distribution when the null hypothesis is true. If the null hypothesis is false and the alternative hypothesis is true, the test statistic then follows a noncentral probability distribution. For instance, a *t* statistic has a central *t* distribution when the population means are equal under the null hypothesis in a two-group comparison study. When the two population means are not equal, the alternative hypothesis is true and the *t* statistic will follow a noncentral *t* distribution. The noncentral probability distribution differs from its central counterpart by a nonzero noncentrality parameter (λ). The noncentrality parameter reflects how prominent the phenomenon stated in the alternative hypothesis appears in hypothesis testing. The larger the noncentrality parameter is, the higher the statistical power. The noncentrality parameter depends on the effect size, error variance, and sample size.

Effect Size

Effect size measures the magnitude of a phenomenon stated in the research hypothesis. It can be a correlation, a ratio, or a mean difference. A correlation shows the association between two things. A ratio can sometimes be used to portray the proportion of the variation due to the treatment over the total variation. The most common effect size is the mean difference between two populations ($\mu_1 - \mu_2$), for instance, the treatment group (μ_1) and the control group (μ_2). The mean difference shows how much the treatment group outperforms the control group—the *treatment effect*. A large treatment effect is easier to detect in hypothesis testing than is a small treatment effect. Therefore, a large treatment effect produces a large noncentrality parameter, which makes it easy to reject the null hypothesis and attain high statistical power.

The mean difference is a simple effect size, and the magnitude of a simple effect size needs to be interpreted with reference to the measurement scale. If the mean difference is divided by the standard deviation (σ), it becomes a standardized effect size or Cohen’s *d*,

$$d = \frac{\mu_1 - \mu_2}{\sigma}.$$

Cohen uses 0.2, 0.5, and 0.8 for small, medium, and large standardized effect sizes, respectively. These rule-of-thumb numbers offer general guidance, but they require a nuanced interpretation depending on the context. A large standardized effect size may either mean a large mean difference or a small standard deviation. A small standard deviation does not necessarily mean a large mean difference; it may simply indicate a small error variance (σ^2). The error variance can influence statistical power in hypothesis testing.

Error Variance

Error variance is the amount of variation due to extraneous variables that may interfere with the assessment of the phenomenon in question. For instance, the effect size is always estimated from a sample, and sampling error adds to uncertainty in the effect size estimate. Other extraneous factors (e.g., measurement error) can also inflate the error variance and obscure the phenomenon. Error variance can therefore be viewed as background noise in examining a hypothesized effect.

Error variance negatively affects statistical power. In a statistical test, the size of the effect size estimate is compared against the backdrop of its standard error. The effect size estimate appears prominent against a small standard error, which favors rejecting the null hypothesis and attaining high statistical power. The standard error is a direct function of the error variance. The larger the error variance, the larger the standard error gets. However, increasing the sample size can reduce the standard error. Because extraneous variables are sometimes beyond human control, researchers can always choose a large sample size to counter error variance.

Sample Size

Sample size affects statistical power much like the zooming function of a camera lens that can change the view of two neighboring objects. Zooming in makes the two objects appear separated and different. However, zooming out blurs the separation of the two objects. In hypothesis testing, the two objects are the group means. A large sample size decreases the standard error of the mean difference or the effect size estimate. The group means appear different from each other, and the null hypothesis of equal means is likely rejected—so statistical power will be high. On the contrary, a small sample size enlarges the standard error of the effect size estimate, which results in a low probability of rejecting the null hypothesis and low statistical power. Statistical power analysis can be used to determine an appropriate sample size in planning a research study.

Example

Suppose that a book publisher wants to know whether test preparatory materials can help students improve their performance on the Scholastic Aptitude Test (SAT). A representative sample of students will be recruited for the study, and they will be randomly assigned into two groups of equal size. An *n* number of students in the treatment group (μ_1) will receive the test preparatory materials in preparation for the SAT; another *n* number of students in the control group (μ_2) will not. Both groups will later take the SAT. The null hypothesis states that there is no difference in average SAT score between the two groups (H_0: $\mu_1 = \mu_2$); in other words, the population mean score of the treatment group is equal to the population mean score of the control group. The alternative hypothesis states that there is a difference between the two mean scores (H_a: $\mu_1 \neq \mu_2$). A statistical power analysis is commissioned to determine an adequate sample size to achieve an 80% chance of detecting a certain mean difference (i.e., 80% power) in a *t* test.

The statistical power in a two-sided *t* test is the probability of obtaining a *t* statistic exceeding the critical value, that is,

$$P\left[\left|T'(2n-2,\lambda)\right| \geq t_0\right],$$

where $T'(2n - 2, \lambda)$ is the noncentral *t* statistic with degrees of freedom $2n - 2$

under the alternative hypothesis. The critical value at the significance level of 5% is t_0, and is the cumulative distribution function of the noncentral *t* with a noncentrality parameter λ:

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma}\sqrt{\frac{n}{2}}.$$

The error variance common to both groups is σ^2, and the standard deviation is σ. Once the standard deviation σ and the simple effect size $\mu_1 - \mu_2$ are conjectured, the noncentrality parameter can be expressed in Cohen’s . If Cohen’s *d* is assumed to be .5, power can be calculated for different sample sizes in statistical software (e.g., SAS, SPSS, and R). When the sample size *n* (the number of subjects in each group) is 50, the statistical power is approximately 70%. For 80% power, the required sample size *n* is 64 in each group.

Xiaofeng Steven Liu

See also Effect size; *p* Value; Power; Sample Size; Significance

Further Readings

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. The American Statistician, 55, 187–193. doi:10.1198/000313001317098149

Liu, X. (2013). Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques. New York, NY: Routledge.

Eduardo Estrada Eduardo Estrada Estrada, Eduardo

Power Tests

Power tests

1285

1286

Power Tests

In the context of educational measurement, a power test usually refers to a measurement tool composed of several items and applied without a relevant time limit. The respondents have a very long time, or even unlimited time, to solve each of the items, so they can usually attempt all of them. The total score is often computed as the number of items correctly answered, and individual differences in the scores are attributed to differences in the ability under assessment, not to differences in basic cognitive abilities such as processing speed or reaction time. This entry describes what a power test is and how it should be applied, provides some examples of power tests, and explains how it is related to the concept of test speediness.

Definition and Application

The term *power test*, together with the opposite concept of *speed test*, was first proposed by Harold Gulliksen in his 1950 book, *Theory of Mental Tests*. To correctly measure the ability of the test takers and assess individual differences among them, it is crucial to apply a power test with an adequate difficulty level for the group under assessment. In a power test, most test takers should be unable to solve a large proportion of the items, and a few items should be solved only by a few test takers, or even not solved by anyone. Ideally, none of the test takers should be able to correctly answer all of the items because this may indicate a ceiling effect for those who do.

One example of a pure power test would be an English vocabulary examination in which a list of words is presented to a group of sixth-grade students who then have to find a synonym for each word without using any external help. If the words are adequately chosen, most students should be easily able to find synonyms for some of them, others should be moderately difficult for the students, and most students should be unable to find a synonym for the most difficult words. Importantly, the students should have a generous time limit for completing the test (e.g., 2 minutes per word on average), so the results yield a pure measure of English vocabulary knowledge, uncontaminated by individual differences in word fluency, processing speed, and reaction time.

The Raven’s progressive matrices test is often used as an example of a power test for evaluating mental ability. It is intended to measure inductive and visual reasoning, and it is composed of a list of items of increasing difficulty. Because there are different versions for children, the normal adult population, and the above-average adult population, the exact number of items and time restrictions depends on the test version. But, under standard conditions of application, examinees are given more than 1 minute per item in every version. Even if the test takers have generous time restrictions, the fact that the last items are very difficult and the test is often completed sequentially implies that many test takers do not have time to consider all items. Thus, although there is a consensus on considering it a “nonspeeded” test, individual differences in speed may still have a minor effect on the examinees’ scores.

Test Speediness

In practice, pure power tests are rarely used in educational assessment. It is extremely uncommon to apply a test without any time constraint, and only in a few scenarios can these time constraints be considered as irrelevant. In fact, most standardized tests have clear time restrictions. This has led some authors to propose that most tests are partly power and partly speeded in unknown proportions. The concept of *test speediness* or *test speededness* has been defined as the extent to which the time restrictions on a maximum-performance test have an impact in the test takers’ achievement.

Following John B. Carroll’s *Human Cognitive Abilities*, published in 1993, many authors began to consider test speediness to be detrimental to the aims of most standardized tests. The rationale for this is that key mental abilities, such as abstract reasoning, language use, mathematical competence, and visual–spatial aptitude, are mainly level (i.e., power) abilities. Therefore, all tests measuring these abilities should be power tests, and any difference in the scores attributable to differences in speed must be considered error variance and irrelevant to the construct of interest. In other words, according to this view, the speed component is a threat to the validity of most tests that measure mental ability.

To solve this problem, various methods have been proposed for studying test speediness. These procedures try to isolate the proportion of the scores' variance that is due to the speed and the power components. Some techniques use information external to the test, such as response-time measures, whereas others rely only on information provided by the test, such as the proportion of unreached items.

Eduardo Estrada

See also Raven's Progressive Matrices; Speeded Tests

Further Readings

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York, NY: Cambridge University Press.

Chadha, N. K. (2009). Speed test versus power test. In Applied psychometry (pp. 39–48). New Delhi, India: SAGE.

Gulliksen, H. (1950). Theory of mental tests (Vol. XIX). Hoboken, NJ: Wiley.

Hunt, E. (2011). The tests. In Human intelligence (pp. 31–78). New York, NY: Cambridge University Press.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. Educational Measurement: Issues and Practice, 26(4), 29–37. doi:10.1111/j.1745-3992.2007.00106.x

Kathryn Weaver Kathryn Weaver Weaver, Kathryn

Pragmatic Paradigm

Pragmatic paradigm

1286

1288

Pragmatic Paradigm

The pragmatic paradigm refers to a worldview that focuses on “what works” rather than what might be considered absolutely and objectively “true” or “real.” Early pragmatists rejected the idea that social inquiry using a single scientific method could access truths regarding the real world. These pragmatists declared that truth could be judged by its consequences. The pragmatic paradigm is useful for guiding research design, especially when a combination of different approaches is philosophically inconsistent.

Evolution of Pragmatism

Pragmatism as a philosophical movement originated in the 1870s by Charles Sanders Peirce (1839–1914). Presenting his basic ideas of pragmatism in the series *Illustrations of the Logic of Science* (1877–1878), Peirce aimed to connect thought and action. Thought produced beliefs, which Peirce defined as entities on which one is prepared to act and not just as a state of mind. To Peirce, doubt hindered action by causing one to continuously inquire until a belief was attained, and inquiry was the practical activity of eliminating doubt to understand an idea in a fruitful way. The meaning of an object or conception could only be fully understood through its practical consequences; for example, one understands what is meant by a timepiece if one knows what a timepiece does.

Pragmatism was further developed by the 19th-and 20th-century classical pragmatists William James (1842–1910) and John Dewey (1859–1952). Writing in 1907, James described pragmatism as a “method for settling metaphysical [theoretical] disputes that might otherwise be interminable.” He proposed tracing the consequences of each idea as a way to enable people to solve problems for themselves.

John Dewey also conceived of inquiry and knowledge as instruments that allowed people to reshape their environment and improve the quality of their lives. Influenced by Darwin’s theory of evolution, Dewey regarded intelligence as power that one could use in facing a challenge beyond one’s usual ways of thinking and acting. Dewey tried to close the gap between action and thought by defining *action* as conducting experiments under controlled situations and *thought* as those theories guiding experiments.

Core Tenets of Pragmatism

The core of pragmatism is the pragmatic maxim, formulated by Peirce and James to clarify intractable metaphysical and epistemological (knowledge-based) disputes through asking, “What concrete practical difference would it make if one theory were true and its rival(s) false?” Where there is no such difference, there is no genuine disagreement and hence no problem. Other tenets of pragmatism are the rejection of all forms of absolutism and the regard for principles as working hypotheses that must produce results in practice. Pragmatists also share a distinctive anti-Cartesian, fallibilist outlook. Rejection of the Cartesian picture of the mind and its contents as subjective, private, and severed from the objective, public world led Dewey to reject dichotomies between fact and value, mind and body, and theoretical judgments and practical judgments.

As Peirce claimed, people possess a background of certainties and everyday beliefs that they trust until given a reason for doubting them. James in 1907 stated that when a new experience challenges established certainties, people will attain a new belief that “preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty.” Dewey emphasized that the utility of a theory is its problem-solving capacity. To the extent that a theory functions to resolve significant difficulties, it makes sense to keep using it—although it may be replaced by one that works better.

According to James, people want to obtain truth and avoid error. The more rigorously the search for truths, the more likely errors will be let in. This possibility of error supports the fallibilist view that a belief may need to be revised in the future. It does not provide reason for skepticism, as the aim of pragmatic inquiry is not on possessing absolute certainty but on possessing methods of inquiry that contribute to human progress.

Pragmatism as a Paradigm

A *paradigm* is a theoretical framework comprising the set of basic beliefs that guide the research or practice of a scientific community and that influence the way knowledge is studied and interpreted within a discipline. In addition to the pragmatic paradigm, there are other paradigms (e.g., postpositivist, constructivist, interpretivist, transformative, emancipatory, critical, participatory, and deconstructivist), which compete for acceptance, according to Thomas Kuhn (writing in 1970). The choice of paradigm is influenced by sets of elements unique to each that define its assumptions of *ontology* (nature of reality), *epistemology* (nature of knowledge), *axiology* (nature of value and how the values of the researcher can influence what is to be studied), and *methodology* (techniques for inquiry and examining practice).

In terms of ontology and epistemology, pragmatism is not committed to any single system of philosophy and reality. Reality is actively created as individuals act in the world, and it is thus ever changing, based on human experience, and oriented toward solving practical problems. Truth is what works at the time and not based on dualism between reality independent of the mind (as with postpositivism and critical paradigms) and within the mind (as with constructivist and deconstructivist paradigms). Most pragmatists embrace a form of *naturalism* (the idea that philosophy is not prior to science but continuous with it). Therefore, their methodology uses the method of science and is open to exploring the different kinds of methods employed in different branches of science. Thus, pragmatism has gained considerable support as a stance for mixed-methods research.

Pragmatic Underpinnings in Social Science and Education

Pragmatism emphasizes that research involves decisions about which goals are most meaningful and which methods most suitable. Dewey opposed any use of force or domination that could limit freedom of inquiry and possibilities for social justice with other social groups. The compatibility of pragmatism with versions of transformative, emancipatory, and participatory research can enable a more detailed understanding of social conflicts.

Advantages

Pragmatic inquiry is concerned with evaluating and transforming features of real-world psychological, social, and educational phenomena. The resulting rich understanding of experience and science offered by pragmatists may reveal ways to find an objective basis for the criticism of institutions and practices. By emphasizing the communal character of inquiry and priority of democracy over philosophy, pragmatism provides a basis for a defense of democratic values. Taking a pragmatic and balanced or pluralistic position will help improve communication among researchers from different paradigms.

Limitations

The early formulations of pragmatism by James and Dewey, particularly regarding the difficulty of determining usefulness or workability, were criticized by Bertrand Russell (1872–1970). This criticism assumes that the usefulness of any particular mixed-methods design can be known in advance of it being used. The question of whether a mixed methods design "works" or not can only be decided once the research product is completed and the findings interpreted.

Kathryn Weaver

See also Constructivism; Mixed-Methods Research; Paradigm Shift; Postpositivism; Scientific Method

Further Readings

Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. Journal of Mixed Methods Research, 4(1), 6–16. doi:10.1177/1558689809349691

James, W. (1907). Pragmatism: A new name for some old ways of thinking.

Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). International Encyclopedia of Unified Science, 2(2). Retrieved from http://projektintegracija.pravo.hr/_download/repository/Kuhn_Structure_of_Sc

Morgan, D. (2014). Pragmatism as a paradigm for social research. Qualitative Inquiry, 20(8), 1045–1053. doi:10.1177/1077800413513733

Peirce, C. S. (1878). How to make our ideas clear. Popular Science Monthly, 12, 286–302. Retrieved from https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_12/January_

Russell, B. (1910). Philosophical essays. New York, NY: Simon and Schuster.

Wright Mills, C. (1964). Sociology and pragmatism: The higher learning in America. (I. L. Horowitz, Ed.). Oxford, UK: Oxford University Press.

Nathan H. Clemens Nathan H. Clemens Clemens, Nathan H.

Kelsey Ragan Kelsey Ragan Ragan, Kelsey

Christopher Prickett Christopher Prickett Prickett, Christopher

Predictive Validity Predictive validity

1288

1289

Predictive Validity

Predictive validity refers to the degree to which scores on a test or assessment are related to performance on a criterion or gold standard assessment that is administered at some point in the future. Predictive validity is often considered in conjunction with *concurrent validity* in establishing the criterion-based validity of a test or measure. Although concurrent validity refers to the association between a measure and a criterion assessment when both were collected at the same time, predictive validity is concerned with the prediction of subsequent performance or outcomes. Educators, researchers, and practitioners are often interested in how well a test or assessment will forecast an individual's future performance in a particular domain; therefore, predictive validity is an important aspect for demonstrating the technical adequacy and practical utility of a test or measure.

Predictive validity is typically established using correlational analyses, in which a correlation coefficient between the test of interest and the criterion assessment serves as an index measure. Multiple regression or path analyses can also be used to inform predictive validity. Because the administration of the test and the criterion assessment may be separated by several weeks, months, or years, it should be noted that predictive validity coefficients are often weaker than concurrent validity coefficients due to maturation, learning, or other variables associated with the passage of time between the assessment occasions.

Several examples of applications of predictive validity to education research and practice can be considered. A measure of toddlers' receptive vocabulary might be evaluated for its ability to predict scores on an assessment of vocabulary proficiency at school entry. An assessment designed to evaluate kindergarten students' emergent literacy skills might be evaluated in terms of how well it predicts their performance on a reading assessment in a subsequent grade. A test of math skills administered at the beginning of the school year might be evaluated for its ability to predict students' scores on a test of overall mathematics proficiency taken later in the year. Or, an assessment of high school students' study habits might be evaluated for the strength at which it predicts rates of school completion or performance on college entrance examinations.

Because predictive validity is a key part of demonstrating the technical adequacy of a test or measure, it is important for test developers to report results of predictive validity studies. Additionally, researchers or educators who are considering the use of a particular test or measure for predicting future performance should determine whether the test has demonstrated evidence of predicting outcomes or performance in the domain of interest. This information is often found in technical support materials provided by the publisher or in independent research studies that have evaluated the predictive validity of the test.

Nathan H. Clemens, Kelsey Ragan, and Christopher Prickett

See also Concurrent Validity; Correlation; Criterion-Based Validity Evidence; Tests; Validity; Validity Coefficients

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5–8. doi:10.1111/j.1745-3992.1995.tb00881.x

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2013). Assessment in special and inclusive education (12th ed.). Belmont, CA: Wadsworth.

Maria Jimenez-Buedo Maria Jimenez-Buedo Jimenez-Buedo, Maria

Pre-experimental Designs Pre-experimental designs

1289

1291

Pre-experimental Designs

Pre-experimental designs are research schemes in which a subject or a group is observed after a treatment has been applied, in order to test whether the treatment has the potential to cause change. The prefix *pre-* conveys two different senses in which this type of design differs from experiments: (1) pre-experiments are a more rudimentary form of design relative to experiments, devised in order to anticipate any problems that experiments may later encounter vis-à-vis causal inference, and (2) pre-experiments are often preparative forms of exploration prior to engaging in experimental endeavors, providing cues or indications that an experiment is worth pursuing. Because pre-experiments typically tend to overstate rather than understate the presence of causal relations between variables, it is sometimes useful to run a pre-experiment (or more commonly, to observe the results of an existing one) in order to decide whether an experiment should be undertaken.

Experimental evidence is defined in contrast to observational evidence: Although the former involves some form of intervention, the latter is limited to recordings of events as they naturally occur, without controlling the behavior of the object being studied. An experiment is normally used to create a controlled environment to aid in establishing valid inferences about the behavior of the object being studied; typically, the control involved in the experiment is used to infer causality among variables. In the case of a pre-experiment, although there is some intervention in the object, the level of intervention does not provide the control required for valid inferences regarding the causal processes involved. Thus, pre-experiments differ from observational data because they are based on some form of intervention. At the same time, they are different from experiments because not enough control is achieved to ensure valid causal inferences.

Types of Pre-Experimental Designs

The category of pre-experimental designs is necessarily an open one because it is defined negatively, in opposition to true experimental designs. Yet, three types of designs are normally considered standard pre-experiments and are used routinely by researchers: the *one-shot case study*, the *single-group before and after*, and the *static group comparison*.

One-Shot Case Study

Also referred to as a *single-group posttest design*, this type of research involves a single group of subjects being studied at a single point in time after some treatments have taken effect, or more broadly, after some relevant intervention that is supposed to cause change has taken place. In order to make inferences about the treatment, the measurements taken in the one-shot case study are compared to the general expectations about what the case would have looked like if the treatment had not been put in place because there is no control or comparison group involved.

In the standard representational language of experimental research design, a one-shot case study is represented as follows:

$$X \rightarrow O,$$

where the *X* represents the treatment or intervention and the *O* represents the observation by researchers of the variable of interest.

Single-Group Before and After

Also known as a *one-group pretest–posttest design*, this method involves a single case observed at two different points in time—before and after an intervention or treatment. Whatever changes happen in the outcome of interest are presumed to be the result of the intervention. Again, there is no control or comparison group involved in this type of study design.

In the standard representational language of experimental research design, a single-group before-and-after design is represented as follows:

$$O \rightarrow X \rightarrow O.$$

Static Group Comparison

Also referred to as a *cross-sectional* or *transversal study*, this type of design involves two groups: one on which a treatment intervention has been carried out (O_1) and another group on which no intervention has been performed (O_2). The difference in the outcome of interest between the two groups is assumed to be caused by the intervention.

In the standard representational language of experimental research design, a static group comparison is represented as follows:

$$\begin{array}{l} X \to O_1 \\ \quad \to O_2. \end{array}$$

Validity and Relevant Comparisons

The main advantage of pre-experimental designs is their cost: The majority of pre-experiments lack a comparison group, which makes them less expensive to run than true experiments. Therefore, they may be the better option if resources are limited. Because of this lack of comparison group, however, they are vulnerable to a number of validity threats. The one-shot case study is vulnerable to the following biases: history, maturation, selection, mortality, and selection treatment. In turn, the single-group, before-and-after design is often affected by biases such as history, maturation, testing, regression, selection maturation, and selection treatment. Finally, the static group comparison often displays problems such as selection mortality, selection maturation, maturation, and selection treatment.

The limitations of pre-experimental design all highlight the importance of comparison groups; this in turn helps to underline the absolute centrality of comparison in making causal inferences. In fact, the kind of inferences that pre-experiments allow, and the inferential difficulties they present, resembles those of observational studies with small sample sizes (small-*N* studies). It is thus no surprise that static group comparisons are on practical grounds analytically indistinguishable from observational cross-sectional studies and that one-shot case studies and single-group, before-and-after designs share many of the features of observational case studies, which are typical of qualitative social sciences research. Thus, for the interpretation of pre-experimental evidence, the analytical strategies that have been made available to researchers dealing with small-*N* studies can often be of use. These are works that develop the comparative method in the tradition initiated by John Stuart Mill (1806–1873) in order to determine which logical conclusions can be supported by a data set composing just a few cases.

In this way, pre-experiments, together with quasi-experiments, demonstrate that there is actually a continuum between observational and experimental studies in terms of the type of inferences they allow. The thread of the continuum is provided by the central notion of *comparison*: Meaningful comparisons are needed in order to establish valid inferences. In the case of pre-experimental research, the absence of relevant comparison groups can be partially circumvented by the background knowledge of the researcher, together with a large dose of caution in the drawing of causal conclusions. In any event, it is always advisable, in the presence of pre-experimental evidence suggesting a causal effect, to run subsequent studies that can rule out validity threats.

Maria Jimenez-Buedo

See also Experimental Designs; External Validity; Internal Validity; Qualitative Data Analysis; Quasi-Experimental Designs; Threats to Research Validity

Further Readings

Marsden, E., & Torgerson, C. J. (2012). Single group, pre-and posttest research designs: Some methodological concerns. Oxford Review of Education, 38(5), 583–616. doi:10.1080/03054985.2012.731208

Mill, J. S. (1856). A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation: Vol. 1. New York, NY: Harper & Brothers. Retrieved from https://books.google.com/books

Nunan, D. (1992). Research methods in language learning. New York, NY: Cambridge University Press

Patten, M. L. (2007). Understanding research methods: An overview of the essentials (6th ed.). London, UK: Routledge.

Ragin, C. C. (2014). The comparative method: Moving beyond qualitative and quantitative strategies. Berkeley: University of California Press.

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. Political Research Quarterly, 61(2), 294–308. doi:10.1177/1065912907313077

Swanborn, P. (2010). Case study research: What, why and how? London, UK: SAGE.

Marilyn M. Ault Marilyn M. Ault Ault, Marilyn M.

Premack Principle

Premack principle

Premack Principle

The Premack principle is an observation about the effectiveness of using certain types of behavior or activities as reinforcement. According to David Premack, “Any response A will reinforce any other response B, if and only if the independent rate of A is greater than that of B” (1959, p. 219). Another way to say this is that a more preferred activity, that with a high independent rate, can be used to reinforce a less preferred activity, that with a low independent rate. Or, a high-probability event can be used to increase the frequency of a low-probability event.

The principle fits within the realm of operant conditioning, the use of a stimulus to elicit a response and the use of a consequence, or reinforcement, to increase the probability of the response occurring again. The principle is noteworthy because it expanded the understanding of what constitutes reinforcement to include activities and behavior. The opportunity to engage in a preferred activity, one that is considered a high-probability event, can be used to increase the probability of, or reinforce, a less-preferred or less frequent activity.

In 1959, this occurrence was first identified by Premack in his early work with primates and replicated with young children. He used children’s preferences to determine reinforcing events and demonstrated the effectiveness of using these events as consequences to strengthen the probability of specific behaviors. Some suggest that Premack identified an obvious relationship and elevated it to a principle of behavior.

This principle has broad application in areas of education and management. The application of the principle can be seen in a wide array of procedures and settings to increase the concurrence of positive or productive behavior. These include, for example, the scheduling of activities in preschool settings that alternate more difficult tasks with more preferred activities, the contingency management of adolescent delinquents to obtain privileges if they engage in productive behaviors, self-management programs for weight reduction by suggesting access to a fun activity after exercise, or management of employees to increase productivity by scheduling breaks contingent on outcomes. The principle is also widely applied in parenting practice and is informally known as “Grandma’s law” or “Do your math homework and then you can play with your toys.”

Considerations of its application include an understanding of satiation and deprivation. As with the use of any type of reinforcement, one element that needs to be considered when selecting and using a particular reinforcement is the motivational strength of the consequence. Satiation suggests that the consequence is overused and that the probability of the event occurring independently is reduced. The concept of deprivation addresses the extent to which an individual is deprived of, or prevented from, performing the more-liked or higher probability event or behavior. This theory suggests that the behavior being used as reinforcement will only be effective if access to that behavior is reduced below baseline levels.

Marilyn M. Ault

See also Behaviorism; Operant Conditioning; Punishment; Reinforcement

Further Readings

Killeen, P. R. (2014). Pavlov + Skinner = Premack. International Journal of Comparative Psychology, 27(4).

Klatt, K. P., & Morris, E. K. (2001). The Premack principle, response deprivation, and establishing operations. The Behavior Analyst, 24(2), 173.

Premack, D. (1959). Toward empirical behavior laws: I. Positive reinforcement. Psychological Review, 66(4), 219–233. Retrieved from http://dx.doi.org/10.1037/h0040891

Premack, D. (1971). Catching up with common sense or two sides of a generalization: Reinforcement and punishment. In R. Glaser (Ed.), The Nature of Reinforcement (pp. 121–150). New York, NY: Academic Press.

Timberlake, W., & Allison, J. (1974). Response deprivation: An empirical approach to instrumental performance. Psychological Review, 81, 146–164.

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Xiaoqin Zhu Xiaoqin Zhu Zhu, Xiaoqin

Pretest–Posttest Designs Pretest–posttest designs

1292

1295

Pretest–Posttest Designs

In single-group pretest–posttest designs, or pretest–posttest designs, the dependent variable or variables are measured before the intervention (i.e., the pretest) and after the intervention (i.e., the posttest). Typically, measures used in the pretest and the posttest are the same, and changes in the dependent variable from pretest to posttest are interpreted to reflect the effectiveness of the intervention (the independent variable). This entry describes the pretest–posttest design compared with classical experimental designs, the role of the pretest–posttest design in human services and education evaluation, possible threats to the internal validity, issues related to the external validity, and ways to strengthen the design.

A pretest–posttest design is a form of pre-experimental design that does not have control or comparison groups. This is a major difference between pretest–posttest designs and processes such as quasi-experimental designs and randomized controlled trials. The logical basis of the pretest–posttest design is in the *method of agreement*, proposed by John Stuart Mill in 1843. Specifically, if *Y* (a change in the dependent variable) regularly follows *X* (an independent variable), then *X* is sufficient for *Y* to happen and could be a cause of *Y* (i.e., if *X*, then *Y*). However, without a control group that is consistent with Mill's *method of differenc*e (i.e., if not *X*, then not *Y*), it is unknown whether *X* is a necessary condition for *Y* to occur. As a result, a causal inference between a change in the dependent and independent variables would be subject to rival explanations in the pretest–posttest design. In other words, the internal validity of the design is subject to threats.

The Pretest–Posttest Design in Evaluation

The Pretest–Posttest Design in Evaluation

Despite its high susceptibility to threats of internal validity, the pretest–posttest design is commonly used across therapeutic and educational settings for several reasons. First, it is unethical to assign participants who need immediate treatment or intervention, such as clients with depressive symptoms or children having posttraumatic stress disorder, to a control group. Second, resources and manpower may not be sufficient to conduct sophisticated experimental studies. Third, teachers without advanced research training can use this design to evaluate their own practice within school settings. Finally, the design can be a cost-effective way to study the preliminary outcomes of a program and decide whether the research should be extended.

Similar to health services, program developers in education research may also use the pretest–posttest design to determine the effectiveness of an education intervention program. For example, the pretest–posttest design was used in the experimental implementation phase of a positive youth development project called *Positive Adolescent Training through Holistic Social Programs* (P.A.T.H.S.) before a randomized group trial was adopted in implementing the program across Hong Kong. In addition, the American Association of State Colleges and Universities has suggested using the pretest–posttest design as a value-added measure in evaluating student learning and adjusting course curricula to better satisfy students’ needs. Typically, a pretest is given to students at the beginning of a course to determine their initial understanding of the measures stated in the learning objectives, and posttest is conducted just after completion of the course to determine what the students have learned. Ideally, there will be a positive change in outcomes.

Threats to Internal Validity

In his 1957 article entitled “Factors Relevant to the Validity of Experiments in Social Settings,” Donald T. Campbell, a pioneering researcher in the field of evaluation, introduced the concepts of internal validity and external validity in research design and stated the factors that influence them. Internal validity refers to the degree of confidence with which plausible rival explanations for research results can be ruled out. Donald T. Campbell and Julian C. Stanley further summarized eight different types of threats to internal validity in experimental studies in their 1963 book, *Experimental and Quasi-Experimental Designs for*

Research on Teaching. The threats most common to pretest–posttest designs include *history, maturation, testing effect, instrumentation*, and *regression to the mean* (RTM).

History

Many events in addition to the intervention may occur between administration of the pretest and the posttest and may account for some or even all of the observed changes. These events might occur either within or outside the context of intervention. For example, in an evaluation study of a school leadership program that aims to improve student awareness of ethical leadership, events occurring within the program context (e.g., participants' education experiences in other classes during the period of implementation) or outside the program context (e.g., a widely disseminated news story of a leadership scandal) may both affect the results.

Maturation

Maturation has a broader meaning than its literal interpretation. The term is used by Campbell and Stanley to cover all possible biological or psychological changes that naturally occur with the passage of time. For example, from pretest to posttest, the students will grow older, and they may become more tired, hungry, and bored; these changes may affect posttest scores regardless of intervention. The threat becomes more serious as the time between pretest and posttest increases and when the outcome measure is unstable over time because of increasing maturity. For instance, in a study investigating adolescents' self-concept, maturation might not be a big factor if the time in between is relatively short (e.g., one or two weeks), but changes in self-concept may occur simply because of maturation if there is a year or longer interval between the pretest and the posttest.

Testing Effect

Testing effect refers to the effect of the pretest itself. Changes in the posttest might result from the pretest independently of the subsequent intervention because (a) participants remember the questions in the pretest, (b) measures in the pretest have sensitized the participants to specific knowledge or problems, or (c) the pretest is raising participants' awareness and motivation to learn the topic

(c) the pretest is raising participants' awareness and motivation to learn the topic after the pretest. For example, students may perform better when they take the test for the second time because they are more familiar with the test or they may change their interaction pattern as a result of observers who are placed in the classroom to evaluate their pretraining interpersonal skills. In general, a greater testing effect can be expected if the test device is more novel and stimulating.

Instrumentation

Instrumentation indicates changes in measuring instruments that may explain pretest–posttest differences. More specifically, when people are involved in providing pretest and posttest scores (e.g., scores are self-reported by participants or rated by human observers), issues such as the evaluators' understanding and fatiguing may produce pretest–posttest differences. For example, students tend to overestimate their skills in the pretest and give a more informed evaluation as a result of a better understanding of the skills in the posttest. Similarly, raters of students' essays or classroom performance may apply different assessment standards in the pretest and the posttest, possibly because the raters are more skillful or familiar with the evaluation tasks in the posttest.

RTM

RTM biases the conclusion of the pretest–posttest design when participants are selected based on extremely low or high pretest scores, as the extreme pretest scores tend to move closer to the average over time. For example, if a program aiming to promote students' reading skills selects a group of students who do poorly in a reading test (the pretest in this case), then their posttest scores will certainly be higher on average than their pretest scores, regardless of any intervention. Likewise, if pretest scores are extremely high, then posttest scores will tend to be lower. These results are not caused by intervention or other effects like test–retest effect but are related to errors of measurement.

Many test results, such as scores on a reading test or an attitude test, are normally distributed with most values clustered around the mean score and only a smaller proportion of values deviating from the mean. Given that the measurement has a random error term, then the scores that markedly deviate from the mean will probably contain a larger error term than those nearer the mean. In this sense, extremely high scores tend to have an unusually large mean. In this sense, extremely high scores tend to have an unusually large positive error, whereas extremely low scores contain an unusually large negative error. However, the unusually large measurement error does not always occur, so in the posttest, the higher scores are expected to decline, whereas lower scores are expected to increase toward the mean.

Although RTM is known to be more evident for participants with extreme pretest scores, it does not mean that RTM will not affect a group formed by participants with a wide range of different scores. Movement from lower scores up to the mean and movement from higher scores down to the mean may not be equivalent, especially when the joint effects of history, maturation, testing effects, instrumentation, and intervention increase the regression effect in one direction while reducing it in the other.

One method to investigate whether RTM has affected the data is to plot change in scores (i.e., posttest−pretest scores) against corresponding pretest scores. If RTM does operate, the higher pretest scores will, on average, decrease toward the mean (i.e., there will be smaller gains), whereas the lower pretest scores will tend to increase toward the mean (i.e., there will be larger gains). As a result, a negative correlation will be observed between pretest scores and gains (i.e., change scores). One issue with using change scores, however, is that they assume the reliability of the pretest and posttest measures is acceptable.

Issues Related to External Validity

External validity is the extent to which the effect of an intervention can be generalized to different conditions: Would the same intervention produce the same results if it is implemented with different participants in a different setting? External validity of the pretest–posttest design, as well as other research designs, may be affected by population factors (e.g., selection, age, and gender) and ecological factors (e.g., settings such as room and time of the day). An issue uniquely related to external validity in pretest–posttest designs is that generalization may be limited to the same testing conditions as a result of interaction between the test (especially the pretest) and the intervention. As mentioned earlier, the pretest itself may sensitize the participants to the intervention; thus when the intervention is implemented without a pretest, results might change as well.

Ways to Strengthen the Design

The pretest–posttest design can be useful in evaluation if it is well conducted and if caution in drawing causal inferences is exercised. Basically, there are two common ways to strengthen the design. First, if all measures consistently change in a predicted direction after the intervention, using several instead of just one valid and reliable outcome measure can make conclusions more convincing. Second, multiple pretests and multiple posttests can provide more credible evidence regarding the participants' status prior to the intervention and can shed light on both immediate and long-term status after the intervention. In fact, if a series of pre-and posttests are employed over time, the pretest–posttest design changes into a quasi-experimental scheme known as the time series design.

Daniel Tan-lei Shek and Xiaoqin Zhu

See also Causal Inference; Internal Validity; Pre-experimental Designs; Program Evaluation; Quasi-Experimental Designs

Further Readings

Banta, T. W., & Pike, G. R. (2007). Revisiting the blind alley of value added. Assessment Update, 19(1), 1–2, 14–16.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54(4), 297–312. doi:10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago, IL: Rand McNally. (Reprinted from D. T. Campbell & J. C. Stanley (Eds.), *Experimental and quasi-experimental design for research*, 1963. Chicago IL: Rand McNally.) Marsden, E., & Torgerson, C. J. (2012). Single group, pre-and posttest research designs: Some methodological concerns. Oxford Review of Education, 38(5), 583–616. doi:10.1080/03054985.2012.731208

Mill, J. S. (2009). A system of logic ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation (8th ed.). New York, NY: Harper & Brothers.

Shek, D. T. L. (2013). Evaluation of the Project P.A.T.H.S. using multiple evaluation strategies. In D. T. L. Shek & R. C. F. Sun (Eds.), Development and evaluation of positive adolescent training through holistic social programs (P.A.T.H.S.) (pp. 53–67). Singapore: Springer.

Shek, D. T. L., & Sun, R. C. F. (2012). Promoting psychosocial competencies in university students: Evaluation based on a one group pretest-posttest design. International Journal on Disability and Human Development, 11(3), 229–234. doi:10.1515/ijdhd-2012-0039

Shek, D. T. L., Yu, L., & Ma, C. M. S. (2014). The students were happy but did they change positively? Yes, they did. International Journal on Disability and Human Development, 13(4), 505–511. doi:10.1515/ijdhd-2014-0348

Suskie, L. (2010). Assessing student learning: A common sense guide. Hoboken, NJ: Wiley.

Thyer, B. A. (2012). Quasi-experimental research designs. New York, NY: Oxford University Press.

Larry Davis Larry Davis Davis, Larry

Primary Trait Scoring Primary trait scoring

1295

1297

Primary Trait Scoring

Primary trait scoring is an approach for evaluating constructed responses, in which scores are based on one or more specific aspects of performance that are essential for the successful completion of the tested task. Primary trait scoring is most typically associated with writing assessment and was originally developed by Richard Lloyd-Jones and colleagues in the early 1970s to score writing in the U.S. National Assessment of Educational Progress (NAEP); since then, it has been used in a variety of contexts such as task-based assessment of second languages. As originally formulated, a key assumption of primary trait scoring was that different types of writing tasks and contexts have different criteria for success. Therefore, it is necessary during test design to carefully define the types of tasks needed to assess the ability of interest and to produce scoring materials that focus on the specific performance aspects that contribute most to the task being assessed. An implication of this specificity is that scoring rubrics may not necessarily generalize across different task types. This entry describes the characteristics of primary trait scoring, along with the strengths and weaknesses of this approach to assessment.

Characteristics

As the name suggests, primary trait scoring targets a limited number of key features of performance that are considered most important for success. Strictly speaking, a score is awarded on the basis of a single criterion, although complex tasks may receive several scores using different rubrics, each addressing a distinct element of performance; multiple rubrics were in fact how primary trait scoring was operationalized in NAEP writing assessments.

As originally conceived, primary trait scoring is more accurately described as an approach to assessment rather than simply a scoring method. The primary trait approach is characterized by a concern for the test taker's ability to successfully complete specific tasks that reflect the real-world situations. Rather than viewing assessment tasks as devices for eliciting a generalizable sample of performance, performance is viewed as task specific. Contextual factors are considered critical for understanding what success means for the task, so the task context should be clearly specified. For example, in writing assessment, contextual features such as the audience and purpose for writing (e.g., persuading a peer or describing a concept to a lay audience) should be made explicit in task design and the instructions to test takers. A full implementation of a primary trait scoring approach would therefore incorporate a focus on tasks not just in scoring but at the test design and development stages as well.

Development of a primary trait scoring system starts with selecting the tasks to be tested, which in turn entails a classification system to identify the tasks needed to cover the domain targeted by the assessment. In the NAEP assessment, for example, a model of discourse was used to identify particular rhetorical functions (e.g., persuasion, description, and personal expression) that were felt to be important for student writing. Once such a framework has been established, prototype tasks are developed and test-taker samples are collected. Scoring criteria are then developed based on the performances elicited from test takers and also informed by theoretical or empirical understanding of the requirements to successfully complete the task. Scoring criteria should be limited in number but should cover key elements required for success in the task, and ideally, scoring criteria should be useful for informing teaching and learning.

Strengths and Weaknesses

An advantage of primary trait scoring is that it makes explicit a specific feature—the one that matters most for successful performance, as compared with holistic scoring, where the relative importance of the qualities being judged may be unclear, or analytic scoring, where several features are judged simultaneously. An explicit description of the key aspect of performance in turn provides more efficient and focused guidance to learners and teachers. A motivation for the development of primary trait scoring was to reinforce the link between assessment and successful real-life written communication, and the targeted guidance provided by primary trait scoring may help to focus learning on practical goals rather than abstract knowledge and abilities. Lloyd-Jones also on practical goals rather than abstract knowledge and abilities. Lloyd-Jones also argued that a focus on effectiveness in specific contexts provides a more realistic view of what people can actually do; for example, someone who is skillful in writing formal arguments may be less successful in writing an expressive narrative.

A drawback of the primary trait approach is the time and expertise required to construct separate scoring materials for multiple tasks. Within the NAEP context, Lloyd-Jones estimated it took 60–80 hours of work to produce rubrics and other scoring materials for each item, not including the time required to obtain initial test-taker responses and pilot the scoring materials with raters. This approach also requires both a conceptual framework for classifying different kinds of tasks as well as an understanding of the aspects important to successful task completion, likely requiring theoretical sophistication during test design as well as a careful analysis of examples of successful and unsuccessful performance. Moreover, while task-specific scoring may support more precise inferences regarding what someone can do in a specific real-world situation, such specificity in turn limits the range of contexts over which the scores apply. Given the considerable demands primary trait scoring places on the test developer, especially the need to produce separate scoring materials for each new task, primary trait scoring is uncommon in large-scale assessments. Even within the NAEP writing assessment, scoring eventually evolved into common scoring rubrics used across a broader range of tasks.

Larry Davis

See also Analytic Scoring; Holistic Scoring; National Assessment of Educational Progress; Performance-Based Assessment; Rubrics; Written Language Assessment

Further Readings

Arter, J., & McTighe, J. (2001). Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. Thousand Oaks, CA: Corwin Press.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27–32. doi:10.3102/0013189X018009027

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging (pp. 33–66). Urbana, IL: National Council of Teachers of English.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. Language Testing, 19(4), 395–418. doi:10.1191/0265532202lt237oa

Allison Jennifer Ames Allison Jennifer Ames Ames, Allison Jennifer

Prior Distribution

Prior distribution

1297

1298

Prior Distribution

There has been an emergence of researchers using Bayesian methods in studies and research on educational measurement, research, and evaluation. Bayesian methods differ from traditional methods in one key aspect—that of parameter uncertainty. All statistical probability models describe a mechanism, or relationship, between unobserved parameters that have given rise to observed data. In Bayesian methods, parameters are regarded as random variables to incorporate the uncertainty, and an entire distribution of possible parameter values is produced. In contrast, traditional methods consider parameters as fixed quantities, the result being a single-point estimate. Another distinction between Bayesian and traditional methods is found in the incorporation of information about the parameter before the data have been observed. This information is the *prior information* and is represented by an entire distribution. The degree of confidence in the prior distribution ranges from quite strong to very low. There are several sources of prior information, all of which can contribute to the estimation of the parameter distribution in Bayesian methods.

Bayesian Modeling Stages

To facilitate discussion of prior distributions, or *priors*, some Bayesian terms must first be introduced. There are three distinct modeling stages in the Bayesian approach, the first of which is specification of a model for the observed data, termed the *likelihood*, which represents the statistical relationship between the parameter and the observed data. Specification of prior information is the next stage. In the final stage, after the data are observed, prior information on the parameters is combined with the likelihood model to provide a distribution of parameter information. This combination of the likelihood and priors comes via Bayes’s theorem, often operationalized via Markov chain Monte Carlo methods. Because the combination occurs after the data are observed, the distribution of parameter values is known as the *posterior* parameter distribution. Posterior distributions specify the probability that each parameter equals a particular value or lies in a certain range of values.

The posterior is determined by the amount of information contained in both the likelihood and the prior. The process of constructing a posterior distribution is a blending between the prior and the likelihood. That is, the prior acts as a weight for the likelihood in the formation of the posterior. In general, if the prior information is weak, then the posterior will be relatively unaffected by the form of the prior because the prior carries little weight in the blending process with the likelihood. Similarly, if the prior information is strong, then the posterior will be significantly affected by the form of the prior because the prior carries considerable weight in the posterior’s formation.

Prior Distributions

Priors define a probabilistic model for the parameters, and a researcher has several options for incorporating the prior information into the Bayesian modeling process. Each option has varying degrees of influence, or *weight*, on the formation of the posterior. What follows is an overview of prior distribution features.

Priors can be strong and narrowly focused. Conversely, they could be weak, reflecting a less focused range of inference, but still with some informative qualities. Typically, the strength of a prior distribution is controlled by the prior distribution’s variance, termed *prior precision* or *informativeness*. Smaller prior variances demonstrate more precision. For instance, a normal (0, var = 1) prior distribution would be considered more precise and informative than a normal (0,var = 100) prior distribution because the former has a smaller variance than the latter.

Being either strong or weak, *elicited priors* are those in which parameter information is obtained from experts with information about the substantive question of interest but who are not involved in the model construction process.

Elicited priors could also arise from a collection of possible values of the parameter informed sequentially through previous studies in the area. This process is referred to as *updating*. Elicited priors could also result from common distributional families, such as the normal or gamma distributions, which are tied to parameter distributional assumptions. An often-cited drawback of elicited priors is that they can be viewed as completely subjective, in that one expert may have a differing belief about the parameter than another.

Another option for the specification of the prior is the use of *noninformative priors*. These noninformative priors are recommended when no reliable information about the parameter exists or if estimates comparable to the maximum likelihood estimate of parameters are desired. However, use of noninformative priors negates many Bayesian advantages by essentially reducing the estimation solution to the traditional one. An additional drawback is that the use of noninformative priors implies that the posterior arose from the data only and that all resulting inferences were completely objective rather than subjective.

A closely related notion to the noninformative prior is that of the *reference prior*. These are treated as a convenient place to begin an analysis. Prior distributions specified as *uniform* are often used as both reference and noninformative priors. These uniform priors are flat and indicate that the value of the parameter is equally likely across the specified range. That is, the prior has equal weight across the parameter space, and no blending occurs. The uniform prior has potential drawbacks, one of which is the possible construction of an improper posterior, resulting in invalid inferences. Caution should be taken when using such a prior. An alternative to the flat prior is the *Jeffreys prior*, which results in a proper posterior.

In the case where the posterior distributions are in the same distributional family as the prior distribution, the prior is called a *conjugate prior*. The benefit of these conjugate priors is that they can help determine posterior distributions without complex numerical integration or sampling techniques such as Markov chain Monte Carlo, a decided benefit for researchers looking to avoid complicated mathematics in the estimation of the model.

Allison Jennifer Ames

See also Bayes’s Theorem; Bayesian Statistics; Distributions; Posterior

Distribution

Further Readings

Congdon, P. (2003). Applied Bayesian modeling. Hoboken, NJ: Wiley.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). Bayesian data analysis (2nd ed.). London, UK: Chapman & Hall.

Haiyan Liu Haiyan Liu Liu, Haiyan

Zhiyong Zhang Zhiyong Zhang Zhang, Zhiyong

Probit Transformation Probit transformation

Probit Transformation

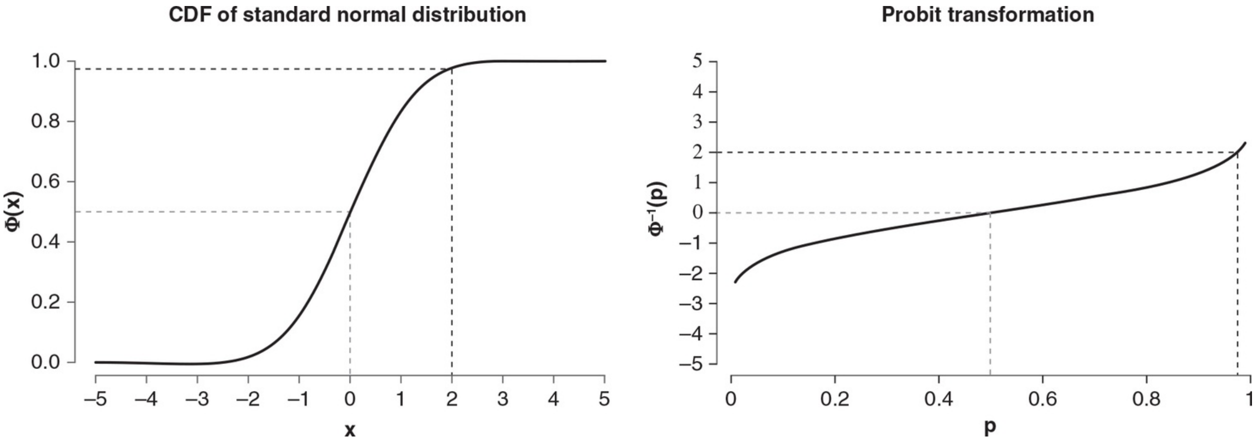
Probit transformation is widely used to transform a probability, percentage, or proportion to a value in the unconstrained interval (−∞,∞), which is usually referred to as a *quantile* in probability theory. Strictly speaking, probit transformation is the inverse of the cumulative distribution function of the standard normal distribution. For any observed value $x \in (-\infty,\infty)$, the cumulative distribution function of the standard normal distribution, denoted by $\Phi(x)$, is defined as follows:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt,$$

with t being a value that the standard normal distributed variable could take. It converts a value in the interval (−∞,∞) to a value p in the interval (0,1) such that $p = \Phi(x)$. For a probability p, or more generally any value between 0 and 1, $\Phi^{-1}(p)$ is its probit transformation to transform p to the quantile x. For instance, $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$. It is true in general that $\Phi[\Phi^{-1}(p)] = p$. For example, when p is .975, $\Phi^{-1}(.975) = 1.96$ and $\Phi^{-1}(1.96) = .975$. An appealing feature of probit transformation is that it converts a sigmoid curve to a line that is almost linear (Figure 1). The linearization brings researchers great convenience because it allows them to model a linear line directly by a linear combination of other variables.

Figure 1 The left panel is plot of the cumulative distribution function (CDF) of the standard normal distribution, and the right panel contains the plot of probit

transformation



Probit transformation is often used in modeling categorical, especially binary, outcome data. The well-known probit regression analysis exemplifies its most notable application. In binary data analysis, one is often interested in predicting the binary outcome variable *Y*. It usually assumes that there is an underlying normally distributed variable *Y** and a threshold τ such that *Y* = 0 when *Y** ≤ τ, and *Y* = 1 when *Y** > τ. Therefore,

$$p = Pr(Y = 1) = Pr(Y^* > \tau).$$

The underlying continuous variable *Y** can be analyzed by a regression model with given predictors *X*,

$$Y^* = X\beta + \varepsilon.$$

To identity the model, it is usually assumed that ε ∼ *N*(0,1). Consequently, the probit model has the following form:

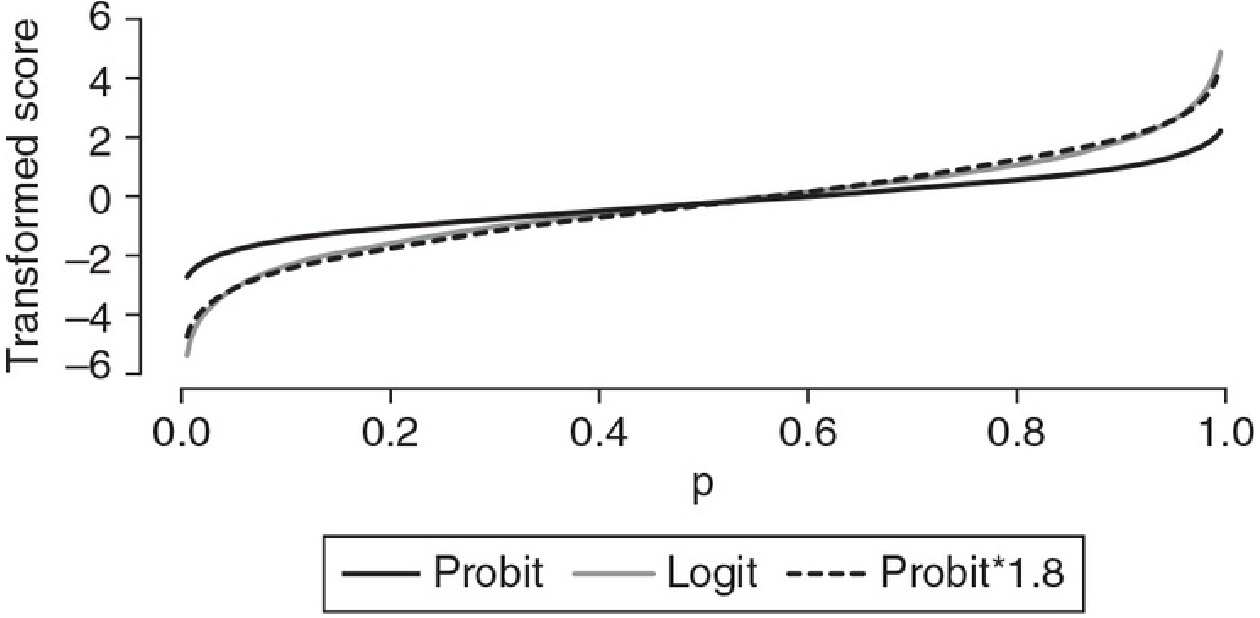
$$p = Pr(Y = 1).$$

$$\Phi^{-1}(p) = -\tau + X\beta.$$

There are other transformation methods to convert the interval (0,1) to the unconstrained interval (−∞,∞) such as the logit transformation. For a value *p* in the interval (0,1), its logit transformation is .

Although both probit and logit transformations linearize a sigmoid curve, the slopes of the two linear lines are different, as shown in Figure 2. The slope from the logit transformation is around 1.8 times as large as the one from the probit transformation. For researchers, both probit and logistic transformations have their own appealing features. In probit transformation, the underlying *Y** is assumed to be normally distributed, which is consistent with the normal assumption on the latent constructs in the social and educational sciences, while in the logit transformation, one assumes the underlying continuous variable *Y** follows a logistic distribution. The results from the logit transformation are more interpretable in terms of the odds ratio. The curves under these two transformations are hardly distinguishable when the probit transformation is scaled by 1.8.

Figure 2 Probit versus logit transformation



Haiyan Liu and Zhiyong Zhang

See also Normal Distribution

Further Readings

Bliss, C. I. (1934). The method of probits. Science, 79(2037), 38–39. doi:10.1126/science.79.2037.38

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. Annals of Applied Biology, 22(1), 134–167. doi:10.1111/j.1744-7348.1935.tb07713.x

Camilli, G. (1994). Origin of the scaling constant *d* = 1.7 in item response theory. Journal of Educational and Behavioral Statistics, 19(3), 293–295. doi:10.3102/10769986019003293

Finney, D. J. (Ed.). (1971). Probit analysis (3rd ed.). Cambridge, UK: Cambridge University Press.

Tyrell Hirchert Tyrell Hirchert Hirchert, Tyrell

Problem Solving

Problem solving

1300

1303

Problem Solving

Problem-solving ability has been called “liquid intelligence” or “street smarts.” It is the process of apprehending information, making a cognitive model of how that information is connected to a possible solution, and using that model to get a desired result. Typically, a problem-solving cycle involves defining and analyzing an issue, developing and implementing a strategy for overcoming the issue, monitoring progress, evaluating results, and repeating the cycle if necessary until a desired solution is achieved. Approaches to problem solving can vary depending on the type of problem being solved and how it is being applied as a cognitive task in the real world. For example, problem solving in reading and writing requires a different approach than it does in math, and problem solving looks different during a game of chess compared with trying to repair electronics. This entry describes the overall process of problem solving, discusses strategies and barriers to solving problems, and reviews methods of measuring problem-solving ability.

The Cognitive Process of Problem Solving

The problem-solving process typically begins by recognizing and defining the problem based on information gathered through perceptual reasoning—inductive, deductive, or some other common reasoning method. Then, working memory is activated by representing the problem cognitively through manipulation and abstraction of information in the brain. At this stage, relevant information is considered and irrelevant information is ignored as it applies to the problem at hand. The variables are then analyzed, labeled, and described, and possible strategies begin to present themselves. After this mental model has been established, the strategies are put to use in hopes of finding an anticipated solution.

Strategies to solve problems can vary widely. Analogies, in which a person finds a solution by comparing the problem to a parallel situation, are a good example of a problem-solving method. Another common strategy, used in the design thinking approach, is brainstorming or ideating, in which an unlimited number of possible solutions and ideas are synthesized into a strategy or strategies that eventually reach an optimal solution. One way that people solve particularly complex problems is by using the divide-and-conquer method, in which the larger problem is broken down into smaller, more easily manageable ones. Working backward, in which a person determines an optimal solution and then considers the steps that can lead to that solution, is an additional approach. Other problem-solving strategies include abstraction, hypothesis testing, reduction, research, root-cause analysis, and trial and error.

There are also many barriers that get in the way of how a person solves problems. *Confirmation bias* is the tendency to more easily accept information that fits a person's beliefs or experiences and reject information that is incongruent with the mental schemes the person has developed over time. *Functional fixedness* is when a person views a problem at face value or in a customary way and struggles to see all of the different options or solutions for solving it. Another barrier is *mental set*, in which a person uses only the strategies that have worked in the past instead of finding alternative ways of solving problems. Although mental set can be a useful heuristic, it can lead to incomplete solutions or ineffective solutions. Irrelevant information (i.e., misleading material presented within the context of the problem but that has no direct relationship to its solution) can also be a barrier in problem solving. Lastly, the assumptions and boundaries that a person creates about a problem can lead to constraints in a person's thinking, thus limiting the ways in which the person tries to solve the problem.

Parameters to Problem Solving

Problems can be ill-defined or well-defined. An ill-defined problem is one that does not always have a clear explanation or a clear result. An example of this would be designing an invention or finding a solution for world peace. A well-defined problem is one that has a clear, expected solution and defined paths that will lead a person to the expected conclusion. Math word problems are a good example of a well-defined problem, as all of the information for finding a solution is there.

Solution procedures are also dependent on the type of problem. Algorithms are set procedures that produce a desired solution for a particular type of problem, such as solving quadratic equations or following the instructions to use a television. Algorithms work well in well-defined problems that have set parameters and logical solutions where there is no question about the appropriateness of the procedure. They do not work well in ambiguous situations or problems that lack clear structure, however. In problems where people still have many questions about what the problem might be or how to proceed, heuristics are more useful. Heuristics are cognitive shortcuts that allow a person to solve problems or make judgments efficiently and effectively. For example, when completing a psychoeducational evaluation on a student, a good place to start would be to do a review of records. Although the solutions may be incomplete when using heuristics, this type of inductive reasoning often leads to better solutions and a better mental model of the problem.

Problem solving can also occur in many different domains and can be somewhat dependent on expertise in a given domain. Analytical problem solving often involves reasoning abilities and algorithms such as those found in math and quantitative reasoning tasks. Interactive problem solving requires the ability to solve problems that have multiple changing variables and possible outcomes based on the strategies used, such as in a game of chess. Some problems that arise involve working with other people within complex systems, such as playing in a soccer match or determining the best educational program for a student with a disability. These may require other types of more collaborative problem-solving skills. The higher the level of expertise a person has within a domain, the increased likelihood that the person will possess more problem-solving strategies in that domain.

Measuring Problem-Solving Ability

Some standardized intelligence tests have scales built into them that attempt to measure problem-solving ability; however, it is evident that nearly all scales on an intelligence test require skills in reasoning, attention, concentration, and using some sort of strategy to get a desired outcome. Reasoning ability on intelligence tests has a high correlation with problem-solving ability, as it requires the use of similar problem-solving strategies to come to a correct conclusion. Where intelligence tests struggle, however, is how to measure the process by which people solve problems, as the tests are designed to reward correct answers rather than the reasoning approaches behind them.

Scales that measure verbal intelligence typically include a task that requires verbal reasoning to reach a solution, which is a type of problem-solving task. On visual/perceptual scales, there are tasks that require an ability to solve problems using some sort of spatial reasoning to determine conceptual relationships between visual/spatial items on the test. Tests that measure fluid reasoning measure inductive and quantitative reasoning abilities as well as some abstract thinking abilities that make up a part of problem solving. One limitation in these intelligence tests currently is that there are clear, correct answers on most questions. In the real world, there can be many possibilities, solutions, and strategies to a particular problem, with an unlimited set of parameters; thus, the testing format itself places restrictions on the ability to measure problem solving as a construct in intelligence.

Some researchers have found a high correlation with general intelligence and problem-solving ability, but researchers have not entirely defined where problem solving fits within general theories of intelligence, as the process may require multiple domains of intelligence and cognitive processes. In the Cattell–Horn–Carroll theory of intelligence, problem-solving aptitude is highly correlated with the third stratum of cognitive ability, or general cognitive ability, but it is difficult to pinpoint where problem solving fits in the second stratum of intelligence or broad cognitive ability (e.g., processing speed, general memory, and visual perception).

Problem solving is also one of the cognitive abilities most susceptible to error, as it requires all of these processes to be functioning at a consistent level. This can make problem solving somewhat unreliable, as it can be error prone, inconsistent, and easily overridden by the mental schemes and biases a person has developed. Problem solving requires working memory and attentional control to work together, and when one of these abilities is inhibited—whether it is from lack of sleep, distraction, illness, anxiety, or other barriers—the ability to solve problems at a high level gets diminished. Problem solving also includes an affective element of motivation and persistence, which further adds to the complexity of measuring it as a construct of intelligence.

Well-designed intelligence tests (e.g., Wechsler Intelligence Scales, Stanford–

Well-designed intelligence tests (e.g., Wechsler Intelligence Scales, Stanford–Binet Intelligence Test, and Differential Ability Scales) measure problem-solving ability as a domain of intelligence. These tests do not rely solely on the memorization of specific facts or patterns, but rather, they measure some of the cognitive tools that people use to solve complex problems, such as processing speed, attention, working memory, and fluid reasoning. The greater a person's reasoning and problem-solving abilities are, the greater the likelihood of scoring high in a number of different scales on these tests.

Given the limited nature of traditional intelligence tests to assess problem-solving ability as a whole, researchers are turning to computer simulations to better address this issue. Some simulations have participants controlling a microworld, such as a small factory, where they are responsible for reaching certain objectives such as sales quotas or worker happiness, whereas other simulations use linear structural equation systems to measure rule identification, rule knowledge, and rule application. Computer simulations allow for an unlimited number of strategies to be used and an unlimited number of variables to be measured, enriching the ability of researchers to assess problem-solving methods, acquisition of causal knowledge, and knowledge application techniques.

Tyrell Hirchert

See also Attention; Cattell–Horn–Carroll Theory of Intelligence; Computer-Based Testing; Metacognition; Stanford–Binet Intelligence Test; Wechsler Intelligence Scales; Working Memory

Further Readings

Bransford, J. D., & Stein, B. S (1993). The ideal problem solver: A guide for improving thinking, learning, and creativity (2nd ed.). New York, NY: W. H. Freeman.

Frensch, P. A., & Funke, J. (2014). Complex problem solving: The European perspective. New York, NY: Psychology Press.

Funke, J. (2010). Complex problem solving: A case for complex cognition? Cognitive Processing, 11(2), 133–142. doi:10.1007/s10339-009-0345-0

Gonzalez, C., Thomas, R. P., & Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. Intelligence, 33(2), 169–186. doi:10.1016/j.intell.2004.10.002

Goode, N., & Beckmann, J. (2011). You need to know: There is a causal relationship between structural knowledge and control performance in complex problem solving tasks. Intelligence, 38(3), 345–552. doi:10.1016/j.intell.2010.01.001

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. Intelligence, 33(4), 347–368. doi:10.1016/j.intell.2005.03.002

Louis Lee, N. Y., & Johnson-Laird, P. N. (2013). Strategic changes in problem solving. Journal of Cognitive Psychology, 25(2), 165–173. doi:10.1080/20445911.2012.719021

Mayer, R. E. (1992). Thinking, problem solving, cognition (2nd ed.). New York, NY: W. H. Freeman.

Organization for Economic Co-operation and Development (OECD). (2010). PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. Paris, France: Author.

Sternberg, R. J., & Frensch, P. A. (2014). Complex problem solving: Principles and mechanisms. New York, NY: Psychology Press.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? Intelligence, 40(1), 1–14. doi:10.1016/j.intell.2011.11.003

Lauren M. Henry Lauren M. Henry Henry, Lauren M.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Process Evaluation

Process evaluation

1303

1306

Process Evaluation

Process evaluation compares presupposed objectives with actual inputs, activities, and outputs to determine whether and why the objectives have or have not been met. It is one of many tools (e.g., needs assessment, outcome evaluation, and impact evaluation) employed in evaluating programs. A *program* can be broadly described as a system of resources and activities allocated to the advancement of one or more goals; programs include interventions, services, and policies initiated and executed by public, nonprofit, or private providers at one or multiple locations. This entry describes (a) the history of process evaluation, including its materialization and proliferation as a facet of program evaluation, (b) the fit of process evaluation within the context of program evaluation, (c) the structure and flow of process evaluation from planning and implementation to analysis and reporting, and (d) the purpose and value of process evaluation, including the ways in which process evaluation informs decision making within programs and the broader scientific community.

History

Under the presidential administrations of John F. Kennedy and Lyndon B. Johnson during the 1960s and 1970s, the War on Poverty and Great Society agendas yielded numerous large-scale, federally funded programs (e.g., Head Start, Job Corps, and Community Action Program) to promote social reform in the United States. From 1950 to 1979, funding for social programs grew by 600% after inflation. Increased economic investment contributed to a desire for empirical confirmation that program outcomes were meeting expectations, leading to the emergence of government-mandated evaluations.

Despite optimism that evaluation would systematically identify effective programs for adoption or continuation and ineffective programs for termination, the majority of the War on Poverty and Great Society programs showed disappointing effects. Furthermore, explanations as to why desired results were not achieved proved impossible to ascertain. Evaluators typically assessed program outcomes to determine the extent to which objectives had been reached, without consideration for implementation fidelity. *Implementation fidelity* (also referred to as *integrity*), or the degree to which programs are carried out as intended by their developers, may moderate the relation between program exposure and intended outcomes. To gain insight as to the reasoning behind program effects or lack thereof, evaluators soon moved beyond "black box" methodology to incorporate process evaluations into their evaluation plans.

By the late 1990s and early 2000s, the number of published studies that included process evaluation components had grown exponentially. Researchers have attributed the proliferation of process evaluation literature to the execution of increasingly complex social and behavioral interventions incorporating multiple levels, sites, and target groups. Moreover, process evaluations collect both quantitative data (e.g., questionnaires and population statistics) and qualitative data (e.g., semistructured interviews, focus groups, and direct observations), and a waxing appreciation for the value of qualitative research in the scientific community facilitated the publication of research using such methods.

The demand for program evaluation has continued to grow as funding agencies (e.g., federal, state, and local government, nonprofit organizations, and private entities) become increasingly interested in understanding how their investments have been used and what they have produced. In some cases and places, program evaluations have been incorporated into the law; the Chief Financial Officers Act, signed into law by President George H. W. Bush in 1990 to improve governmental management of finances, requires U.S. federal agencies to report on program evaluations.

Process Evaluation in Context

Evaluation methods typically file under one of the two functional categories: *formative* or *summative*. During the development of a new program, or the adaptation or modification of an existing one, formative evaluations establish whether the program is feasible and appropriate and guide necessary modifications prior to full-scale implementation. Needs assessments (identifying and prioritizing gaps between current and desired conditions and selecting the most important ones to resolve) and evaluability assessments (determining whether a program is ready for summative evaluation) are examples of formative evaluations. Following implementation, summative evaluations are used to assess program effectiveness and inform decision making about program modification, continuation, or termination. Summative evaluations include outcome evaluations (measuring the degree to which programs meet short-term goals), impact evaluations (measuring the degree to which programs meet long-term goals), cost-effectiveness analyses (relating program expenditures to outcomes), and cost–benefit analyses (relating program expenditures to dollar value outcomes). Process evaluations are unique in that they can be used for both formative and summative purposes.

Structure and Flow of Process Evaluation

Considering their dual formative and summative roles, process evaluations can be introduced during program development, implementation, analysis, or modification. However, process evaluations are most valuable when integrated throughout the life cycle of a program. Process evaluations should be planned and monitored by a team of evaluators in collaboration with key stakeholders or those who have a vested interest in the program (i.e., funders, designers, and delivery staff). The team should be composed of people who embrace the iterative nature of process evaluations; techniques, tools, and designs must often be selected, tested, and revised. Although several program evaluation plans (e.g., prevention plus III, community coalition action theory, getting to outcomes, and the Centers for Disease Control and Prevention framework) have been developed and are available to teams, few process evaluation frameworks exist. Those that have been established (e.g., Allan Steckler and Laura Linnan’s 10-step guide) recommend that programs adhere to the following steps when designing and implementing a comprehensive process evaluation:

1. solidify the theory underlying program goals and create a logic model (i.e., a graphic outlining the intended relation between program inputs, activities, outputs, and projected outcomes) cataloging realistic and measurable objectives for each activity;

2. select and prioritize process evaluation questions to be answered (e.g., regarding reach, dose, and fidelity);
3. identify or create measurement tools to assess process objectives;
4. design, implement, and administer quality control assurances for process data collection and management;
5. collect, manage, and analyze process data; and
6. report findings to key stakeholders and use takeaways to inform decision making.

Process evaluations can be designed to answer any number of questions about program implementation. With program goals and evaluation efforts in mind, stakeholders must select for inclusion and, due to likely financial and human resource constraints, prioritize questions within their comprehensive process evaluation plan (as indicated in Step 2). Process evaluations often investigate program reach (portion of the target population that receives the program), dose delivered (units of each program activity provided to the target population), dose received (extent to which the target population engages with the program activities that reach them), contamination (whether the target population receives interventions from outside the program and the extent to which the control group inadvertently receives the program), and fidelity (extent to which the program delivery adheres to the original protocol). To further illuminate the nature of the relation between program exposure and outcomes, process evaluations can also collect data on variables such as context (aspects of the physical, social, political, and economic environment that influence implementation) and participant satisfaction.

Purpose and Value of Process Evaluation

Process evaluation allows stakeholders to determine *whether* their program does or does not function as intended. Although outcome evaluation can be used to determine the extent to which programs produce intended effects, in isolation, it is impossible to know whether those effects stem from program design or extraneous variables. Although statisticians use “Type I error” to describe the probability of rejecting the null hypothesis when it is true and “Type II error” to describe the probability of failing to reject the null hypothesis when it is false, evaluators use “Type III error” to describe the probability of correctly rejecting the null hypothesis for the wrong reason. Process evaluations are used to prevent Type III errors.

Process evaluation also allows stakeholders to determine *why* their program does or does not function as intended. Careful planning and execution of process evaluations can determine the mechanisms by which effects are produced. For example, a smoking cessation intervention for adolescents may be deemed ineffective based on outcome data showing no difference between pretests and posttests of self-reported cigarette use. However, an analysis of process data may indicate implementation failure in dose delivered, such that only 30% of teachers distributed informational packets detailing the harmful side effects of exposure to nicotine to their students. Program evaluation facilitates decision making, and the incorporation of process evaluation allows for those decisions to be strategic. Using process evaluation data, stakeholders can make informed value judgments, which may simplify seemingly difficult choices about whether to continue, discontinue, or modify current practices and programs.

Process evaluation necessitates rigorous planning, the results of which have impacts beyond enabling process data collection. As a program is described, variables operationalized, and constructs refined, organization of the program itself improves, and a deeper understanding of underlying theory and structure is achieved among stakeholders. In addition, the consensus required on various decisions throughout the planning process can facilitate stakeholder buy-in. Awareness of and commitment to the program and evaluation process increase the likelihood that stakeholders will support evaluation efforts and advocate for the program and decrease the odds that evaluation results will be ignored, criticized, or resisted. The thoughtful planning associated with process evaluation also benefits the community at large. Detailed program descriptions amplify the probability of fidelity in program replication. Furthermore, enhanced organization increases the possibility of and accuracy in knowledge transfer for scholarly endeavors, including the advancement of theories for constructs implicated in specific programs (e.g., interventions to promote coping in children of depressed mothers) as well as new contributions to the program evaluation literature.

Lauren M. Henry and Marc H. Bornstein

See also Applied Research; Evaluation; Formative Evaluation; Program Evaluation; Type III Error

Further Readings

Haracchi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: Using process evaluation measures to assess implementation and theory building. American Journal of Community Psychology, 27(5), 711–731. doi:10.1023/A:1022194005511

McLaughlin, M. W. (1984). Implementation realities and evaluation design. In R. L. Shotland & M. M. Mark (Eds.), Social science and social policy (pp. 96–120). Beverly Hills, CA: SAGE.

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. BMJ (Clinical Research Ed.), 332(7538), 413–416. doi:10.1136/bmj.332.7538.413

Royse, D., Thyer, B. A., & Padgett, D. K. (Eds.). (2015). Program evaluation: An introduction to an evidence-based approach (6th ed.). Boston, MA: Cengage Learning.

Saunders, R. P., Evans, M. H., & Joshi, P. (2005). Developing a process-evaluation plan for assessing health promotion program implementation: A how-to guide. Health Promotion Practice, 6(2), 134–147. doi:10.1177/1524839904273387

Scheirer, M. A. (1994). Designing and using process evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), Handbook of practical program evaluation (pp. 40–68). San Francisco, CA: Jossey-Bass.

Steckler, A. B., & Linnan, L. (Eds.). (2002). Process evaluation for public health interventions and research. San Francisco, CA: Jossey-Bass.

Valeisha M. Ellis Valeisha M. Ellis Ellis, Valeisha M.

Professional Development of Teachers Professional development of teachers

Professional Development of Teachers

Professional development of teachers involves a wide variety of specialized workshops, training, education, and advanced professional learning opportunities intended to help teachers improve their professional knowledge, competence, skill, and effectiveness. Professional development of teachers has been a major focus of school reform efforts in recent years. This entry first describes the core features of effective professional development of teachers, and the challenges to ensuring professional development are effective. It then looks at how professional development programs for teachers are evaluated.

Effective professional development of teachers deepens content knowledge, transforms teaching practices, and fosters individual and group learning. There are several key factors to consider in designing professional development programs for teachers. Professional development is most effective when teachers are treated as professionals and active learners who construct their own understanding and the development programs are situated in classroom practice.

Several core features have a positive impact on teachers’ self-reported knowledge, skills, and changes in classroom practice. These include a focus on content knowledge and numerous opportunities for active learning (e.g., reviewing student work samples, reviewing feedback on teaching). Also, professional development activities are most effective when they are congruent with other learning activities that teachers take part in, including professional conversation with other teachers.

Certain structural features of professional development also can positively affect teacher learning. The collective participation of teachers from the same school, grade, or subject can have a positive impact on professional learning. In addition, activities that are sustained over time support professional learning and addition, activities that are sustained over time support professional learning and classroom practices.

Although the benefits of effective professional development for teachers are numerous, there are also several challenges. High-quality professional development for all teachers is expensive. In addition, it is difficult to assess and evaluate teacher knowledge and instructional practices.

The poor reputation of traditional, professional development often overshadows effective professional development models. Although professional development of teachers is often required, facilitators and administrators cannot mandate professional learning. The constant negotiation of content, purpose, control, and discourse style is often discouraging for administrators and facilitators of professional development. The intense and sustained hard work and unstable nature of funding are also challenges that require consideration.

Evaluation of professional development can provide evidence that programs were effective in strengthening teachers’ content knowledge, instructional practices, and student learning. The planning evaluation is the groundwork for all other evaluations that provide understanding of what should be accomplished, implementation guidelines, and desired outcomes.

Formative evaluation is ongoing and provides evidence of whether professional development is going as planned, documents progress, and identifies areas for improvement. Summative evaluation is completed at the end of professional development and describes what was accomplished, the strengths and weaknesses of the program, and its value to the participants. The participants’ reaction, participants’ learning, organization support and change, participants’ use of new knowledge and skills, and student learning outcomes are the five levels of evaluating professional development.

Valeisha M. Ellis

See also Teacher Evaluation; Teachers’ Associations

Further Readings

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. Educational Researcher, 33(8), 3–15.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. Educational Evaluation and Policy Analysis, 24(2), 81–112.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38(4), 915–945.

Guskey, T. R. (1999). Evaluating professional development. Thousand Oaks, CA: Corwin Press.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4–14.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. Review of Research in Education, 24, 173–209.

Valeisha M. Ellis Valeisha M. Ellis Ellis, Valeisha M.

Professional Learning Communities Professional learning communities

1307

1308

Professional Learning Communities

A professional learning community (PLC) is a collaborative work culture for teachers whereby reflective understanding is gained with the aid of peers who have the same experiences, thereby improving professional knowledge and student learning. Professional learning communities are often used in education as a model for program evaluation and continuous program improvement. This entry describes the essential characteristics of PLCs, their effectiveness, the processes involved in creating or developing them, and their significance in education research, measurement, and evaluation.

Professional learning communities have several central features. Shared values and a common vision are qualities that are considered foundational to PLCs, as is collective responsibility for student learning. Effective PLCs in districts and schools address the current challenges or problems that are most critical for the school. There is reflective professional inquiry within the PLC in which dialogue regarding curriculum, instruction, and student growth is constant. PLCs support teachers’ self-efficacy and level of professionalism; this support is given from all levels of the school system (i.e., all stakeholders). PLCs not only foster an atmosphere of trust, but the ongoing work of the PLC is monitored by the stakeholders, and constructive feedback is provided. The process to create or develop a PLC is unique and should include all of the characteristics described above.

Studies have shown that there is a strong relationship between effective PLCs in schools and improved student achievement and teacher learning and instruction. PLCs promote advances in teaching practices, such as documented changes in the professional culture of the school (e.g., collaboration, a focus on student learning, teacher authority, and continuous teacher learning) associated with them. The improvements in student learning brought about by PLCs indicate that them. The improvements in student learning brought about by PLCs indicate that targeted focus on learning is critical for achievement gains.

Professional learning communities provide a significant contribution to education research on improving student learning, teacher learning, and instructional practices. They are a shift away from traditional models of professional development, combining the knowledge and theory of teachers with current understanding on how to best make an impact on student and teacher learning. In other words, collaboration is the process, and the goal is student learning (analysis of student work) as well as improved teacher learning and instructional practices.

Valeisha M. Ellis

See also Professional Development of Teachers

Further Readings

Averitt, S. (2017). Raising student achievement using a multi-tiered system of supports: A problem-based organizational study (Doctoral dissertation). East Carolina University, Greenville, NC.

DuFour, R., & DuFour, R. (2013). Learning by doing: A handbook for professional learning communities at work. Bloomington, IN: Solution Tree Press.

Green, T. (2017). Dimensions of professional learning communities and student achievement in data coaching schools (Doctoral dissertation). Delta State University, Cleveland, MS.

Watson, C. (2014). Effective professional learning communities? The possibilities for teachers as agents of change in schools. British Educational Research Journal, 40(1), 18–29. doi:10.1002/berj.3025

Yuko Goto Butler Yuko Goto Butler Butler, Yuko Goto

Proficiency Levels in Language

Proficiency levels in language

1308

1309

Proficiency Levels in Language

In language education, *language proficiency* refers to one’s general language knowledge and skills for using the target language for various communicative purposes. However, researchers disagree about what proficiency entails, with the differences reflecting their theoretical orientations.

For example, generative linguists, influenced by the work of Noam Chomsky, restrict their focus to grammatical knowledge. Dell Hymes, a sociolinguist, expanded the scope to emphasize the social appropriateness of language use. He proposed as a model *communicative competence*, which includes both knowledge (not only grammatical knowledge but also sociolinguistic knowledge) and the ability for use (one’s potential ability to use the language in socially appropriate ways).

Although Hymes’s proposed model was situated in the first-language (L1) contexts, his notion of communicative competence greatly influenced succeeding models of language proficiency for second-language (L2) learners. For example, Michael Canale and Merrill Swain’s communicative competence, one of the most influential models in L2 acquisition, was composed of grammatical, sociolinguistic, and strategic competencies. Lyle Bachman and Adrian Palmer further elaborated Canale and Swain’s model and proposed a hierarchical model of communicative language ability, which consisted of organizational knowledge (grammatical and textual knowledge), pragmatic knowledge (functional and sociolinguistic knowledge), and strategic competence (a series of metacognitive strategies).

Jan H. Hulstijn, in an attempt to explain individual differences among L1 users (native speakers) as well as among L2 users, proposed a language proficiency model composed of core and periphery elements. Core elements are linguistic knowledge and processing skills (speed) in phonetic, phonological, morphosyntactic, and lexical domains. Periphery elements are metacognitive competencies, including metalinguistic knowledge, strategic competence, and so forth. In Hulstijn's model, core elements can be divided into two types: basic language cognition and high language cognition. Basic language cognition is restricted to oral domains only (listening and speaking) and is attainable by all L1 users, whereas the degree of attainment of high language cognition differs substantially among L1 users with different age, educational, and literacy levels. Hulstijn predicts that L2 users can develop a high level of high language cognition depending on their educational and literacy levels in their L1 but may not be able to fully develop basic language cognition in their L2.

Although these multiconstruct models of language proficiency have been widely adopted in language education, researchers have found it rather difficult to provide empirical evidence to support the construct configurations of their models.

Despite lack of consensus on what accounts for language proficiency, the notion of proficiency levels continues to be frequently employed in practice. Learners may be arranged in different programs according to their proficiency levels, such as beginning, intermediate, and advanced. Similarly, curriculum and material may be developed in sequence from lower to higher proficiency levels. Various kinds of L2 assessments are available to identify learners' proficiency levels as well. The Foreign Language Assessment Database developed by the Center for Applied Linguistics, for example, contains information on more than 200 proficiency tests for identifying one's proficiency levels in more than 90 languages. It is important to note that language proficiency tests, unlike achievement tests, are not tied to a particular course, curriculum, or program.

Language proficiency assessments are based on various kinds of language proficiency scales. Such scales were developed differently depending on how language proficiency is conceptualized and how the assessments are intended to be used. Many scales are organized by four skill domains (i.e., listening, speaking, reading, and writing), but others can be further broken into different contexts of use (e.g., academic contexts and professional work contexts), whereas still others take holistic approaches. Some of the major proficiency scales include the American Council on the Teaching of Foreign Languages'

Proficiency Guidelines in the United States, and the *Common European Framework of Reference for Languages* (CEFR) developed by the Council of Europe.

American Council on the Teaching of Foreign Languages' *Proficiency Guidelines* are designed to capture what an individual can do in a foreign language in listening, speaking, reading, and writing in the real-world language-use situations and have been used primarily in academic and professional workplace contexts. The guidelines are composed of five major levels (novice, intermediate, advanced, superior, and distinguished), and each level (except superior and distinguished) is further divided into three sublevels (low, mid, and high); there are 11 levels altogether for each skill domain. Since its publication in 1986, American Council on the Teaching of Foreign Languages' *Proficiency Guideline*s have been influential in U.S. foreign language education.

CEFR is a framework of language learners' functional abilities to use the target language, aiming to provide common and comprehensive standards for curriculum/material development, instruction, and assessment practices and use across Europe. CEFR is organized into three dimensions: language activities (receptive, productive, interactive, and mediation activities); domains where the activities are conducted (educational, occupational, public, and personal domains); and competence levels at which learners can perform and develop when they engage in the activities (A1, A2, B1, B2, C1, and C2 levels). CEFR has had a substantial impact on language education in Europe, and increasingly in other regions across the globe, by prompting the creation of various professional development and language-learning activities, most notably in the field of assessment. But, it has been criticized for a number of reasons, including that (a) CEFR's level descriptors were developed primarily by relying on professionals' judgment without having a basis in second-language acquisition research; (b) because of its original intention to be flexibly applicable across different programs and languages, the wording in descriptors is too vague; and (c) CEFR is inadequate as an assessment framework. Considering that the highest levels (C1 and C2) in CEFR require language use with high intellectual knowledge and skills, Hulstijn has argued that such high levels cannot be attainable unless a learner has a high educational level.

Yuko Goto Butler

See also *Reading Comprehension*; *Reading Comprehension Assessments*

Further Readings

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge, UK: Cambridge University Press.

Hulstijn, J. H. (2015). Language proficiency in native and non-native speakers: theory and research. Amsterdam, the Netherlands: John Benjamins.

Websites

American Council on the Teaching of Foreign Languages: http://www.actfl.org

Foreign Language Assessment Database: http://webapp.cal.org/FLAD/

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Li Lin Li Lin Lin, Li

Jianqiang Liang Jianqiang Liang Liang, Jianqiang

Program Evaluation

Program evaluation

1310

1314

Program Evaluation

Program evaluation refers to the systematic, scientific, and rigorous investigation of a program’s effectiveness. In education research, for example, such evaluations examine the goal attainment and outcomes of programs designed to promote student, teacher, and/or school performance. Through the evaluation of educational programs, the credibility and accountability of related education entities (i.e., curricula and educational services) and educational systems can be assessed and improved. This entry outlines the development of program evaluation, discusses some of the guidelines established for effective evaluation, and introduces three main approaches: quantitative design, qualitative design, and mixed-methods design.

History and Development

The modern development of program evaluation dates back to the 1960s. In 1967, an American sociologist, Edward Suchman, suggested using Donald Campbell and Julian Stanley’s book *Experimental and Quasi-Experimental Designs for Research* as an appropriate guide for developing evaluation designs. This suggestion brought the qualitative evaluation approach, featured by experimental and quasi-experimental designs, to the forefront of program evaluation methods. From the 1960s to the 1990s, researchers debated whether experimental and quasi-experimental designs should be the standard usage for program evaluation, as opposed to approaches such as needs assessment, the client satisfaction approach, and cost–benefit analyses.

In 1978, Michael Patton published a seminal book titled *Utilization-Focused Evaluation*, which argued for the use of a more qualitative approach to program evaluation. Patton contended that qualitative evaluation methods, such as interviews and observations, generated insights beyond numerical data, enriching the understanding of a program for its participants and other stakeholders. In 1980, Patton wrote another book titled *Qualitative Evaluation Methods*, which became the first textbook for the application of qualitative methods to program evaluation.

Researchers with a more pragmatic outlook, wanting to assess particular goals within a particular context rather than using independent and objective measures of assessment, argued that in order to have a comprehensive understanding of program effectiveness, both quantitative and qualitative methods should be used. Since then, many evaluators have combined quantitative and qualitative approaches to triangulate the results of one topic or to address the different facets of program effectiveness. In the field of evaluation, it is believed that the worthiness of a program is better understood with multiple approaches and multiple program stakeholders; thus, combining the quantitative and qualitative methodologies into a mixed-methods approach was proposed in the 1970s.

In the 2000s and 2010s, evaluation was influenced by the evidence-based practice movement. Evidence-based practice, initially applied to medical research, argues for the use of scientific rigor as a basis for the assessment of a program. Advocates maintained that more emphasis needed to be placed on high-quality evidence to inform decisions regarding program outcomes, as opposed to the opinions and theories of policy makers, professionals, and researchers. As such, systematic and rigorous program evaluation has become increasingly important, and evaluation has become a specialized and interdisciplinary field involving multiple methods.

Standards for Program Evaluation

With more attention focused on methodological appropriateness in evaluation decision making, many professional societies and associations established national and international standards to improve the quality of program evaluation. For example, in 1999, the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education jointly proposed and updated the *Standards for Educational and Psychological Testing*, which clarified validity, reliability, and errors in measurement and highlighted the importance of testing standards for education professionals and researchers.

In addition to guidelines for methods and procedures, standards for measuring whether an evaluation has met its intended objectives were established. In 2011, for example, the Joint Committee on Standards for Educational Evaluation, a coalition of professional North American associations, proposed five attributes indicative of sound and fair evaluation of educational programs: utility (the extent to which the evaluation meets the needs of the stakeholders), feasibility (effectiveness and efficacy of evaluation), propriety (appropriateness, legitimacy, and justice in evaluation), accuracy (truthfulness of evaluation), and accountability (responsibility for evaluation process and products). The Joint Committee on Standards for Educational Evaluation also highlighted the importance of ongoing stakeholder involvement in educational evaluation research.

A third area of focus was on the professional standards of the evaluators themselves. For example, the American Evaluation Association's *Guiding Principles for Evaluators*, first issued in 1994, proposed that the ethical standards of the evaluation profession be defined by five general areas:

1. Systematic inquiry: Assessments should be presented clearly and be based on data-based, reproducible approaches.
2. Competence: Evaluators should be adequately trained and have the necessary experience, including demonstrating cultural and social competence, when working with participants and stakeholders.
3. Integrity/honesty: Evaluators should communicate clearly with clients and relevant stakeholders, should not misrepresent their findings, and should disclose any potential conflicts of interest with their role in a particular project.
4. Respect for people: Evaluators should follow professional ethics standards regarding informed consent for participants and should respect the dignity and self-worth of participants and stakeholders.
5. Responsibilities for the general and public welfare: As evaluators often have obligations that encompass the public interest, particularly when they

> are supported by publicly generated funds, they should consider the perspectives and interests of not only the full range of stakeholders but at times the welfare of society as a whole.

Program Evaluation Methods

Quantitative Evaluation Designs

Quantitative evaluation relies on the statistical analyses of numerical data in an aggregated manner to infer conclusions regarding the success of a specific program, with the intention of generalizing the findings to the population. Relevant evaluation questions include “How successful are the outcomes of a program?” “How well has the program been implemented?” and “What factors have contributed to the program’s effectiveness?” Program evaluators often collect data from program participants, program implementers, or other stakeholders via surveys, polls, second-hand documents, or other objective measurements to answer these questions.

There are three types of quantitative evaluation methods: *true experimental design, quasi-experimental design,* and *preexperimental design*. In a true experimental design, such as randomized controlled trials, participants are randomly assigned to either receive treatment (the experimental group) or receive no treatment or placebo (the control group). These two groups are assessed both before and after the program using valid measures. By comparing the changes in the experimental group with the control group during the same period, program evaluators can understand the effect of the treatment on outcomes. Through randomization, the intrinsic differences between the experimental group and the control group are expected to be minimized, and thus one can draw conclusions that the outcome is due to the intervention of the program.

There are several strengths of true experimental design. First, it minimizes selection biases through randomization. Second, it eliminates the influence of unwanted extraneous variables on program outcomes. Third, it helps the experimenters understand the causal effect of independent variables on the dependent variables. Therefore, using true experimental design minimizes the threats to internal validity (the extent to which a casual-effect inference based on a study is true). On the other hand, such designs can result in increased threats to external validity (the extent to which the study findings can be generalized to other populations) and ecological validity (the extent to which the study findings can be generalized to the real world). With so many factors being controlled, the results of an experimental evaluation may not represent a similar case in the actual world. In addition, employing true experimental design may be ethically problematic, as it may be difficult to justify why some participants are assigned to receive the treatment while the others are not. It is also time consuming and costly to administrate.

Compared with experimental design, quasi-experimental design is more practical and flexible with reference to the constraints of real life. Quasi-experimental design is similar to true experimental design, yet the allocation of people to different conditions is not based on randomization but on preexisting criteria. One typical example is the nonequivalent groups design. To evaluate the outcomes of taking a leadership course, for instance, evaluators can compare the students in the course (i.e., the experimental group) with other comparable students who have never taken the course (i.e., the control group). During the recruitment of the control group, evaluators can match their sociodemographic backgrounds with the experimental group to ensure the groups have comparable background demographics. Another type of quasi-experimental design is the time series design. By collecting data at multiple time points before and after the implementation of a program, researchers can use this method to understand trends in participants' performance across time. Although quasi-experimental design has greater ecological validity, it suffers from lower internal validity relative to true experimental design. Specifically, it is difficult to draw causal-effect conclusions because participants have not been randomly assigned to the groups.

The least rigorous quantitative method is preexperimental design, in which program outcomes are examined without comparison to a control group. The single-group pretest–posttest design is one example. For instance, to understand change in elderly people in an active-aging educational workshop, evaluators can invite the participants to report their attitudes toward active aging both before and after the workshop and compare their results. Preexperimental design is highly susceptible to threats of internal validity. Therefore, causal conclusions about the influence of the program on outcomes can hardly be drawn. However, considering its high feasibility, evaluators adopt this design to understand the outcomes of a program among participants.

There are multiple statistical methods for analyzing quantitative evaluation data

There are multiple statistical methods for analyzing quantitative evaluation data. To understand the changes in participants before and after a program is implemented, methods for treating longitudinal data can be employed. In a balanced repeated-measures design with equal group sizes, generalized linear models such as analysis of variance and analysis of covariance can be used to analyze the time changes. In more complex designs with more than two assessment time points, particularly when the assessment intervals are not constant and the sample sizes are not equivalent across time, advanced longitudinal models such as a linear mixed model can be used. Evaluators can compare the initial responses and the developmental trajectories of participants and nonparticipants over time to understand a program's effect on them.

In addition, as evaluators are often interested in what factors contribute to program effectiveness, statistical methods that test the relations between variables can be employed. For example, evaluators may want to know if participants' previous academic achievements will influence the effect of a new teaching method or if years of teaching experience will influence the outcomes of a new curriculum. Correlation analysis, multiple linear regression, and structural equation modeling can be applied to understand the impact of individual characteristics on program outcomes.

Qualitative Evaluation Designs

Although quantitative evaluation helps evaluators obtain an aggregate picture of program effectiveness through data analysis, it is less useful at answering questions about the "meaning" behind the overall picture. In addition, when the responses to a program's effectiveness are obtained using predefined measures, such as Likert-type scales, responses that are out of the evaluators' expectations are often excluded; program stakeholders may have different perspectives that the evaluators have not considered. Qualitative evaluation methods provide evaluators with detailed, in-depth, and personal information via narratives, excerpts, diaries, images, and/or written material from different program stakeholders, which can aid in understanding what happens during program implementation. More importantly, the information is based on the perceptions of the respondents, which enables evaluators to capture ideas not constrained by any preexisting mind-sets and experiences.

There are three kinds of qualitative data: interviews, documents, and observations. Interview data can be obtained via informal conversational discussions, through a general interview guide, or by using a standardized open-discussions, through a general interview guide, or by using a standardized open-ended interview approach; the interview unit can be either an individual or a group. For instance, program evaluators may initiate a spontaneous conversation with a few participants (the informal conversational interview approach), interview the participants based on a guide listing major topics that need to be covered (the general interview guide approach), or interview the participants following a structured guide with fixed questions given in a fixed sequence (the standardized open-ended interview approach). Written documents for program evaluation can include responses to open-ended questions in a survey. Other written materials, such as personal diaries, program records, memoranda, correspondence, official publications, and reports, can also be used. To obtain observation data, evaluators can serve as independent observers to see how a program is operating, such as how a new training course is delivered (i.e., nonparticipant observation). Evaluators can also take an active role in the program while observing its operation (i.e., participant observation), for example, by working as a teaching assistant during an educational program. Such qualitative evaluation methods can yield rich and deep information from the viewpoint of the participants or other program stakeholders being studied.

Qualitative evaluation has some intrinsic limitations. First, it is labor and time intensive to conduct qualitative evaluations and analyze qualitative data. For an interview transcript or personal diary, the identification, processing, and categorizing of relevant information (a procedure known as *coding*) usually requires several rounds of iteration, with multiple coders involved. Second, the process of collecting the data and interpreting the findings is highly affected by the experiences, skills, knowledge, and even the personal views of the evaluators. For example, the quality of interviews is subject to the skill of the interviewers to elicit sufficient information and enable interviewees to share their ideas in an accurate and honest way. Finally, the generalizability of the findings from a qualitative evaluation is limited, as it is usually based on a small sample of participants. Therefore, the insights generated from qualitative evaluations may contain personal biases.

Mixed-Methods Designs

As quantitative and qualitative evaluation both have advantages and disadvantages, mixed-methods designs, which combine both methods of evaluation, has become increasingly more common. By using multiple evaluation approaches, informants, and data sources, researchers can overcome the intrinsic disadvantages of each approach and generate more accurate and valid results for a topic. This is the process of *triangulation*. The mainstream view is that educational evaluators should triangulate findings from different evaluation methods, program stakeholders, and data sources to generate convergent results about the effectiveness of a program.

According to the degree of synthesis of the different methods, there are two broad classes of mixed-method designs: component designs and integrated designs. In a component design, the mixing of methods occurs at the final stage, during result interpretation and conclusion making. Researchers usually carry out different studies and analyze the data sets separately. Different methods can be used to address a single research question, with the objective of accumulating convergent evidence (triangulation designs), to supplement one major method by clarifying or extending the results (complementary designs) or to meet different aspects of the inquiry (expansion designs).

In contrast, in an integrated design, the mixing of methods occurs throughout the evaluation process. Researchers synthesize multiple methods into the design, implementation, and data analysis of the evaluation. Disparate methods can be interplayed over time, during which one method may inform the development of another (iterative designs). In addition, a given method can be implemented within another (e.g., using observation to measure the outcomes of an experimental design), with the different approaches mutually strengthening each other (embedded designs). They can also be conducted under a substantive conceptual framework, known as a *concept map*, and all study designs and data interpretation can be derived from this framework (holistic designs). Lastly, they can be conducted to serve divergent program positions and value stances, where dialog across the different ideologies is highly encouraged (transformative designs).

With the growth of program evaluation as a field of research, evaluators have developed a repertoire of methods and techniques as well as sophisticated approaches to choosing and combining methodologies. According to the utilization-focused evaluation approach, the adoption of methodologies needs to serve the usability of the intended users. Therefore, evaluators should work with the intended users and the stakeholders of a program in selecting questions for investigation, determining the appropriate methodology to address these questions, and implementing evaluation procedures.

Daniel Tan-lei Shek, Li Lin, and Jianqiang Liang

See also Experimental Designs; Joint Committee on Standards for Educational Evaluation; Mixed-Methods Research; Preexperimental Designs; Process Evaluation; Program Theory of Change; Qualitative Research Methods; Quantitative Research Methods; Quasi-Experimental Designs

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Evaluation Association. (1994). Guiding principles for evaluators. Washington, DC: Author. Retrieved from http://www.eval.org/p/cm/ld/fid=51

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally.

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms (New Directions for Evaluation, No. 74). San Francisco, CA: Jossey-Bass.

Joint Committee on Standards for Educational Evaluation. (2011). The program evaluation standards. Thousand Oaks, CA: SAGE.

Patton, M. Q. (2008). Utilization-focused evaluation (4th ed.). Thousand Oaks, CA: SAGE.

Patton, M. Q. (2015). Qualitative research & evaluation methods: Integrating theory and practice (4th ed.). Thousand Oaks, CA: SAGE.

Royse, D., Thyer, B. A., & Padgett, D. K. (2010). Program evaluation: An introduction. Belmont, CA: Wadsworth.

Shek, D. T. L., Tang, V. M., & Han, X. Y. (2005). Evaluation of evaluation studies using qualitative research methods in the social work literature (1990–2003): Evidence that constitutes a wake-up call. Research on Social Work Practice, 15(3), 180–194. doi:10.1177/1049731504271603

Stufflebeam, D. L. (1994). Professional standards for educational evaluation. In T. Husen, T. N. Postlethwaite, & H. J. Walberg (Eds.), International Encyclopedia of Education (2nd ed.). Oxford, UK: Pergamon Press.

Stufflebeam, D. L. (2001). Evaluation models: New directions for evaluation (No. 89, Spring 2001). San Francisco, CA: Jossey-Bass.

Suchman, E. A. (1967). Evaluative research: Principles and practice in public service and social action programs. New York, NY: Russell Sage Foundation.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2010). The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.). Thousand Oaks, CA: SAGE.

Natalie D. Jones Natalie D. Jones Jones, Natalie D.

Benjamin D. Rosenberg Benjamin D. Rosenberg Rosenberg, Benjamin D.

Program Theory of Change Program theory of change

1314

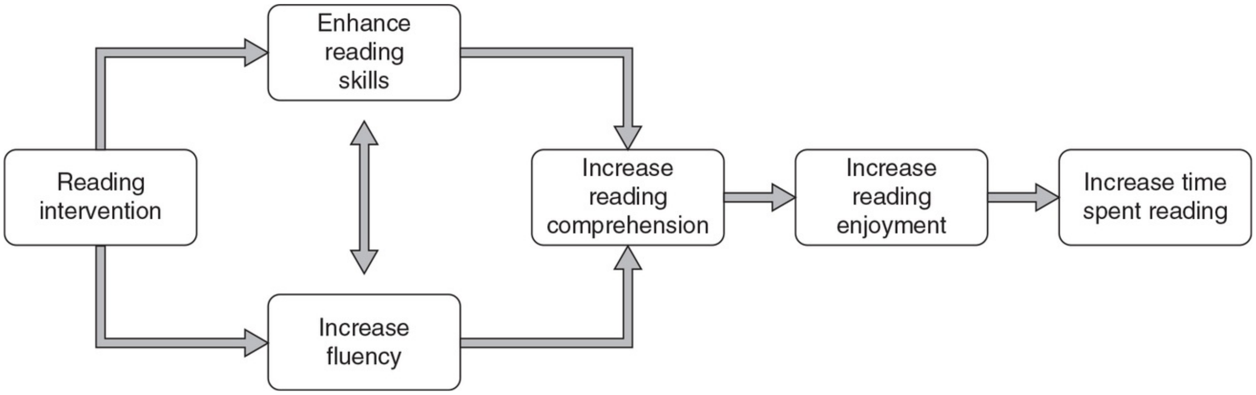
1318

Program Theory of Change

A program theory of change (PToC), also referred to as an action theory, causal pathway, intervening mechanisms theory, logic model, program theory, or theory of change, is a set of statements that describe the process and the mechanisms (i.e., the how and why) through which a program is thought to work and the outcomes it intends to affect. Program theories of change are built upon beliefs and assumptions developed through such means as personal experience, social science theories, or previous programs. Typically depicted through a diagram or model, a PToC explains the causal process through which change is expected to occur as a result of a program's intervention. Generally, a PToC uses a set of "if, then" statements to represent the mechanisms of change and their intended results. The "if" statements should indicate what the program intends to do; the "then" statements represent the results. Simply stated, these "if, then" statements provide the step-by-step causal process that is assumed to underlie the program.

As an illustration, a program aimed at increasing elementary students' time spent reading independently could create a series of "if, then" statements depicting the mechanisms by which it intends to reach its goals. One such statement might read, "If students participate in a reading intervention program, then their reading skills and fluency will increase; as a result, their reading comprehension and enjoyment of reading will increase, ultimately leading to more time spent reading independently." These statements can be articulated through the use of boxes and arrows that depict the causal pathways, as shown in Figure 1.

Figure 1 Example of a reading intervention's program theory of change


Enhance reading skills
Reading intervention
Increase reading comprehension
Increase reading enjoyment
Increase time spent reading
Increase fluency

The PToC and Logic Models

A PToC is most readily understood in contrast to a logic model. Even though the terms *PToC* and *logic model* are often used interchangeably, scholars in the evaluation community have begun to address the difference between the two. One of the main concerns with the conflation of these two terms is that they serve different purposes and consist of different, although often overlapping, information.

A primary distinguishing difference is that a program’s logic model focuses on the actual operation of the program and allows the evaluator to examine the implementation of the program’s components. In contrast, the PToC provides the theoretical foundations of change processes and can be used to aid the evaluator in examining the mechanisms that cause change or in answering the questions of why and how a program works.

Most notably, a PToC represents the conceptual causal model of the pathways between the program’s outcomes and activities; it also provides explanations for these hypothesized relationships, which may include indicators of change. In contrast, logic models are typically used to identify specific program components, consisting of linear descriptions of the program’s actual (as opposed to the conceptual) resources (i.e., inputs), activities, outputs, outcomes, and sometimes impacts. This depiction is often used to identify the way in which these program components fit together.

Types of PToC Models

There is no standardized model for what a PToC should include; as a result, there are no established unifying criteria for the components in a PToC. Instead there are no established unifying criteria for the components in a PToC. Instead, there are various ways in which a PToC can be depicted. Two prominent approaches to creating a PToC are distinguished by the level of detail included in the model.

The first approach is a streamlined method that places emphasis on articulating the change pathway through a series of cause-and-effect statements and may include indicators of change at each level. This approach includes the program activities (i.e., the *how*), the mechanisms by which program activities are understood to lead to the intended outcomes (i.e., the *why*), and the intended outcomes. Generally, this approach begins by identifying the outcomes and working backward through the “if, then” statements to identify the prerequisites of the causal changes (i.e., the mechanism that will cause the stated change). This modeling approach is also often used in conjunction with a logic model.

The other prominent approach to creating a PToC takes into account the complex and dynamic system within which the program operates. This approach views the program as part of a holistic process that involves both contextual factors as well as the assumptions of how the program or intervention is believed to bring about the desired change. In this method, the PToC is often separated into two connecting parts: (1) the causal pathway (i.e., the “if, then” statements) and (2) contextual process factors (e.g., the program’s target population, mechanism, and resources for delivery of program services as well as the activities or intervention). The relationship between each part of the PToC is then connected, often illustrated through the use of directional arrows. Similarly, the individual components of the contextual process are connected by single-or double-headed arrows, indicating a unidirectional or bidirectional relationship.

Developing a PToC Model

Sources of Information

Creating a PToC model is an iterative process, often involving several rounds of revisions. There are many sources of information that can be used alone or in combination to develop a PToC, one of the most common being the program’s stakeholders. Including stakeholders early and often in the iterative development process is highly encouraged—in fact, a PToC is often referred to as a *stakeholder’s theory* due to the emphasis on stakeholder input. The inclusion of various stakeholders’ perspectives in the formation of the PToC is also considered a means of gaining stakeholder engagement in the evaluation process.

Another prominent resource for the development of a PToC is social science research, which can help to articulate the underlying theory as well as support the causal statements. This plausibility check of the underlying programmatic assumptions is also one of the most prominent benefits of using a PToC. Other potential sources of information include program observations as well as program documents such as grant applications or program designs. The choice of sources depends on the overall evaluation approach and the developmental stage of the program (i.e., whether it is a new program or a mature one).

Approaches

Although there are various approaches to developing the PToC model, it is often beneficial to begin with the outcomes or impact the program intends to have and proceed in reverse to the activities. One method that can be used, alone or in combination with other approaches, is to ask a series of questions that provide information regarding underlying assumptions and program activities. Examples of questions an evaluator might ask during this process include the following:

- What is the program trying to achieve?
- What are the changes that should be expected to occur as a result of the program?
- How will these changes take place?

One prominent systematic approach proposed by Stewart Donaldson involves a six-step process that begins with engaging relevant program stakeholders to build evaluation buy-in and provide a more in-depth and holistic understanding of the program. The next steps involve developing a preliminary draft of the PToC model and providing it to stakeholders for review and feedback. The fourth step, once stakeholders reach a consensus about the depiction of the PToC, consists of conducting a plausibility check, intended to provide an examination of the soundness of each of the causal links in the model. The next step is to probe the hypothesized links in the PToC for greater specificity, such as the amount of intervention (e.g., 2 hours of reading every night) necessary to effect change. The sixth and final step is to create a completed model and provide it to the stakeholders for approval.

The Role of PToC in Evaluation

A PToC can aid program developers, implementers (e.g., program staff and administrators), funders, and evaluators in articulating the underlying processes that enable a program to achieve its goals. It is especially useful when developing a program, framing an evaluation, and monitoring progress. One particular role for the PToC is providing an outline of the program, around which an evaluator can then structure an assessment. The questions that are raised during such an assessment can then guide the design of the evaluation and help to identify the various stakeholder groups that should be involved in answering them. Such questions can address various aspects of the program, including those that examine formative, summative, or process aspects. For instance, with the reading program depicted in Figure 1, an evaluator could use the PToC to generate questions that probe the effects of the intervention on reading skills and fluency or questions that examine the overall program process.

A PToC may also enable the evaluator to identify any gaps in the causal process (i.e., conduct a plausibility check) before a program is implemented or an evaluation takes place. In the reading program example, perhaps there is an additional step needed between the objective of increased reading comprehension and the assumed result of increased reading enjoyment. Flawed program theory is a frequent cause of program failure—namely, because the outcomes or impacts that a program intends to achieve may be unarticulated or based on dubious assumptions about how change will occur to achieve the program's intended results. Returning to the original example, the evaluator could check the underlying assumption that enhancing reading skills and fluency leads to increased reading comprehension against current educational research. Examining these assumptions prior to program or evaluation implementation could save time and resources, both for the program and the evaluator.

Even though the aforementioned uses are prominent, the PToC is not limited to them; it can also be used to

- gain consensus among various stakeholder groups as to how the program intends to achieve its goals and the mechanisms that are responsible for achieving these goals;

- help the evaluator determine whether the program is ready to be evaluated (i.e., an evaluability assessment);
- improve stakeholder buy-in and encourage engagement in the evaluation process;
- identify performance dimensions most critical to program success; and
- aid in program planning and design or redesign, if necessary.

PToC models that are most useful and accessible to a wide stakeholder audience are those that are streamlined and clear. Models that need to depict a more complicated program theory can include subcomponents of the larger program or can be illustrated in an online interactive format. Providing these more complicated PToC depictions online or in an interactive form allows stakeholders to explore each component and connection while still retaining the model’s parsimonious form.

Natalie D. Jones and Benjamin D. Rosenberg

See also Data Visualization Methods; Formative Evaluation; Goals and Objectives; Logic Models; Outcomes; Process Evaluation; Program Evaluation; Stakeholders; Summative Evaluation

Further Readings

Azzam, T., Evergreen, S., Germuth, A. A., & Kistler, S. J. (2013). Data visualization and evaluation. In T. Azzam & S. Evergreen (Eds.), Data visualization, Part 1 (New Directions for Evaluation No. 139, pp. 7–32). Hoboken, NJ: Wiley.

Chen, H. T. (2014). Practical program evaluation: Theory-driven evaluation and the integrated evaluation perspective. Thousand Oaks, CA: SAGE.

Coryn, C. L., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. American Journal of Evaluation, 32(2), 199–226. doi:10.1177/1098214010389321

Donaldson, S. I. (2007). Program theory-driven evaluation science: Strategies and applications. New York, NY: Routledge.

Raikes, H. H., Roggman, L. A., Peterson, C. A., Brooks-Gunn, J., Chazan-Cohen, R., Zhang, X., & Schiffman, R. F. (2014). Theories of change and outcomes in home-based Early Head Start programs. Early Childhood Research Quarterly, 29(4), 574–585. doi:10.1016/j.ecresq.2014.05.003

Vogel, I. (2012). Review of the use of “theory of change” in international development (Report). London, UK: Department for International Development. Retrieved from http://www.theoryofchange.org/pdf/DFID_ToC_Review_VogelV7.pdf

Weiss, C. (1998). Evaluation (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Gregory J. Marchant Gregory J. Marchant Marchant, Gregory J.

Programme for International Student Assessment Programme for international student assessment

1318

1319

Programme for International Student Assessment

The Programme for International Student Assessment (PISA) is an achievement test given in more than 70 countries and in certain cities and special administrative regions in China such as Hong Kong, Macau, and Shanghai. The assessment is given every 3 years to 15-year-old students; it was given in 2015 and was scheduled to be given again in 2018. The sample of over 500,000 students is randomly selected and then weighted to be equivalent to about 28 million students who are representative of both the country and the students' school.

The test does not reflect any specific school curriculum, and it goes beyond simple achievement to determine the students' ability to apply content to authentic situations. The test covers reading, mathematics, and science with an emphasis on one of the areas for each testing year (the key domain for the 2015 assessment was science). The content is reviewed by panels of experts from each participating country.

The test would be considered a low-stakes or no-stakes test in that there are no negative consequences for students, teachers, or schools based on the quality of performance. However, there is an element of competition among countries, and relative performance does drive policy for some countries.

The test is available on computer and as a paper-and-pencil test. Each student spends 2 hours taking the test with a mixture of multiple-choice and open-ended items. In addition to the test, a survey is given to every participant, their parents, teacher, and their school's principal. The surveys include items regarding the student’s home background and their approaches to learning, instructional approaches from the teacher, and school characteristics.

Not every student’s test contains the same items. In fact, there are 66 different forms of the test. The total testing time for all of the items would be about 6½ hours. Because each student is representative of other students and each test contains a subset of items from the entire test, the students’ scores are estimates rather than specific scores. The weightings and estimates yield five plausible values. Special software is required to run the plausible values and all of the weighing that essentially runs the analysis 5 times and averages the results. In practice, many researchers either average the plausible values or simply use the first one as the students’ scores.

Finland’s strong performance on PISA has generated interest among journalists and researchers in its educational practices. Most of the other top-scoring countries and economies are in Asia.

PISA has its critics. Despite the assessment’s low-stakes consequences for students, teachers, and schools, the resulting rankings of countries impact national educational policies. The negative consequences of concern include an increase in other standardized testing, a narrowing of curriculum to focus on PISA-based content, differential student familiarity with test format and technology, and a lack of cultural sensitivity.

Gregory J. Marchant

See also Achievement Tests; Organization for Economic Co-operation and Development; Progress in International Reading Literacy Study; Rankings; Trends in International Mathematics and Science Study

Further Readings

Andrews, P., Atkinson, L., Ball, S., Barber, M., Beckett, L., Berardi, J., & Zhao, Y. (2016, May 6). OECD and PISA tests are damaging education worldwide—Academics. The Guardian. Retrieved from http://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics

Organization for Economic Co-operation and Development. (2014). PISA 2012 results in focus: What 15-year-olds know and what they can do with what they know. Paris, France: PISA, OECD. Retrieved from https://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Organization for Economic Co-operation and Development. (2016). PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy. Paris, France: PISA, OECD. Retrieved from http://dx.doi.org/10.1787/9789264255425-en

Hong Jiao Hong Jiao Jiao, Hong

Chen Li Chen Li Li, Chen

Progress in International Reading Literacy Study Progress in international reading literacy study

1319

1320

Progress in International Reading Literacy Study

The Progress in International Reading Literacy Study (PIRLS) assesses students' reading achievement after their fourth year of elementary schooling and was approved by the General Assembly of the International Association for the Evaluation of Educational Achievement (IEA). It assesses students' reading for literary experience and reading for information. Further, it integrates four types of comprehension processes: retrieving explicitly stated information, making inferences, interpreting and integrating ideas and information, and evaluating and critiquing content. The results of the study tell how well students can read and report on elements of literacy involving the home, student, classroom, school, community, and national curriculum policy. These results can be used to inform policy makers and researchers in every country about how to improve reading achievement and literacy.

PIRLS started in 2001 as a continuation of IEA's 1991 Reading Literacy Study. It is conducted every 5 years with administrations in year 2001, 2006, 2011, and 2016. It is a complement assessment to IEA's assessment of mathematics and science at the fourth grade of elementary schooling. Over 60 countries participated in the 2016 PIRLS. The countries that participated in the previous years' studies have access to data to evaluate progress in reading achievement across multiple years.

PIRLS is a collaborative effort among multiple agencies. The Trends in International Mathematics and Science Study and PIRLS International Study

Center at Boston College directs PIRLS in collaboration with the IEA Secretariat in Amsterdam and IEA’s Data Processing and Research Center in Hamburg. Sampling is monitored and implemented by Statistics Canada. Item development is supported by the National Foundation for Educational Research in England and the Australian Council for Educational Research, while Educational Testing Service in the United States conducts psychometric analysis.

The framework and the types of data provided by PIRLS evolved over the years. PIRLS 2011 included student, teacher, and school questionnaires; the Learning to Read Survey completed by students’ parents or caregivers; and the PIRLS Encyclopedia describing the reading curriculum and instruction by each participating country. PIRLS 2006 reported results on comprehension processes as well as literacy and informational reading purposes with the PIRLS Curriculum Questionnaire. The same fourth-grade students participated in PIRLS 2011 and Trends in International Mathematics and Science Study 2011 assessments, and data are available on their achievement in reading, mathematics, and science.

The PIRLS 2016 suite included two new assessments of reading comprehension: PIRLS Literacy and ePIRLS in addition to PIRLS 2016. PIRLS Literacy targets at the lower end of the ability scale with shorter passages and more straightforward questions, which are linked to PIRLS 2016, so that comparisons could be made between the two assessments. Countries can participate in either or both of these two assessments with the goal of providing policy makers with more relevant information about how to improve teaching and learning. ePIRLS is a computer-based online assessment of reading. ePIRLS simulates authentic school-like assignments on topics related to science and social studies. PIRLS data, framework, released items, and scoring guides are often used to guide selection of curriculum and textbooks for improving classroom instruction and research in reading and literacy.

Hong Jiao and Chen Li

See also Programme for International Student Assessment; Trends in International Mathematics and Science Study

Further Readings

Martin, M. O., & Mullis, I. V. (2013). TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—

implications for early learning. Herengracht, the Netherlands International Association for the Evaluation of Educational Achievement.

Topping, K. (2006). PISA/PIRLS data on reading achievement: Transfer into international policy and practice. The Reading Teacher, 59(6), 588–590.

Valtin, R., Roller, C., & Else, J. (2015). How international assessments contribute to literacy policy. In Handbook of research on teaching literacy through the communicative and visual arts (Vol. 2, pp. 73–77). New York, NY: Routledge.

Theodore J. Christ Theodore J. Christ Christ, Theodore J.

Stephanie Snidarich Stephanie Snidarich Snidarich, Stephanie

Jordan Thayer Jordan Thayer Thayer, Jordan

Progress Monitoring

Progress monitoring

1320

1322

Progress Monitoring

Progress monitoring is the assessment of performance across time. Repeated assessments of progress provide information about changes, which is often used to inform formative evaluations. Progress monitoring is used by educators, psychologists, and others to evaluate responses to academic or behavioral programs. This entry first discusses the types of measurement used in progress monitoring and the assessments designed for this purpose. It then looks at how progress monitoring is implemented in schools.

In a multitiered system of support, progress monitoring is essential to evaluate whether a program is effective for individuals or groups, so that effective programs can be maintained and ineffective programs can be modified or terminated. Individuals who consistently fail to respond with demonstrated progress are provided increasingly intensive levels of support. In that way, progress monitoring data are used to make decisions as to whether program modification is warranted. Ideally, progress monitoring measures should be designed so they are efficient to administer and produce scores that are reliable, valid, and sensitive to intervention effects over brief periods of time. These characteristics ensure that the process is minimally intrusive, yet helpful to inform and evaluate instruction.

Types of Measurement

Progress monitoring typically is done through one of the two types of assessment: a general outcome measurement and a specific outcome measurement, sometimes referred to as specific skill measures. General outcome measurements are designed to assess performance over extended periods of time, such as a school year. They are useful for monitoring progress toward annual goals. Specific outcome measurements are designed to assess performance within a brief period of time, such as an instructional unit that is only a sample of the annual curriculum. Approximately 1–10 weeks is considered brief. Specific outcome measurements are useful for monitoring progress toward short-term goals. The two methods are often paired to monitor annual goals with general outcome measurements and multiple short-term objectives with specific skill measures.

Examples of Measures

Measures and scores should have documented evidence for use as progress monitoring tools. The scores should be reliable and valid, which are important psychometric considerations. Progress monitoring assessments should also match the target skill(s) or disposition(s). Assessments have been designed and validated to monitor student progress in reading, writing, math, and behavior. In the area of reading, curriculum-based measurement for reading probes are a common way to measure oral reading fluency in the form of rate and accuracy. A student reads a grade-level passage aloud for 1 minute, while the administrator records the number of words read correctly and errors. Curriculum-based measurement for oral reading fluency scores are moderately predictive of general reading ability such as national norms measures and state accountability tests.

In the area of writing, curriculum-based measurement writing probes such as the number of words written and number of words spelled correctly are common ways to measure writing skill. A student is given 1–5 minutes to respond to a writing prompt, and an administrator subsequently scores the student’s writing sample. Words written and words spelled correctly scores are moderately correlated with measures of general writing ability such as the Stanford Achievement test or the Test of Written Language.

In the area of math, curriculum-based measurement uses one of the two approaches. The *curriculum sampling* approach draws material directly from the curriculum or intervention program of the student(s) assessed. The *robust indicators* approach uses measures that are correlated with a measure of general math ability, but the tasks alone are not representative of a particular mathematics domain. Probe tasks vary, as does the amount of time allotted for completion of the tasks (i.e., 1–8 minutes). Math curriculum-based measurement scores are moderately correlated with measures of general math ability.

In the area of behavior, direct behavior rating is a method for assessing student behavior that combines systematic direct observation with behavior rating scales. Specifically, observers (e.g., teachers) use a rating scale to rate student behavior in situ. According to Sandra Chafouleas and colleagues, direct behavior rating scores are moderately correlated with systematic direct observation.

Implementation in Schools

Proper implementation of progress monitoring involves the collection of data that quantify a student’s performance and rate of improvement and can be used to assess the appropriateness of an intervention to a student’s need(s). When and how often to monitor progress is still a matter of debate, but general practice guides suggest that high-risk students should be monitored at least weekly and prior to intervention sessions to capture how well the student’s skills are retained between sessions. Monitoring every 2 weeks is the least possible for students at moderate risk for difficulties, so weekly monitoring is recommended if possible. Adhering to these recommendations ensures a minimum of data are available to guide intervention selection and modification, including the removal of interventions.

Once a reliable and valid measure is selected, and an assessment schedule determined that produces at least a minimum of data needed, educators must then decide how to present the data for data-based decision making. Graphs are common methods for data presentation because they concisely summarize multiple sources of relevant data. Line graphs are the most common graphs used for progress-monitoring purposes. In general, they present baseline data that capture the observed level of a target behavior prior to intervention, vertical phase change lines that indicate instances where procedures were adjusted (e.g., implementation of an intervention), breaks in data collection, goal lines representing desired changes in a target behavior, and trend lines representing a target behavior’s projected performance level. In addition, line graphs can include useful statistical supports for interpretation, including mean scores and estimates of measurement error.

Progress monitoring is a simple and useful process; however, it relies on educators to engage in a cyclical process involving consistent collection of valid data, appropriate data-based instructional decisions, and revision of student progress toward goals. With proper implementation, progress monitoring improves educators' ability to positively affect a student's educational outcome.

Theodore J. Christ, Stephanie Snidarich, and Jordan Thayer

See also Benchmark; Curriculum-Based Assessment; Curriculum-Based Measurement; Evaluation; Formative Assessment; Screening Tests; Summative Assessment

Further Readings

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C., & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. Journal of School Psychology, 48(3), 219–246.

Deno, S. L. (1980). Relationships among simple measures of written expression and performance on standardized achievement tests.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. The Journal of Special Education, 41(2), 121–139.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. Exceptional Children, 57(6), 488–500.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. Scientific Studies of Reading, 5(3), 239–256.

Hixson, M., Christ, T. J., & Bradley-Johnson, S. (2008). Best practices in the analysis of progress monitoring data and decision making. Best Practices in School Psychology-VI (pp. 2133–2146).

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. The Journal of Special Education, 41(2), 68–84.

Silberglitt, B., Burns, M. K., Madyun, N. I. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. Psychology in the Schools, 43(5), 527–535.

Mira B. Kaufman Mira B. Kaufman Kaufman, Mira B.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Projective Tests

Projective tests

1322

1325

Projective Tests

A projective test is a type of personality assessment that examines an individual’s responses to ambiguous stimuli. Associated with psychodynamic and psychoanalytic theories, projective tests are believed to reveal a person’s unconscious thoughts or emotions as related to the test stimuli; these responses are in turn thought to be connected to the individual’s personality and psychological makeup. Projective tests are analyzed for meaning based entirely on the open-ended responses given by individuals, as opposed to other types of psychological tests, including self-report assessments, whose response options are shaped by and compared to a more limited and universal standard of meaning. Most commonly administered in clinical settings, projective tests are also used in schools to determine levels of behavioral and socioemotional functioning in student populations. This entry discusses the incorporation of projective tests into psychodynamic and psychoanalytic theory and treatment, the common types of projective tests, and the controversial implications of this psychological assessment in both clinical and educational settings.

History of Projective Testing

Projective tests have their origins in psychodynamic and psychoanalytic psychology, two overlapping psychological fields that emphasize the significance of the unconscious as a motivator of personality, emotion, and behavior. Sigmund Freud, the father of psychoanalytic theory, developed and modified many ideas relating to the unconscious and its influence over psychological states. One such idea was that defense mechanisms, or unconscious coping strategies, are employed to combat anxiety that is brought on by uncomfortable or harmful stimuli. Freud defined *projection* as a defense mechanism that involves the misattribution of an individual’s own undesired thoughts or emotions onto another person or object. The projective hypothesis, adapted from this defense mechanism, states that when individuals attempt to understand an ambiguous stimulus, they will assign meaning to the stimulus that is consistent with individuals’ *own* unconscious thoughts, attitudes, or needs. This projection of the unconscious serves as the basis for the design of all projective tests.

Projective tests were first used as clinical assessments of personality in the early 20th century. Francis Galton’s initial work with psychometrics, or the science of psychological measurement, and his consequent research in personality assessment motivated clinicians and researchers in the psychodynamics field to utilize projective measures as tools for evaluating underlying attitudes. Consistent with the projective hypothesis, these measures were designed to provide individuals with an ambiguous stimulus acting as a blank slate on which the individuals could attribute their own unconscious processes. Projective tests were quickly integrated into psychodynamic and psychoanalytic treatment procedures, and clinicians used the data gathered from these measures as a basis for diagnosing and treating people’s varying psychological needs and conflicts.

Since their induction into the realm of clinical assessment, projective tests have been used in a variety of contexts that extend beyond the typical evaluation of a patient’s personality. In schools, for example, projective tests are used to assess socioemotional and behavioral functioning in students who may have already exhibited difficulties in either of these domains (e.g., the inability to maintain positive social relationships with peers, abnormal levels of anxiety, or frustration related to schoolwork). Despite growing controversy over the validity and reliability of projective tests, they remain a common form of assessment in clinical, educational, and research settings.

Common Types of Projective Tests

All types of projective tests use ambiguous stimuli (e.g., images, words, or scenes) to expose unconscious mental processes, most commonly in clinical contexts. However, certain types of projective tests, including assessments involving picture drawing, sentence completion, and storytelling tasks, have been integrated into educational settings to assess students’ levels of functioning in school.

Rorschach Inkblot Test

The Rorschach inkblot test was one of the first projective tests to be developed, and it remains one of the best known. Originally designed by Hermann Rorschach to diagnose schizophrenia, the Rorschach test evolved primarily into a clinical assessment of personality. The test contains 10 cards, each with an ambiguous and symmetrical inkblot. Individuals are shown each card and asked to describe what they see. Responses are analyzed for a variety of features, including content (what the individual sees in the inkblot), location (which part of the inkblot the individual sees the object in), and determinants (which characteristics of the inkblot inform the individual’s response, such as form, color, movement, or shading), as well as the originality of the response and additional reactions to the images (e.g., gestures, tone of voice, or time taken to respond). Location, determinants, and additional reactions are often considered even more valuable than content when analyzing the individual’s response, as people are thought to exercise the most conscious control over the content of their responses during the assessment; thus, content by itself is less likely to accurately indicate an individual’s unconscious attitudes or desires. Although arguably the most recognized projective measure, the Rorschach inkblot test is used less frequently in schools because of the amount of time and extensive training necessary for the administration and interpretation of this test.

Thematic Apperception Test (TAT)

In the 1930s, Henry Murray and Christiana Morgan created the TAT. TAT consists of 32 cards that depict different scenes, some with and some without human characters. A participant is shown each card and asked to provide a detailed explanation for the scene the participant observes. This procedure is adapted depending on the individual. For example, as the scenes represent a variety of themes (e.g., sexuality, aggression, success, and failure), some clinicians recommend using between 8 and 12 cards that portray scenes most closely related to the participant’s specific situation. Other clinicians follow the original advice by Murray and Morgan to use a set of 20 cards with characters that most closely resemble the participant. The participant’s responses, including the events that occurred before the scene took place, the events occurring in the scene, the outcome of the scene, and the thoughts and feelings of the characters depicted in the scene, are evaluated to better identify unconscious issues, needs, or motivations that the participant might be preoccupied with. The TAT and similar storytelling tests are administered in a variety of contexts, including clinical, educational, forensic, and vocational settings. Clinicians in schools may substitute the children’s apperception test for the TAT when assessing students; this test depicts either animals or humans in common social situations, including those involving the family, and is used to pinpoint the child’s specific needs and emotions.

Draw-a-Person Test

The draw-a-person test is used to assess personality traits, cognitive and emotional development, intelligence, and psychopathology in children and adolescents. A person is asked to draw pictures of a man, a woman, and himself or herself. These drawings are evaluated using 64 scoring items, including the presence, absence, level of detail, and proportion of various parts of the body; certain details, inclusions, or exclusions signify certain developmental or pathological tendencies in the participant. This projective test faces criticism both as a measure of intelligence and psychopathology, as critics argue that test administrators may attribute low levels of intelligence or high levels of psychopathology to people who are merely poor artists. However, this test has proven valuable as a nonverbal assessment of cognitive development that is able to sidestep barriers such as primary language and language-based learning disabilities.

Sentence Completion Test

Sentence completion tests are assessments of an individual’s unconscious attitudes as demonstrated by individual completion of a given sentence. A participant is provided with the beginning of a sentence, known as a *stem*, and asked to finish the sentence in any way the participant wishes. These responses are thought to indicate the participant’s motivations, conflicts, or emotions. Sentence completion tests have been criticized for potentially eliciting both conscious and unconscious responses, and thus debate exists over whether these assessments can be categorized exclusively as projective tests. Nevertheless, they are used in a range of settings, including schools, as assessments of personality, unconscious mental processes, intelligence, language comprehension, and cognitive development.

Word Association Test

Developed by Carl Jung, the word association test uncovers patterns between objects or concepts and unconscious emotions. This projective test provides a participant with a word and then requires the participant to name the first word the participant thinks of in response as quickly as possible. In addition to the word response itself, response time and involuntary reactions to the test are analyzed. Word association is used to identify psychological complexes or the thoughts and memories pertaining to a specific theme that evoke strong emotional responses in people. For example, participants might provide a strange response to a given word that they cannot explain. This response might therefore expose the presence of a psychological complex relating to that word because the participant unconsciously experienced a great amount of emotion upon hearing it that interfered with the participant’s response. In this way, the participant’s unconscious attitudes are revealed.

Controversial Implications of Projective Testing

Despite their prevalence in clinical and therapeutic settings, including educational contexts, projective tests are considered highly controversial as psychological tools. Most notably, critics of projective methods cite the general lack of standardized scoring systems as an obstacle in obtaining statistical validity (success in measuring what the test is designed to measure) and reliability (consistency of the results of the test). Projective measures are most commonly evaluated using only the practitioner’s interpretation of an individual’s responses. Although some projective tests include scoring systems, such as Murray and Morgan’s complex guide to scoring the TAT, many practitioners rely on their own subjective evaluations and opinions to arrive at a clinical diagnosis. As a result, it is difficult to compare assessments by multiple practitioners who often administer and evaluate projective tests in markedly different ways and without consulting standardized scoring criteria, if available. This absence of universality in assessment leads to data with collectively lower levels of reliability and validity and thus with less consistent supporting evidence for the value of projective testing in clinical evaluation.

However, many scientists and clinicians argue that projective testing provides insights into the human mind that cannot be obtained using more standardized assessments. Projective tests are designed specifically to elicit unconscious responses from people, a task that is not easily accomplished using a more structured and less open-ended measure. From a clinical perspective, projective measures are not dissimilar to therapeutic interviews between a clinician and a patient, in which solely qualitative data are gathered and later analyzed. Projective tests also offer benefits that are exclusive to their design. For example, the ambiguous nature of these tests leads to an increased likelihood that participants answer honestly and without editing, as opposed to more straightforward assessments, such as self-report questionnaires, that participants might attempt to answer in a more socially desirable manner. Many projective tests also manage to avoid language-related barriers, including verbal delays, while still gathering sufficient information for clinical assessment; this feature of projective testing is particularly valuable in educational settings, where students may exhibit a wide range of verbal abilities. Finally, some versions of projective tests have been shown to demonstrate statistical validity in some domains of assessment, particularly with regard to personality traits (e.g., implicit motivation, or an emphasis on needs that are largely driven by the unconscious). Projective tests should be used by practitioners as assessment tools, rather than diagnostic tests, and supplemented with both structured quantitative measures and evidence-based clinical judgment to obtain a more accurate clinical evaluation.

Mira B. Kaufman and Marc H. Bornstein

See also Educational Psychology; Psychometrics; Testing, History of; Tests

Further Readings

Butcher, J. N. (2010). Personality assessment from the nineteenth to the early twenty-first century: Past achievements and contemporary challenges. The Annual Review of Clinical Psychology, 6, 1–20. doi:10.1146/annurev.clinpsy.121208.131420

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). Psychological testing and assessment: An introduction to tests and measurement (3rd ed.). Mountain View, CA: Mayfield.

Freud, S. (1921). A general introduction to psychoanalysis. New York, NY: Boni and Liveright.

Galton, F. (1879). Psychometric experiments. Brain, 2, 149–162.

Jung, C. G. (1953). In H. Read (Ed.), Two essays on analytical psychology (Vol. 7, R. F. C. Hull, Trans.). London, UK: Routledge & Kegan Paul.

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: The thematic apperception test. Archives of Neurology and Psychology, 34, 289–306. doi:10.1001/archneurpsyc.1935.02250200049005

Rorschach, H. (1922). Psychodiagnostik. The Journal of Nervous and Mental Disease, 56, 306.

S. Jeanne Horst S. Jeanne Horst Horst, S. Jeanne

Heather D. Harris Heather D. Harris Harris, Heather D.

Propensity Scores

Propensity scores

1325

1328

Propensity Scores

A propensity score is the probability that an individual received a particular treatment, given a set of researcher-identified variables related to self-selected treatment participation. The use of propensity scores is grounded in the logic of the counterfactual: To know the true effect of a treatment on the outcome, a researcher must also know what the outcome would have been had participants not received treatment. However, this is not something researchers will ever know. Accounting for the propensity for treatment, given a set of variables related to self-selection into that treatment, can increase precision in estimating the effect of the treatment on the outcome.

In the context of educational and social sciences research, *treatment* typically refers to a course, intervention, or program. Propensity scores are used in observational or quasi-experimental studies, in which researchers are unable to randomly assign participants to conditions, making accurate evidence-based causal claims challenging. Instead, students typically self-select or choose to participate in educational programs. Propensity scores provide a means of accounting for variables related to self-selection bias, allowing the researcher to create matched, or balanced, treatment and comparison groups. Propensity scores can also be used to adjust, or assign weight to, outcome analyses to account for self-selection bias, thereby mimicking a randomized controlled study.

Example

As an illustration, perhaps education researchers are interested in implementing a new afterschool program and then comparing the academic performance of students in the program with the performance of students who did not attend it. Because students are not randomly assigned to the new afterschool program (i.e., the treatment), the researchers are conducting an observational or quasi-experimental research study. In order to make claims about the impact of the new afterschool program, educational researchers need to take into account variables related to the students' self-selection into the program. Such variables may include the incoming interest, motivation, or prior ability of the students, and they may be related to the outcome of interest to the program (e.g., a test score). Once self-selection variables are identified, the researcher computes the probability that a given student was enrolled in the afterschool program, taking into account each student's levels on those variables.

When the self-selection variables (i.e., the covariates) are properly identified, propensity scores can be used to control for bias related to self-selection into treatment. In the afterschool program example, propensity scores could be used to form a matched comparison group of students with similar levels of interest, motivation, and prior ability. The performance of the two matched groups can then be compared. Alternatively, researchers could compute propensity scores and use them as weights to examine differences in performance between the students who attended the program and those who did not. It is for situations such as this example that propensity scores were developed.

Calculation of Propensity Scores

There are numerous methods of estimating propensity scores; however, logistic regression is the most common. Although logistic regression is usually considered a statistical inference technique, in this case, it is used simply for the purpose of computing propensity scores. Variables related to self-selection into the program, the covariates, serve as predictors of treatment participation. From the logistic regression analysis, propensity scores are the probability that an individual participated in treatment, given the set of covariates. Individuals with the same propensity scores have similar distributions on the covariates, regardless of whether they were in the treatment group. Consequently, matching nonparticipants to participants with similar propensity scores provides a means of creating a comparison group that is balanced with the treatment group on the covariates.

Consider the afterschool program described earlier. Perhaps the researchers identified several demographic and aptitude characteristics that were related to reasons for self-selecting into the program. Data on the demographic and aptitude variables would serve as predictors of students' program participation. Regardless of whether they enrolled in the afterschool program, each student would have a propensity score representing the probability of participation in the program. Once each student had a propensity score, a matched comparison group could be formed by selecting comparison students with the closest propensity to the students in the program.

Logistic regression is not the only method that can be used to compute propensity scores. Other methods of calculating propensity scores include Mahalanobis distance, discriminant analysis, and generalized boosted models. As is the case with propensity scores calculated via logistic regression, propensity scores calculated with these methods are used to balance the treatment and comparison groups on the set of covariates.

Use of Propensity Scores

There are numerous ways to create matched groups using propensity scores. Once propensity scores are computed, matching algorithms can be used to create a matched comparison group. Algorithms commonly used to create a matched comparison group are nearest neighbor matching, nearest neighbor with a caliper adjustment, optimal matching, and genetic matching. The most common matching method is nearest neighbor, which uses a "greedy algorithm." The greedy algorithm typically used to create matches moves sequentially through treatment cases to find the closest possible out of all possible matches in the comparison group. Unless the researcher specifies otherwise, matched comparison group members early in the matching sequence are not rematched with treatment members late in the sequence, even if the treatment group member late in the sequence would be the closest match. Moreover, with nearest neighbor matching, a match will be made for each treatment member, regardless of how far from the participant the match is on the propensity score.

It is common for researchers to specify how close a comparison match must be to the treatment member in order to be considered an acceptable distance between propensity scores. This distance is often referred to as a *caliper*, and it is measured in standard deviation units on the logit of the propensity score. In other words, the researcher is creating matched treatment–comparison groups with propensity scores that are nearly identical and fall within a designated range. As the greedy algorithm proceeds through treatment group members, those without a comparison group member with a similar propensity score (i.e., within the caliper distance) are dropped from the final matched group. Consequently, it is common for sample sizes to decrease when using nearest neighbor with caliper distance. In the case of the afterschool program, there may be 30 students in the prematching afterschool program group. If there are no comparison students with propensity scores within the caliper distance of a given student in the afterschool program, that student will not be included in the final sample. Thus, the number of students from the original afterschool program sample retained in the final comparison would drop to 29.

Two other commonly used matching algorithms are optimal and genetic matching. Unlike the greedy algorithm used with nearest neighbor, optimal matching uses an algorithm that reevaluates the overall closeness of matches. Similarly, genetic matching algorithms search for the best weight for covariates to achieve optimum balance between groups.

Rather than using propensity scores to create matched groups, some researchers prefer to employ *subclassification*. In this method, the researcher computes propensity scores but then evenly divides treatment and control participants into five or six groupings, or subclasses, based on the treatment members’ individual propensity scores. The researcher then compares the treatment and control group outcomes within the subclasses. In this instance, the researcher is using data from all those in the treatment and control groups. In the afterschool program example, all students in the treatment and control groups would have a propensity score and would be grouped together with those who have similar scores. Comparisons between the afterschool students and comparison group students would then be made at each level.

Another use for propensity scores involves using the scores to create weights to compare the outcomes of the treatment and comparison groups. Analyses involving propensity scores can be viewed from two different perspectives: the average effect of treatment on the treated (ATT) or the average effect of treatment. In situations when an estimate of the treatment effect on participants is of interest, the ATT may be calculated. When the ATT is calculated, the entire population of interest (i.e., the treatment participants) has observed outcomes. Thus, the propensity scores serve as a balancing measure to create a qualitatively similar comparison group. For example, if only the performance of the students similar comparison group. For example, if only the performance of the students enrolled in the afterschool program is of interest, a researcher would calculate the ATT.

However, researchers may instead be interested in estimating how the effect of treatment would generalize to individuals who did not participate in the treatment. In this situation, the average effect of treatment is of interest to the researcher. Several techniques allow researchers to extrapolate the estimated effect of treatment to nonparticipants using propensity scores to calculate the average effect of treatment, including weighted regression models and inverse probability of treatment weighting.

Inverse probability of treatment weighting allows researchers to generalize the estimated effect of treatment across different levels of the propensity score. This technique allows researchers to approximate how the treatment might affect individuals with lower propensity scores—relative to the treatment participants—to estimate what the treatment effect may have been had they participated in the treatment. Because participants and nonparticipants often vary in their distribution of propensity scores, the treatment effect is essentially extrapolated across levels of the propensity score for nonparticipants. That is, the effect of the program or intervention is generalized to nonparticipants with qualitatively different propensity scores. For example, participants in the afterschool program might have propensity scores ranging from 0.7 to 0.8 and might have improved 10 points on a math exam. A researcher could use inverse probability of treatment weighting to see how the change of 10 points on the outcome (math exam) would generalize to other students in school with lower propensity scores than the students in the program.

Logic Underlying Propensity Scores

There are several assumptions that must be met in order to make valid interpretations of analyses involving propensity scores. One underlying assumption is the *ignorable treatment assignment* (or the *assumption of strong ignorability*). That is, we are assuming that the set of covariates identified by the researcher accounts for all bias associated with self-selection and that there are no unobserved confounding variables. Although a researcher never knows whether all key covariates have been measured, evaluating the extent to which treatment is ignorable is necessary to ensure that estimates of the treatment effect are unbiased. A second assumption is that there is adequate overlap in propensity scores between the treatment and comparison groups (i.e., adequate common support) to ensure that there are students with similar characteristics to which the treatment group can be compared. If these assumptions are met, then treatment assignment is described as strongly ignorable, which means that after accounting for the variables related to self-selection, the treatment and comparison groups vary only randomly from one another.

Finally, as is also the case with random assignment, the appropriate use of propensity scores also assumes the stable unit of treatment value. That is, the treatment should have no influence on the comparison pool participants. In the case of the afterschool program, if students in the program shared what they learned with students who did not participate in the program, it could potentially contaminate the findings when comparing the two groups of students. In order to minimize biased conclusions, researchers who use propensity scores should keep each of these assumptions clearly in mind.

S. Jeanne Horst and Heather D. Harris

See also Control Variables; Discriminant Function Analysis; Logistic Regression; Quasi-Experimental Designs; Random Assignment; Selection Bias

Further Readings

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424. doi:10.1080/00273171.2011.568786

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22(1), 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Guo, S., & Fraser, M. W. (2014). Propensity score analysis: Statistical methods and applications (2nd ed.). Thousand Oaks, CA: SAGE.

Harris, H., & Horst, S. J. (2016). A brief guide to decisions at each step of the propensity score matching process. Practical Assessment Research & Evaluation, 21(4), 1–10. Retrieved from http://pareonline.net/getvn.asp?

v=21&n=4

Pan, W., & Bai, H. (2015). Propensity score analysis: Fundamentals and developments. New York, NY: Guilford Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55. doi:10.1093/biomet/70.1.41

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1), 1–21. doi:10.1214/09-STS313

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inferences. In J. W. Osborne (Ed.), Best practices in quantitative methods (pp. 155–176). Thousand Oaks, CA: SAGE.

Ryan J. McGill Ryan J. McGill McGill, Ryan J.

Psychometrics

Psychometrics

Psychometrics

Psychometrics is a branch of scientific psychology that is concerned with the theory and technique of psychological and educational assessment and measurement and the construction of instruments that are developed to appraise psychological and educational constructs (e.g., multidimensional achievement batteries, intelligence tests, and behavior rating scales). As a result, psychometric techniques are routinely employed throughout the corpus of quantitative educational research.

Historical Development

Although the genesis of these methods has been debated extensively within the technical literature, Sir Francis Galton (1822–1911) has been called the “father of psychometrics” and is credited by many as the first to apply early versions of these techniques when attempting to measure individual differences during the Victorian era. As a result of these experiments, Galton developed the *correlation coefficient*, which later served as the focal point for Charles Spearman’s research on intelligence and the subsequent discovery of the general intelligence factor (*g*) in 1904, a finding that Arthur Jensen in 1998 argued was one of the greatest discoveries in the history of the social sciences.

Theoretical Approaches and Fundamental Concepts in Measurement

As a general principle, measurement consists of rules for assigning symbols to objects so as to (a) represent objects numerically (scaling) and (b) determine whether objects fall within a particular category (classification). Although the vast majority of psychometric research has traditionally been devoted to the task of scaling, classification research has intensified, as psychometric techniques have been applied more extensively to assess the quality of diagnostic and other decision-making models in a variety of educational contexts; these include but are not limited to the following: providing a diagnosis of a learning disorder, determining whether a student is at risk for educational failure, and evaluating whether a student has met a priori standards for educational attainment within a high-stakes accountability model.

The field of psychometrics is bifurcated by two divergent theoretical approaches to measurement: classical test theory and modern test theory. The classical test theory model posits that any observed score is produced from two hypothetical components expressed in the form

$$X = T + E,$$

where *X* represents the observed score, *T* reflects the hypothetical true score for that construct, and *E* denotes a random error term. This model provides the foundation for estimating the reliability of a measure. *Reliability* refers to the degree to which differences in the observed score are consistent with differences in true scores and are not the product of measurement error. Reliable measurement is also necessary, but it is not singularly sufficient for establishing the validity of a measure or the degree to which a test measures the construct of interest. *Validity* is a multidimensional concept that requires analyzing elements such as the internal structure of a test and relationships between test scores and external criteria. In 1995, Samuel Messick argued that validity is of more importance than reliability, as it provides the basis for how a measure should be interpreted in clinical practice.

Modern test theory, also known as item response theory, is a psychometric approach emphasizing that an individual’s response on a test item is influenced by individual standing on the construct being sampled as well as the degree of difficulty that item samples that particular construct. Item response theory techniques are commonly used in education and psychology to document test bias and to develop computerized adaptive tests.

Ryan J. McGill

See also Classical Test Theory; Confirmatory Factor Analysis; Exploratory Factor Analysis; Item Response Theory; Reliability; *Standards for Educational and Psychological Testing*; Validity

Further Readings

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. American Psychologist, 61(1), 27–41. doi:10.1037/0003-066X.61.1.27

Furr, R. M., & Bacharach, V. R. (2014). Psychometrics: An introduction. Thousand Oaks, CA: SAGE.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. doi:10.1037/0003-066X.50.9.741

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York, NY: McGraw-Hill.

Lindsay Till Hoyt Lindsay Till Hoyt Hoyt, Lindsay Till

Jeanne Brooks-Gunn Jeanne Brooks-Gunn Brooks-Gunn, Jeanne

Puberty

Puberty

1329

1330

Puberty

Puberty is a process that occurs between childhood and adulthood through a complex series of neuroendocrine changes, resulting in extensive internal and external physical changes and eventual reproductive competence. Puberty consists of two associated yet independent processes. *Adrenarche*, maturation of the adrenal gland and subsequent production of adrenal androgens (starting around ages 6–8 years in girls and 1 year later in boys), is primarily responsible for axillary and pubic hair during adolescence. *Gonadarche* begins a few years later with the secretion of estradiol and testosterone. In gonadarche, the primary sex organs develop (ovaries and testes) and the external signs of puberty begin, culminating in reproductive capability. Puberty also coincides with other, related biological processes, such as brain development and physical growth, which together have important implications for adolescent health and development. This entry briefly describes the literature around the process of puberty and summarizes common methods of pubertal measurement.

Pubertal Timing

The onset of puberty can vary by as many as 4–5 years among healthy individuals, with recognized sex and race/ethnicity differences. Pubertal timing reflects a strong genetic component, with other factors such as childhood weight, nutrition, stress, and epigenetic programming having additional effects. Early timing of puberty has often been shown to be associated with or be a risk factor for negative mental and physical health, including depression, anxiety, risky behaviors, and cardiometabolic outcomes. There are a number of proposed mechanisms, including direct influences of underlying sex hormones on brain development, that occur at different times during puberty and in turn may affect neuropsychological function and behavioral changes. Additionally, adults or peers who notice external pubertal changes often respond differently to the adolescent, which may indirectly affect outcomes via stress or low self-esteem. The tempo of puberty (progression to the established milestones after entry into puberty) may also have implications for adolescent development.

Puberty Measurement

Puberty can be studied directly through hormone concentrations or indirectly through pubertal staging, as the hormones influence the degree of physical development. Classical empirical work by James M. Tanner and colleagues in the middle of the 20th century established five graded categories of both adrenarche (pubic hair) and gonadarche (breast development for girls and testicular volume for boys), and Tanner staging (1 = prepubertal; 5 = full maturity) remains a primary system for measuring puberty. A physical examination is the gold standard, but other studies utilize drawings, photographs, or descriptions of the five stages to solicit parental or self-reports. The Pubertal Development Scale is currently the most widely used pubertal questionnaire. This measure includes some of the early pubertal changes (e.g., breast development) as well as developmental markers that become evident in mid-to late adolescence, such as facial hair and voice change for boys and menarche for girls. It is important to consider the outcome variable of interest when selecting a measure of puberty.

Lindsay Till Hoyt and Jeanne Brooks-Gunn

See also Adolescence; Anxiety; Demographics

Further Readings

Hayward, C. (Ed.). (2003). Gender differences at puberty. Cambridge, UK: Cambridge University Press.

Marshall, W. A., & Tanner, J. M. (1969). Variations in pattern of pubertal changes in girls. Archives of Disease in Childhood, 44(235), 291–303.

Marshall, W. A., & Tanner, J. M. (1970). Variations in the pattern of pubertal change in boys. Archives of Disease in Childhood, 45(239), 13–23.

Mendle, J., Harden, K. P., Brooks-Gunn, J., & Graber, J. A. (2010). Development's tortoise and hare: Pubertal timing, pubertal tempo, and depressive symptoms in boys and girls. Developmental Psychology, 46(5), 1341–1353. doi:10.1037/a0020205

Petersen, A. C., Crockett, L., Richards, M., & Boxer, A. (1988). A self-report of pubertal status: Reliability, validity, and initial norms. Journal of Youth and Adolescence, 17(2), 117–133. doi:10.1007/BF0153796

Shirtcliff, E. A., Dahl, R. E., & Pollak, S. D. (2009). Pubertal development: Correspondence between hormonal and physical development. Child Development, 80(2), 327–337. doi:10.1111/j.1467-8624.2009.01263.x

Meagan M. Patterson Meagan M. Patterson Patterson, Meagan M.

Punishment

Punishment

1330

1332

Punishment

Punishment is a consequence that decreases the frequency of a given behavior. Understanding of punishment is important for effective classroom management and promoting student motivation. This entry discusses types of punishment, the use of punishment in schools, research on the efficacy of punishment for changing behavior, and strategies for the effective use of punishment.

Researchers in the behaviorist tradition classify punishments as either positive or negative. In positive punishment, a stimulus is added; in negative punishment, a stimulus is removed. Spankings, lectures, or failing grades are examples of positive punishment because a stimulus is being added. Monetary fines and removal of privileges (e.g., no television for a week) are examples of negative punishment because a stimulus is being removed as a consequence of behavior.

Common punishment strategies used in schools are verbal reprimands, time-outs, removal of privileges (such as recess time), suspension, and expulsion. Some schools allow teachers or administrators to use physical punishment (spanking or paddling) with students.

Typically, punishments are intended to be unpleasant or aversive. However, in certain cases, a stimulus that is intended to be pleasant and reinforcing in fact acts as a punishment. For example, a teacher may publicly praise a student’s performance on an academic task intending to reinforce the students’ academic achievement, but the students may find this praise embarrassing and thus reduce their effort on academic tasks in order to avoid such embarrassment in the future.

Research on the Efficacy of Punishment

Research on the Efficacy of Punishment

Some researchers argue that punishment is largely ineffective at changing behavior and that behavior management strategies should instead focus on reinforcing desired behaviors. Other researchers argue that punishment can be an important behavior management strategy and that combining punishment with reasoning and reinforcement is more effective than using reasoning and reinforcement alone.

Research indicates that generally effective forms of punishment are verbal reprimands, restitution (in which the student must return the environment to the state it was in before the misbehavior; for example, cleaning up a mess one has made), time-outs, and response costs (such as loss of privileges). Ineffective forms of punishment are physical punishment, psychological punishment (making embarrassing or humiliating statements directed at the student), assigning extra classwork, withholding recess, and suspension from school.

Using Punishment Effectively

Research indicates a variety of factors and strategies for effective use of punishment. Using punishment as a component of classroom management works best when rules and consequences are clear and consistent and when the overall classroom environment is warm and supportive.

Clear Rules and Consequences

Rules are most effective when desired and undesired behaviors are described in clear, concrete language. Teachers often create rules that are too vague, such as “be respectful.” If respectful and disrespectful behaviors are clearly described, students will have an easier time understanding and complying with the rule. Punishment is also most effective when combined with reasoning about why rules exist (including discussion of the consequences of breaking rules for self and others).

Genuine Punishments

Many punishments are ineffective because they are not actually punishing for students. For example, if a student does not like school, being suspended is unlikely to be an effective punishment.

Logical Punishments

In selecting and enforcing punishments, logical consequences tend to be more effective than arbitrary consequences. Logical consequences are related to the specific misbehavior and are intended to help the students develop an understanding of the reasons for rules along with the desire and ability to regulate their own behavior. Logical consequences frequently include apologies, repairing damage done, or withdrawal of a relevant privilege. For example, if students draw on their desk with markers during art class, logical consequences could include cleaning the ink off of the desk and not being allowed to use the markers for the remainder of class.

Consistency in Delivering Punishment

Punishment is more likely to extinguish an undesired behavior if the punishment is delivered after every instance of misbehavior. In addition, consistency across students is important. If students are punished for a behavior that other students are allowed to “get away with,” the students are less likely to see the punishment as legitimate and may become resentful of the teacher, rather than changing their behavior.

Role of the Classroom Environment

A warm, supportive classroom environment allows for better classroom management, including the use of punishment. Punishment is more effective when a positive relationship with the learner has been established prior to the administration of punishment.

In addition, classroom environments should be structured so as to reduce the likelihood of misbehavior. Removing the temptation to misbehave can reduce undesired behavior without the need for punishment. For example, if two students frequently talk to each other during independent seatwork time, the teacher might opt to modify the classroom seating arrangement such that these students are not seated near each other.

Along with discouraging undesired behaviors, teachers can teach and reinforce appropriate behaviors. For example, in addition to punish talking out of turn, a teacher might reinforce students who raise their hands before speaking or who are spending time on task appropriately.

Meagan M. Patterson

See also Applied Behavior Analysis; Behaviorism; Reinforcement; School-Wide Positive Behavioral Support

Further Readings

Gershoff, E. T. (2013). Spanking and child development: We know enough now to stop hitting our children. Child Development Perspectives, 7, 133–137.

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. Review of Educational Research, 86(3), 643–680.

Larzelere, R. E., Gunnoe, M. L., Roberts, M. W., & Ferguson, C. J. (2016). Children and parents deserve better parental discipline research: Critiquing the evidence for exclusively "positive" parenting. Marriage & Family Review. Retrieved from http://dx.doi.org/10.1080/01494929.2016.1145613

Ormrod, J. E. (2012). Human learning (6th ed.). Upper Saddle River, NJ: Pearson.

Kyrsten M. Costlow Kyrsten M. Costlow Costlow, Kyrsten M.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Pygmalion Effect

Pygmalion effect

1332

1334

Pygmalion Effect

The Pygmalion effect refers to the phenomenon whereby having higher expectations of others leads to an increase in their performance. Pygmalion effect research often focuses on the relation between teacher expectations and student academic performance. This entry describes the Pygmalion effect's origins, possible mechanisms, and applications both in and outside of the classroom.

The counterpart to the Pygmalion effect, the Golem effect, occurs when lower expectations of others lead to a decrease in performance. Both the Pygmalion and Golem effects represent self-fulfilling prophecies or expectations that influence people's behaviors in ways that cause those expectations to be fulfilled. The Pygmalion effect is used to characterize leader–follower relationships, such as those found in the classroom and the workplace.

The Pygmalion effect is named after the sculptor Pygmalion in Greek mythology and in Ovid's narrative poem *Metamorphoses*, who falls in love with an ivory statue of his own creation. Based on the myth, in 1913, George Bernard Shaw penned a play he called *Pygmalion* about a professor who becomes infatuated with a low-class flower girl after training her to pass for a duchess. Shaw's play later inspired the musical and film *My Fair Lady*. The Pygmalion effect is also known as the Rosenthal effect due to its origins in a study of teacher expectations on students' academic performance conducted by psychologists Robert Rosenthal and Lenore Jacobson.

The Pygmalion Effect in the Classroom

Rosenthal and Jacobson’s 1968 study of the Pygmalion effect looked at the effects of teacher expectations on students’ academic performance. In this study, also known as the Oak School experiment, Rosenthal and Jacobson gave IQ tests to students in a California elementary school. They then told teachers they were administering the Harvard Test of Inflected Acquisition and provided teachers with the names of their students who had scored in the top 20%. The teachers were told that these students were expected to bloom academically that year when, in actuality, the names provided to the teachers were randomly selected. After taking the IQ test once more at the end of the school year, the “bloomers” confirmed their teachers’ expectations by showing increased IQ score than control students. Teachers also rated the experimental group of “bloomers” as more interesting, curious, appealing, and well-adjusted than the control group at the end of the year.

Although Rosenthal and Jacobson’s study met with criticism, the Pygmalion effect has been supported by numerous replication studies and meta-analyses and has been tested across a variety of contexts. Studies of undergraduate and graduate students have shown the Pygmalion effect to persist beyond elementary school to higher education. The influence of teacher expectations on student academic performance is stronger for younger students, however, because they have yet to establish fixed conceptions of their academic abilities.

The Pygmalion effect can extend from individual teachers to entire academic departments as well when dominant teacher attitudes and expectations spread within these larger contexts. The effect can also work in a reverse direction, where teachers’ performance is influenced by the expectations of their students. For example, studies have shown that teachers perform better when provided with positive student nonverbal behavior, such as student attentiveness, compared to negative nonverbal behavior.

Beyond the Classroom

The Pygmalion effect can also be applied to leader–follower environments outside of the classroom, such as the workplace. Managers and supervisors can use the Pygmalion effect to boost productivity by raising their expectations of subordinates to high but achievable levels. If managers set unrealistic expectations for their subordinates, however, this can actually result in a expectations for their subordinates, however, this can actually result in a productivity decline. The Pygmalion effect is also most effective in boosting productivity in younger subordinates, particularly those in their first year at a given workplace, who have yet to create a self-image based on their career reputations. High managerial expectations and subordinate performance in the first year have been found to correlate with later subordinate success.

The Four-Factor Theory

The Pygmalion effect has been studied across various contexts, but the mechanisms behind the effect are less widely explored. Rosenthal proposes a four-factor theory, identifying four possible mediators of the Pygmalion effect: climate, input, response opportunity, and feedback. The first factor, climate, supposes that teachers create a warmer environment for students when they expect more of them. The second factor, input, posits that teachers teach more material to students if the teachers have higher expectations of the students.

The third factor, response opportunity, suggests that students who are expected to perform better get more opportunities to respond in the classroom. The final factor, feedback, proposes that students receive more praise when correct and more constructive feedback when incorrect if teachers expect more of them. Students who are not expected to perform well in the classroom receive less feedback because teachers do not expect them to be able to understand and utilize corrections as effectively.

Applications

The Pygmalion effect can be applied in both the classroom and the workplace to increase performance and productivity. In the classroom, initiatives such as the Common Core State Standards have been developed to provide teachers with realistic, age-determined expectations for their students' performance. If teachers expect all of their students to achieve these standards, the Pygmalion effect predicts an increase in student performance. The Pygmalion effect can be applied similarly to the workplace, where managers can maintain positive expectations to encourage subordinate success. The expectations of teachers and managers therefore have powerful effects that can be used to increase performance in the classroom and the workplace.

Kyrsten M. Costlow and Marc H. Bornstein

See also Common Core State Standards; Giftedness; Scientific Method; Standards-Based Assessment; State Standards

Further Readings

Eden, D. (1990). Pygmalion in management: Productivity as a self-fulfilling prophecy. Lexington, MA: Lexington Book.

Feldman, R. S., & Prohaska, T. (1979). The student as Pygmalion: Effect of student expectation on the teacher. Journal of Educational Psychology, 71(4), 485.

Livingston, J. S. (2009). Pygmalion in management. Boston, MA: Harvard Business Review Press.

Rosenthal, R. (2002). The Pygmalion effect and its mediating mechanisms. In J. Aronson (Ed.), Improving academic achievement (pp. 25–36). San Diego, CA: Academic Press.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. New York, NY: Holt, Rinehart & Winston.

Q

Joseph A. Maxwell Joseph A. Maxwell Maxwell, Joseph A.

Qualitative Data Analysis Qualitative data analysis

1335

1339

Qualitative Data Analysis

Data analysis in qualitative research is quite different from that in quantitative research due not only to differences in the data themselves but also to substantial differences in the goals, assumptions, research questions, and data collection methods of the two styles of research. Because qualitative approaches and methods are an important part of educational research, both researchers and practitioners need to understand these differences, the strengths and limitations of the two approaches, and how they can be productively integrated. Data analysis may be the least understood aspect of qualitative research, partly because the term *qualitative analysis* has several different meanings. This entry reviews the aspects of qualitative research that are most important for data analysis, describes the history of its development, and surveys the current diversity of approaches to analysis in qualitative research.

Data Analysis

The phrase *qualitative analysis* in the physical sciences, and in some quantitative research in the social sciences, refers to categorical rather than numerical analysis. For example, qualitative analysis in chemistry simply determines what elements are present in a solution, while quantitative analysis also measures the amount of each element. Some quantitative researchers have assumed that this distinction also applies to the social sciences—that qualitative analysis deals with data that are simply categorized, rather than measured numerically, and that the basic principles of quantitative research can be applied to both. This represents a profound misunderstanding of qualitative research and analysis, which rests on quite different premises from quantitative research, and uses distinct strategies for analyzing data.

These strategies are grounded in the primarily inductive, rather than hypothesis testing, nature of qualitative research. This, in turn, is shaped by the nature of qualitative data. Such data are primarily descriptions of what people did or said in particular contexts—either observations of actual settings and events or transcripts of interviews. Instead of converting these descriptions to variables and measuring or correlating these, as quantitative researchers do, qualitative researchers retain the data in their original, descriptive form and analyze these in ways that, at least to a greater extent than in quantitative research, retain their narrative, contextualized character. Qualitative research reports tend to contain many verbatim quotes and descriptions, and the analysis process is to a substantial extent devoted to *selecting* these as well as to aggregating, comparing, and summarizing them. The use of numbers, to make more precise statements of how often something happened or how many participants reported a particular experience or event, is legitimate and common in qualitative research, but such uses are supplementary to the primary descriptive and interpretive goals of analysis.

Because of the inductive character of qualitative research, and its particularistic focus, data analysis is not a “stage” that occurs in a sequential order with theorizing, research design, data collection, and writing up results. Data analysis should begin as soon as any data are collected and should be continued as long as any significant questions remain about the meaning and implications of the data. Although the relative emphasis on the different aspects of the research process varies over time, they are not chronologically separated components of a linear series.

History

Although the term *qualitative research* is a more recent development, its actual practice has a long history, extending back at least to the 19th-century work of anthropologists and the study of social problems by Charles Booth, Jane Addams, and others and to later community studies such as *Middletown* and *Yankee City*. Despite this, the analysis of qualitative data has, until relatively recently, received little theoretical attention. This is in striking contrast to quantitative research, which has a well-developed theory, statistics, which informs quantitative analysis.

The first widely recognized, named, and systemically developed method for analyzing qualitative data was analytic induction (AI). It was created by the sociologist Florian Znaniecki in the 1930s, during his research with W. I. Thomas for their classic work *The Polish Peasant in Europe and America*, and was further developed by Alfred Lindesmith in his research on opiate addiction in the 1940s. In contrast to quantitative research, which typically collects and analyzes data in order to test previously developed theories, AI proceeds inductively to generate categories, concepts, and theories from the data. These inductively developed theories specify the necessary preconditions for a type of case (e.g., of people who embezzle money from a firm to deal with unexpected personal financial problems); the theory is tested by seeking negative instances, and revising the theory, or limiting its scope, until no negative cases are found.

The goal of AI was to develop *explanatory* theories about the phenomena studied. This was done by iteratively examining cases to see whether the theorized conditions were present; any case that lacked one of these preconditions required revision of the theory. However, the view that *any* exception to the preconditions necessitated revision of the theory is now seen by most researchers as too stringent. However, the inductive development of categories for sorting and classifying (coding) data has been a feature of most subsequent strategies for qualitative analysis.

Approaches

The most influential and widely used strategy for qualitative analysis, grounded theory, was presented by Barney Glaser and Anselm Strauss in their 1967 book *The Discovery of Grounded Theory*. Their work was in part a response to the growing prestige of quantitative research in sociology and some other social sciences, as sophisticated statistical analyses of survey data became dominant in academic influence and funding. The book challenged the growing separation of theory development from research, in which broad abstract theories, often generated without reference to actual data, were then tested by researchers, using quantitative data to establish correlations between variables derived from the theories. As with AI, grounded theory emphasized the inductive development of theory but established a more systematic and flexible way of doing this. The phrase *grounded theory* was intended to emphasize the generation of theory that was “grounded” in, and developed in interaction with, the collection and analysis of data.

A key concept for grounded theory was the constant comparative method, a strategy that Glaser and Strauss distinguished from both the quantification of data in order to test existing theory and the simple examination of data to generate theory. Constant comparison integrates the coding of data with the development of theory and hypothesis generation in an iterative process. This strategy, a radical departure from standard practice (at least as theorized) when it was first presented, is now a fairly typical part of most qualitative research.

A second innovation that Glaser and Strauss introduced was the use of memos (written reflections on methods, data, or other aspects of the research) as an explicit data analysis strategy. Although memos were used informally in earlier research, Glaser and Strauss recognized these as a distinct strategy for qualitative analysis. *The Discovery of Grounded Theory* treated memos very briefly, in only a few paragraphs, but Strauss's later work (*Qualitative Analysis for Social Scientists*, 1987; Strauss and Corbin, *Basics of Qualitative Research*, 1990), as well as that of Matthew Miles and A. Michael Huberman, provided a much more extensive discussion of the uses of memos for data analysis and theory development.

In his later work, Strauss also developed additional strategies for analysis, including what he called axial and selective coding. The terminology he used for these is potentially confusing, because neither involves "coding" in the usual sense of creating categories and sorting data by category; Strauss used "coding" to mean broadly "the process of analyzing data." In axial coding, the researcher connects a categorized phenomenon to the conditions that gave rise to it, its context, the strategies by which it is handled, and the consequences of these; selective coding involves relating a category to the core categories of the emerging theory. These are both ways of *connecting* a category to other categories; such strategies are discussed in more detail later in this entry.

There are now at least three different versions of grounded theory in use: Glaser's development of traditional grounded theory, Strauss's and Juliet Corbin's subsequent elaboration of this approach (which Glaser rejected), and constructivist grounded theory, as developed by Kathy Charmaz. The latter combines the grounded theory approach with social constructivism, the epistemological position that people construct the realities in which their lives are embedded. The latter view, a reaction to the positivism that has dominated quantitative research, has become widespread (though by no means universal) in qualitative research. It emphasizes research relationships, participants' subjectivity, and the social context of the research.

Another major contribution to the development of qualitative analysis was Miles and Huberman’s *Qualitative Data Analysis: A Sourcebook of New Strategies* (1984). This work, although it covered most traditional forms of analysis, emphasized what they called *displays*—visual ways of presenting and analyzing data. Most of these strategies were qualitative adaptations of two forms of data analysis and presentation that had been used in quantitative research: matrices (tables) and networks (concept maps or flowcharts). In contrast to quantitative displays such as numerical tables or structural equation models, Miles and Huberman presented numerous examples of genuinely qualitative displays. Matrices are formed by crossing lists of categories (including individuals, groups, or times) to create cells; but rather than numbers, the cells contain qualitative data, either verbatim quotes or field note excerpts, or summaries of these. Networks, on the other hand, can display relationships among categories (similar to what are called concept maps) or the sequence of actual events. Networks can be used to display both sequences of specific events or properties of a particular group or institution (what they called an event-state network) and hypothesized relationships (usually causal) among categories. Both types of displays can be used either within particular cases or in cross-case analysis.

Charles Ragin’s qualitative comparative analysis, a method originally developed in political science and sociology but more recently used in other fields as well, is a way of analyzing a collection of cases (traditionally done using qualitative case study methods) in a more systematically comparative way to identify cross-case patterns. It is actually a combination of qualitative and quantitative strategies for identifying different combinations of causal conditions (variables) that can generate the same outcome. It is most useful when the number of cases is larger than qualitative researchers can easily handle but too small for rigorous statistical analysis. Ragin’s 2014 presentation of this approach dropped the term *qualitative*, titling the book *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*.

All these approaches are based on some form of coding and categorization of data. However, there are other ways of doing qualitative data analysis that draw more from the humanities than the social sciences. The most widespread of these is narrative analysis, but this is really a loose collection of rather different approaches to analyzing narrative forms of data. Some of these approaches involve coding and thematic analysis and are thus similar to the types discussed previously. Others focus on the structure of narratives, using strategies drawn from literature or linguistics. However, all of these tend to be more holistic in their approach than are approaches that primarily involve coding, which intrinsically segment or “fracture” the data and re-sort the segments into categories; they focus more on identifying connections within the data and retaining these connections in the analysis.

The more holistic types of narrative research result in rather different forms of presentation of the results of the analysis, and the creation of these forms of presentation may largely constitute the analysis. For example, Irving Seidman, in his book *Interviewing as Qualitative Research*, described two types of presentation of life history interviews, which he called vignettes and profiles. These are created by rearranging and condensing the interview transcripts, to generate a clearer flow to the narrative, while retaining the interviewee’s own words. Similarly, what Frederick Erickson called ethnographic microanalysis of interaction involves taking observations (usually videotaped and transcribed) of some event, analytically decomposing these, and then *reconnecting* them to create a holistic portrayal of social interaction. This sort of analytic segmentation and rearrangement of data is common in qualitative case studies, as well as in much narrative research, but has rarely been discussed as a type of analysis.

Other researchers have used poetry as a way to communicate the meaning of interviews, but the analytic strategies that are involved in this are rarely explicit. An exception, which Carolyn Mears called the gateway approach for analyzing and displaying interview material, is presented in her book *Interviewing for Education and Social Science Research*. Drawing on humanity-based practices, including oral history interviewing, poetic forms of transcription and display, and Elliot Eisner’s educational connoisseurship, Mears created poetic renditions of her interviews, retaining the interviewee’s own language, but editing and rearranging this to better convey the experience and emotion that may be obscured or missed in a verbatim transcription.

It is also possible to combine categorizing and connecting strategies in analysis—not simply by connecting the results of a prior categorizing analysis, as Strauss did with axial and selective coding, but by integrating connecting strategies from the beginning of the analysis. An example is the listening guide approach to analysis, developed by Carol Gilligan and her associates, for analyzing interviews. This approach, which they describe as a voice-centered relational method, involves a series of “listenings” that attempt to identify the “plot” (the story that is being told), the stance of the speaker (identifying the “I” statements and creating a separate document from these, an “I poem”), and different “voices” in the interview; in Gilligan’s original use of this approach different "voices" in the interview, in Gilligan's original use of this approach, which contrasted men's and women's views on moral judgment, these were the voices of justice and of caring, and of a separate and a connected self. However, the particular voices identified depend on the goals of the research and may be inductively developed during the study. Such an approach interweaves categorizing and connecting steps rather than keeping these separate.

The analysis of qualitative data has been substantially transformed by the development of computer-assisted qualitative data analysis software that facilitates many of the processes involved in analyzing data; many different programs are available, with different strengths and limitations. However, unlike quantitative analysis programs, which actually carry out the chosen statistical procedures, qualitative software simply automates some of the actions involved in the analysis; every decision about what categories to use in coding the data, and selecting which segments to code, must still be made by the researcher, although the software can then display the results. In addition, such software is most useful for categorizing analysis; this can lead the researcher to employ this strategy even when the research purposes would be best served by a connecting approach. It is important to keep in mind, though, that the development of qualitative analysis software is progressing rapidly and that any attempt to characterize the field may quickly become out of date.

Joseph A. Maxwell

See also Case Study Method; Mixed Methods Research; Narrative Research; Qualitative Research Methods; Single-Case Research

Further Readings

Bazeley, P. (2013). Qualitative data analysis: Practical strategies. London, UK: SAGE.

Charmaz, K. (2014). Constructing grounded theory (2nd ed.). Thousand Oaks, CA: SAGE.

Coffey, A., & Atkinson, P. (1996). Making sense of qualitative data. Thousand Oaks, CA: SAGE.

Gilligan, C., Spencer, R., Weinberg, M. C., & Bertsch, T. (2003). On the listening guide: A voice-centered relational method. In P. M. Camic, J. E. Rhodes, & L. Yardley (Eds.), Qualitative research in psychology: Expanding perspectives in methodology and design (pp. 157–172). Washington, DC: American Psychological Association.

Glaser, B., & Strauss, A. (1967). The discovery of grounded theory: Strategies for qualitative research. New Brunswick, NJ: Aldine Transaction.

Lieblich, A., Tuval-Mashiach, R., & Zilber, T. (1998). Narrative research: Reading, analysis, and interpretation. Thousand Oaks, CA: SAGE.

Maxwell, J., & Miller, B. (2008). Categorizing and connecting strategies in qualitative data analysis. In S. Hesse-Biber & P. Leavy (Eds.), Handbook of emergent methods (pp. 461–477). New York, NY: Guilford Press.

Miles, M. B., Huberman, A. M., & Saldana, J. (2013). Qualitative data analysis: A methods sourcebook (3rd ed.). Thousand Oaks, CA: SAGE.

Ragin, C. C. (2014). The comparative method: Moving beyond qualitative and quantitative strategies. Berkeley: University of California Press.

Strauss, A. (1987). Qualitative analysis for social scientists. Cambridge, UK: Cambridge University Press.

Strauss, A., & Corbin, J. (2014). Basics of qualitative research: Techniques and procedures for developing grounded theory (4th ed.). Thousand Oaks, CA: SAGE.

Linda S. Behar-Horenstein Linda S. Behar-Horenstein Behar-Horenstein, Linda S.

Qualitative Research Methods Qualitative research methods

1339

1345

Qualitative Research Methods

Qualitative research methods, in the broadest sense, refer to the approaches of conducting qualitative inquiry; the methods used to analyze data; and the conventions used to ensure the rigor of data collection, analysis, and researcher transparency. Qualitative methods can bring to the forefront knowledge about phenomena that are not understood or that have not been explored. For example, little is known about the impact of illnesses or injuries (i.e., intimate patient experiences; daily living adjustments; or the social, financial, personal, or emotional tolls) beyond mechanistic effects of therapies other than laboratory measures of titers and other body fluids, pharmaceutical or manipulative impacts via physical therapy or occupational therapy that are provided by quantitative outcomes. Qualitative research is also a unique role to play in better understanding the context of classroom environments and how the quality of teaching influences student outcomes. By giving credence to student or educator experiences, or exploring the dynamic of classroom culture, qualitative findings can inform practice, education, and outcomes. The use of collaborative methods and processes that are inherent to qualitative research can empower individuals at local levels to make changes in the classrooms and school systems.

Widely accepted in educational research is that educators are expected to base practice on evidence. Sources of evidence that can be helpful include meta-analytical studies and randomized controlled trials. However, there is a dearth of evidence that portrays (a) what teaching looks like as it is being delivered, (b) the type of learning activities that students are asked to complete and to what degree those activities stress lower or higher cognitive thinking skills, and (c) how assessment practices are used (e.g., Is assessment used simply to assess student performance or is it used to provide students insight into ways that they can improve their outcomes?)

can improve their outcomes?).

Four types of qualitative research are discussed in this entry: case studies, ethnography, narrative, and grounded theory.

Case Study Method

The case study method is used to provide in-depth understanding of a phenomenon as it occurs in real time to portray the complexities, interactions, and occurrences that explain how, what, and why. Grounded in a theoretical perspective of interpretivism, case study may use any or all of the following methods: participant observation, nonparticipant observation, questionnaires, interviews or focus groups, and document or content analysis. Advantages of the case study approach lie in its ability to investigate complex social units consisting of multiple variables. Disadvantages related to the case study approach include the potential for considerable expenditure of time and resources. Also, there are no agreed upon guidelines for constructing the final report. A reported lack of reliability, validity, and generalizability can threaten the believability of the findings. Finally, the case study can be quite labor intensive.

Ethnography

Ethnography focuses on describing, analyzing, and interpreting a culture-sharing group’s collective patterns of behavior, beliefs, and language. Ethnography is grounded in the epistemology of constructionism, a belief that meanings are constructed through engagement with the world. Meaning is not created, but rather constructed. Moreover, social realities are constructed and sustained by observing social rules. Social realities are a function of shared meaning, or the collective generation of meaning, not the meaning making of the individual mind.

Ethnography is grounded in the theoretical perspective of interpretivism. Approaches to ethnography are grounded in the binary perspectives of phenomenology or hermeneutics. Phenomenology concentrates on illuminating the detailed description of conscious experience, without recourse to explanation, metaphysical assumptions, and traditional philosophical questions. Hermeneutics refers to different schools of thought on its meaning.

The oldest of the qualitative approaches, ethnography evolved from anthropology, and there is no singularly agreed method for undertaking ethnography. The processes that researchers adopt typically reflect their training and belief systems. As a qualitative method, it provides a lens for (a) seeing the whole, rather than a glimpse of group behavior, and (b) discerning practices among individuals with common learning challenges. Sources for ethnography include field observations (participant or nonparticipant); interviews; researcher journal notes; and artifacts of the group, documents, and photographs.

There are two types of ethnographies: classical and critical. Classical ethnography refers to spending extensive time in the field observing behavior and explicating why and under what circumstances they occur. Critical ethnography is often used to study culture with critical lenses that identify internal and external power relationships that influence how a group behaves. Characterized by having political intent, its aim is to empower a cultural group by raising individual and group awareness. Findings from critical ethnography may offer heightened awareness to provide a group with evidence to effect change. Often motivated politically, critical ethnographers use their credentials to assume power and authority to become spokespersons for the oppressed. Cultural immersion differentiates this approach to ethnography from other forms of qualitative inquiry. Spending time in a cultural setting is used to listen, observe, ask questions, document and collect data, and acquire insight into the cultural milieu and day-to-day relationships. Observing as an outsider, the *etic* perspective, leads to a single dimension of understanding. Adding depth and legitimacy is accomplished through the *emic*, or insider perspective, by asking questions and seeking clarity. Ways to legitimize findings include supplementing observations with focused interviews and maintaining a research journal of observations, while documenting personal reflections, interpretations, and questions. Cultural informants may be used to capture social meanings and nuances of interactions that occur in ordinary member activities. Researchers who conduct ethnography must question their own interpretation and be deliberate in describing their positionality.

The focus of the ethnography means that researchers ask different questions during inquiry and data collection. For example, when ethnographers are interested in what group members do, they ask what beliefs and practices inform how the group constructs their world. If the ethnographer is interested in what members make and do, the focus of inquiry is on what it is like for each person in this context. If the ethnographers are interested in cultural speech or what people say, they ask how individuals shape their life within the context and what people say, they ask how individuals shape their life within the context and what environmental factors influence coping and adaptation.

Effective engagement in ethnographic research requires that the researcher remember that the focus is on developing a complex and complete description of the culture-sharing behavior. From an analytical perspective, the researcher must look for patterns, including rituals, customary social behaviors, and regularities. The researcher must also describe beliefs and attitudes as expressed through language or material items and behavior that is observed through the group members’ actions, as the researcher observes it. Moreover, the researcher is tasked with looking for patterns of social organization within social networks and ideational systems expressed within worldviews or by ideas. Researchers must provide a complete description of how data were managed, a rationale for data that were analyzed, and justification for how conclusions were reached. Relying on the researcher’s interpretation and cultural immersion approach are considered limitations.

Criteria for “good” ethnography include providing a detailed description of the cultural group. Using interviews, observations, and other eliciting procedures, an authentic view of a cultural group’s reality is rendered. The ethnographic interviewer must not predetermine participant responses by the kinds of questions asked. Themes derive from a collectivist (researcher–participant) understanding of the cultural group. An overall explanation of how the culture-sharing group works is explicated. Self-disclosure and reflexivity by the researchers about their position relative to the research is also described.

In educational settings, such as K–12 schools, ethnography permits insight into social actions as well as how and why groups of people construct their worldview and make life choices. It provides a window for viewing and making sense of life worlds including the rules, beliefs, or ideas that inform understanding and a group’s behaviors that they make be enacted, for example, as a school faculty and staff grapple with ways to raise the performance level of students who are not reading at grade level. How do the faculty and staff define the issue? What evidence do they use to analyze individual student and group performance within the classroom and across the school? How do they use this information to make changes to instruction or assessment practices?

Narrative Approach

The narrative approach is grounded in the epistemology of constructionism, a

The narrative approach is grounded in the epistemology of constructionism, a belief that meanings are constructed through our engagement with the world. Narrative is aligned with the theoretical perspective of interpretivism. It is used when the researcher wants to describe the lives of individuals by collecting and retelling their stories using a structured approach. Using narrative, the researcher collects multiple stories, retells events chronologically while looking for causal links with attention to (a) context, environment, conditions, and time of year; (b) characters, archetypes, behaviors, personalities, and patterns; (c) actions and movement that illustrate characters' behaviors; (d) problems and questions that call for answers; and (e) resolution and answers to questions.

Narrative analysis is both a method and an approach. In educational research, the terms *story*, *narrative*, and *voice* are used interchangeably. Sources for narrative research include the storyteller, the story, and the listener. There are several defining features. First, researchers gather participants' accounts, hence stories, of their lived and told experiences. Stories are the spoken word, hence text of what is told to researchers or what is coconstructed between researcher and the participant. These stories/accounts may be individual experiences or events that shed light on individuals' identities. Primary data forms are interviews, although observations, pictures, or documents can also be data sources. Researchers may retell or restory data chronologically. The analysis can be thematic identification or structural. Participant stories often contain turning points, specific tensions, or interruptions. Because narratives occur in specific places or situations, context is an important element that must be fully described in the researcher's retelling of the story.

There are four types of narrative: biographical study, autoethnography, life history, and oral history/personal narrative. A biographical study is a recording and writing about another person's life events. Individuals who record and write their own personal stories develop autoethnography. These stories range broadly from social critique to larger cultural meanings or may illustrate the inherent tensions between an individual's present-day experiences and how others interpret their life experiences. A life history portrays an individual's entire life. Oral history/personal narratives refer to gathering personal reflections regarding events, cause, and effect from one or several individuals. These may have a specific contextual focus or be stories about organizations.

Alternative frameworks for narrative research include researcher's use of (a) storytelling, (b) story and life approach to narrative, (c) narrative practices/environments, and (d) the researcher and the story. In the use of storytelling, the researcher’s interest is in *how* people tell their stories in the construction of meaningful selves, identities, and realities. In the story and life approach to narrative, there is a focus on the relationship between individuals’ life stories and quality of their life experiences. The narrative practices/environment framework depicts the reflexive interplay between how personal narratives shape and are shaped by environments. In the researcher and the story framework, the researcher becomes part of the story. In this approach, the researcher seeks equitable position in the researcher–researched relationship includes the subjecting one another to the research analysis along with the participant and sometimes researcher’s own storied experiences in the research.

Procedures for doing narrative research include asking if the research problem or question best fits narrative research. Another technique commonly used is restorying the act of gathering and analyzing stories for time, place, plot, or scene. Stories told to researchers may not be expressed chronologically. In this instance, the researcher often provides a casual link among ideas. Stories can be deconstructed to expose dichotomies, silences, contradictions, or inconsistencies. Stories can also be coconstructed or performed by the participant to convey a specific agenda or message.

There are several challenges associated with doing narrative research. Narrative research requires collecting extensive information about participants to develop a clear understanding of context and individuals’ life. Concerted efforts must be undertaken to allow essential information to be revealed, differentiating it from inconsequential or consequential within multilayered context. The researcher also needs to actively collaborate with the participant and discuss the participant’s story to ensure accurate representation. This requires a patient and committed participant. The researchers are also expected to reflect upon their own personal or political background, which inherently shapes how the participant’s story is retold.

Issues pertaining to representation must also be addressed. These issues include the following: Who owns the story? Who can tell or change it? Whose version is convincing? What happens when narratives compete? For communities, how do the stories told impact them? How does the researcher distinguish between reliable and unreliable participant accounts? How do the researchers avoid making themselves simply the voice of participants or having their interpretation substitute for the participants’ story?

Two criteria are used to evaluate good narrative. First, the researcher tells a story that reports what was said (themes), how it was said (describes how the story unfolds), and explains how the speakers interact or perform the narrative verbally or nonverbally. Second, the researchers’ transparency is evident in how their voices are embedded in the story.

Narrative research highlights what can be learned from lived experiences, history, or society. In educational research, narrative offers vast opportunities to learn from educator or student groups how school-based practices impact teaching and student outcomes. One example of this is the reduction or elimination of recess to provide more instructional time for students to prepare for state accountability tests. Narrative is the systematic exploration of stories with a specific focus or set of foci that is not predetermined. The temporality of narrative indicates that it is not a precise history, nor is this intended.

Grounded Theory

Grounded theory is a systematic procedure used to generate a theory that broadly explains, at a conceptual level, a process, action, or interaction about a substantive topic. Situated in the epistemology of constructionism, a belief that meanings are constructed through our engagement with the world, grounded theory is aligned with the theoretical perspective of social constructionism. Considered as both a method and an approach to analyzing data, there are basically two approaches to grounded theory. A postpositivistic/scientific/objectivist inquiry approach is driven by determinism, objectivity, tireless cross checking, and faith in the generation of absolute knowledge, all of which is labeled, located, put in place, and explained objectively. The process is grounded by objectivity and deduction. In grounded theorists’ view, one cannot construct reality; it already exists, waiting for emergence, by being discovered, analyzed, and interpreted objectively. In the objectivist approach, the researcher is expected to be detached, unbiased, and value-free to ensure the proper use of the scientific method.

The constructivist approach calls for a relativistic view of epistemology and asserts that human phenomena are socially constructed. The researcher works actively and reflexively through multiple, complex, and contextual issues in selecting, collecting, and analyzing data. Thus, the analytical process is messy. The grounded theorist’s approach is to use flexible, open-ended strategies to conduct systematic, directed inquiry, while engaging in imaginative theorizing from empirical data. The process is iterative and inductive. Thus, often the theorist interweaves rich, full, and detailed data coding line by line for close study before sorting for meaning and connection to aid in the theory construction.

Describing the data and moving toward conceptualization is the work of the researchers as they move data into an analytic framework. The analytical process involves coding, refining codes, identifying examples to support the categories, analyzing within categories, looking for themes across categories, and locating quotations to support the grounded theory.

Grounded theory relies heavily on the use of interviews. Typically, 20–30 participants are interviewed individually until saturation is reached or when no new findings are revealed. Thus, this approach requires conducting interviews, reading and analyzing each one, and repeating this process multiple times until saturation is reached. Knowledge is considered partial, incomplete, subjective, and ever-changing according to time, place, and person. The researcher must “explain” the context of the research, cycling between the interpretation and data, and grounding the interpretation with examples. Also, the researcher must seek to ensure that the data closely represent participants’ voices and experiences. During the focused coding process, data are moved to better fitting codes or where they are more suitable within other themes or categories.

Grounded theory is defined by several specific activities. Its intent is to generate or discover a theory by studying participants who experience the same process. The resulting theory might help explain practice or provide a framework for further study. Often *memoing* is used. This is where the researcher writes down ideas, as data are collected and analyzed. These ideas might be to sketch out the process, articulate the researcher’s reactions, tentative interpretations, or describe reflections that are responses to data collected. Memos can become the basis for generating the grounded theory.

Data collection is grounded by the researcher’s constant comparison of data gleaned from participants with ideas about an emerging theory. The analysis can be structured or developmental/emergent. When grounded theory is guided by structured data analysis, it follows a prescribed pattern of development and uses open coding (taking data and segments and moving them into categories of information). Next, the researcher selects one category (a representative unit of information composed of events, happenings, and instances) to be the focus of theory. Following this step, additional categories (axial coding) are identified to theory. Following this step, additional categories (axial coding) are identified to form a theoretical model. The intersecting categories become the theory, referred to as a process of selective coding. The grounded theory is presented as a figure or diagram, as propositions (hypotheses), or as a discussion. The development of theory is guided by piecing together implicit meanings about a category or weaving together several categories to identify a central theme.

There are several challenges to conducting grounded theory. A critical step, as in all qualitative forms of inquiry, is that a researcher must set aside, as much as possible, any preexisting ideas so that the analytical, or substantive, theory emerges from the data. Determining saturation can be difficult. One strategy used to move toward saturation is discrimination sampling and gathering information from individuals different from those initially interviewed. This step is undertaken to ascertain if the theory holds true for additional participants. The primary outcome of grounded theory is a theory characterized by (a) specific components, (b) a central phenomenon, (c) casual conditions, (d) strategies, (e) context, and (f) consequences. The use of categories may limit flexibility during analysis of data.

Nine criteria are used to adjudicate good grounded theory: (1) Does the theory explain definitions of major categories? (2) Are links between theoretical links and categories strongly established? (3) Has the understanding of studied phenomena been advanced? (4) Do the implications of the analysis move theoretical edges? For its theoretical reach and breadth? For methods? For substantive knowledge? For action or intervention? (5) Is a theoretical model or figure provided? (6) Is there a story line or propositions that connect categories in the theoretical model, which also raise unanswered questions? (7) Do the researchers disclose their reflexivity, its potential impact on the processes and interpretation, and explain how this was mitigated? (8) Does the author provide justification for the type of grounded theory used? (9) Is there congruence between the stated grounded theory assumptions and the reported outcomes (results, findings, insights, implications, and recommendations)?

Ensuring Standards Rigor of Qualitative Methods

Any discussion of qualitative methods must address standards of rigor. Types of indicators are used as a means of estimating the quality of the research's findings and its interpretative value. *Trustworthiness* or the believability of the findings is used in place of reliability and validity. The need to establish trustworthiness is the overall degree to which the study’s findings can be authenticated by other researchers. In quantitative studies, its parallel is internal validity. Trustworthiness is achieved by showing some or all of the following: credibility, transferability, dependability, and confirmability. Credibility is used to ensure that reconstructions are accurate representations. Strategies to establish credibility include prolonged field experience; reflexivity, which can be developed using a field journal; triangulation; and member checking. Reflexivity relates to the researchers’ discussion of how their biases, values, and experiences with central phenomenon might shape interpretation and how the researchers averted that possibility. Triangulation is achieved through the use of multiple and different sources, methods, investigators, and theories to provide corroborating evidence that validates the accuracy of the findings. Respondent validation, a process used to establish the credibility of findings, is often achieved through member checking. Returning transcribed interviews or sharing the researcher’s emergent findings with participants to authenticate their accuracy are methods used to conduct member checking. Peer debriefing refers to the process of discussing observations or primary findings with a coresearcher or observation. This process is used to enhance credibility.

Transferability, similar to external validity in quantitative studies, refers to the extrapolation or applicability of findings to similar contexts and by describing to what degree, if any, the findings might be replicated in similar contexts. Strategies to establish transferability include describing a selected sample in depth or the use of dense description.

Dependability refers to the consistency within the findings. Its parallel in quantitative research is reliability. Peer debriefing, a process of discussing initial impressions of what was heard, observed, or understood from a review of interviews, focus groups, or review of participant writings such as journals or reflection papers are used to enhance dependability. Strategies that are used to establish dependability include a dependability audit (also referred to as an inquiry audit or audit trail), triangulation, or coding and recoding. The inquiry audit or audit trail is used to illustrate how the research moved from open codes to categories and then to themes during the analytical process. Confirmability refers to the believability of the study’s findings. Strategies to establish confirmability or the neutrality of the findings include triangulation and reflexivity. In quantitative studies, the parallel process is objectivity.

Principles Guiding Qualitative Research

There are several guiding principles for engaging in qualitative research. First, the methodologies need to be linked to theoretical assumptions, which underline the selection of data collection and analysis methods. The methods are dependent upon choice of methodology. Methods are what researchers do to collect and analyze data. The processes for engaging in qualitative research are how researchers go about doing them. The criteria for assessing good qualitative research are (a) substantive contribution, (b) aesthetic merit, (c) reflexivity, and (d) impact. Regarding substantive contribution, the researcher must ask: Does the study contribute to an understanding of social phenomena? Do the findings demonstrate a deeply grounded scientific perspective? Do the findings seem true? To assess the aesthetic merit of the study, the researcher asks: Does use of creative analytical practices open up the text and invite interpretive responses? Is text artistically shaped, satisfying, and complex—not boring? With respect to reflexivity, the following questions are asked: Is researcher/author positionality a product and producer of text? Is researcher/author positionality transparent? Do the researchers/authors hold themselves accountable to the standards of knowing and telling the people studied? To assess the impact of the study's findings, the following questions should be asked: What is the emotional/intellectual influence of the study? Does the study generate new questions? Motivate one to write? Try new research studies? or Move the reader to action?

Linda S. Behar-Horenstein

See also Case Study Method; Ethnography; Focus Groups; Grounded Theory; Interviews; Narrative Research

Further Readings

Creswell, J. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.). Boston, MA: Pearson.

Charmaz, K. (2014). Constructing grounded theory (2nd ed.). Thousand Oaks, CA: SAGE.

Corbin, J. M., & Strauss, A. L. (2015). Basics of qualitative research: Techniques and procedures for developing grounded theory (4th ed.). Thousand Oaks, CA: SAGE.

Glaser, B. G., & Strauss, A. L. (1999). The discovery of grounded theory: Strategies for qualitative research. New Brunswick, Canada: Adeline Transaction.

Merriam, S. (1998). Qualitative research and case study applications in education: Revised and expanded from case study research in education. San Francisco, CA: Wiley.

Strauss, A. L., & Corbin, J. (1990). Basics of qualitative research: Techniques and procedures for developing grounded theory. Thousand Oaks, CA: SAGE.

Strauss, A. L., & Corbin, J. (1998). Basics of qualitative research Techniques and procedures for developing grounded theory (2nd ed.). Thousand Oaks, CA: SAGE.

Yin, R. K. (2014). Case study research: Design and methods (5th ed.). Thousand Oaks, CA: SAGE.

Francisco L. Rivera-Batiz Francisco L. Rivera-Batiz Rivera-Batiz, Francisco L.

Quantitative Literacy

Quantitative literacy

1345

1348

Quantitative Literacy

Quantitative literacy or numeracy refers to the arithmetic knowledge and skills that are required by individuals to function effectively at work and in society. This entry discusses the concept and measurement of quantitative literacy, the results of surveys examining the quantitative literacy skills of adults in various countries, and the links between quantitative literacy and social and economic outcomes.

Historically, the concept of literacy included only basic skills connected to reading and writing. Over the years, however, it has expanded to include other skills. In 1978, the United Nations Educational, Scientific and Cultural Organization adopted its definition of literacy, still used today, which incorporates quantitative literacy or numeracy, stating:

> A person is functionally literate who can engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing and *calculation* for his own and the community’s development.

Similar definitions have been adopted by the Organization for Economic Cooperation and Development as well as international organizations and many national education agencies. This entry discusses the measurement of quantitative literacy, effects of literacy skills, and policy directions.

Measuring Quantitative Literacy

Measuring Quantitative Literacy

Quantitative literacy is focused on essential skills that involve computations, arithmetic operations, and mathematics concepts, either alone or sequentially, and that are required to function on a job or to carry out activities performed in daily living, such as figuring out a tip, calculating the overall cost of a list of products purchased, or computing taxes owed from a tax table. Although measurement of these skills can be determined by simply asking individuals whether they know how to add, subtract, multiply, and divide, the inaccuracy of self-identification and the more complex quantitative operations required by jobs and in the economy has led to the growing use of test-based measurement. Standardized tests of the quantitative skills of children in school are widely available through the efforts of national education agencies as well as international assessments of student achievement, such as the Programme for International Student Assessment and the Trends in International Mathematics and Science Study. But data for adults (16 years of age or older) are more scarce.

The earliest studies adopting a comprehensive test-based measure of quantitative literacy were developed by Educational Testing Service and the U.S. Department of Education and included the 1985 Young Adult Literacy Assessment Survey, the 1990 Workplace Literacy Survey, and the 1992 National Adult Literacy Survey. Currently, the Organization for Economic Cooperation and Development utilizes the same approach in its Programme for the International Assessment of Adult Competencies (PIAAC). The PIAAC has conducted quantitative literacy surveys for random samples of the population aged 16–65 years in over 40 countries, with results published for 32 countries so far. Test scores are standardized across countries and range from 0 to 500, gathered into five levels of increasing proficiency, with scores at or below Level 1 corresponding to a very rudimentary literacy, with the person able to carry out only operations with whole numbers, one at a time and in very concrete situations. Higher levels of proficiency progressively require two or more steps in calculations, more than one operation, use of decimals, and—at the top level—the use of abstract thinking, the ability to use data to construct graphs and statistical representations, and utilizing more complex mathematical problem-solving strategies.

Of the sample of countries included as part of the PIAAC, Japan scores the highest—with an average score of 288—followed by Finland (282), Belgium and the Netherlands (280), Denmark (280) and Sweden (279). The lowest scoring countries include Chile (206), Indonesia (210), and Turkey (219). The

United States has an average score of 253, below the Organization for Economic Cooperation and Development average of 263. In terms of the percentage of the population performing at or below the rudimentary skills Level 1, among the countries tested, those with the lowest percentages were Japan (8.1%), Finland (12.9%), and the Czech Republic (12.9%), while those with the highest proportion were Chile (53.4%), Turkey (50.2%), Italy (31.7%), and Spain (30.8%). The average for the United States was 27%. These figures suggest that even among high-income countries, a substantial proportion of the population does not appear to have more than the rudimentary quantitative literacy skills.

The Effects of Quantitative Literacy Skills

In a world of advancing use of digital technology and information in the workplace and at all the levels of society, it is not surprising that quantitative skills are rapidly becoming a significant determinant of success in the labor market. Indeed, as economists David Autor, David Card, Alan Krueger, and Richard Murnane, among others, have documented, the forces of demand and supply have operated since the early 2000s to sustain an upward trend in the employment and salaries of jobs that require more technical skills, of which quantitative literacy is at the top. As the demand for high school and college graduates with minimum levels of mathematics proficiency has steadily increased, those without such skills are being displaced from the labor market, especially young workers.

In the United States, for example, a mismatch has emerged between the numeracy skills demanded by employers and the supply of those skills by many workers entering the labor market. An early study by Francisco Rivera-Batiz at Columbia University showed that, holding other things constant, such as reading proficiency, higher quantitative literacy scores are positively and significantly associated with the likelihood of employment of young adult men and women in the labor market. More recently, results from the PIAAC surveys show that among unemployed high school dropouts in the United States, 59% performed at Level 1 or lower in the quantitative literacy scale, and even among those who had a college credential, 46% of the unemployed performed at Level 1 or below. Economists Ross Finnie and Ronald Meng have found similar results for Canada. More generally, for all countries in the PIAAC survey, persons scoring at a Level 4 or 5—compared to Level 1—in the quantitative literacy scale are 2.2 times more likely to be employed.

Higher quantitative literacy skills are also associated with higher wages. Using data compiled from the latest round of the PIAAC surveys, Stanford University’s Erik Hanushek and his coauthors find that, controlling for other individual characteristics that might influence earnings, an increase of one standard deviation in quantitative literacy test scores is associated with an 18% increase in wages among prime-age workers, with the impact among U.S. workers equal to 28%. Similar results are obtained by Marguerita Lane and Gavan Conlon in their 2016 research. Other studies focusing on individual countries (e.g., Australia, Canada, Finland, and the United Kingdom) or even among subgroups of the population (e.g., immigrants or racial and ethnic minorities) have produced the same results.

The impact of quantitative literacy is not restricted to the labor market. Financial decisions, for example, require the capacity to understand quantitative concepts or carry out mathematical operations that go beyond the rudimentary knowledge many persons have. The research by economists Douglas Bernheim, Annamaria Lusardi, and others finds that failure to have the necessary numeracy skills can lead to a variety of effects, ranging from frequent misunderstandings regarding credit and borrowing, which can generate serious personal indebtedness, to the lack of adequate retirement financial planning, which can seriously affect the standard of living of the elderly.

Policy Directions

The rising importance of quantitative literacy in the labor market, in the financial sector, and everywhere in the economy and society, has led to an increasing emphasis in developing those skills in schools and in adult education. Countries with high quantitative literacy rates, like Denmark, Finland, and the Netherlands, also have high rates of participation in adult education programs. Financial literacy programs, such as those fostered by the Federal Reserve in the United States, have also been essential in developing the applied numeracy skills required by the complex financial transactions often confronted by consumers.

Francisco L. Rivera-Batiz

See also Literacy; Programme for International Student Assessment; Trends in International Mathematics and Science Study

Further Readings

Further Readings

Broecke, S. (2016). Do skills matter for wage inequality? (IZA World of Labor, No. 232).

Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessman, L. (2015). Returns to skill around the world: Evidence from PIAAC. European Economic Review, 73(1), 103–130.

Kirsch, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (2002). Adult literacy in America. Washington, DC: National Center for Education Statistics.

Lane, M., & Conlon, G. (2016). The impact of literacy, numeracy and computer skills on earnings and employment outcomes (OECD Education Working Papers, No. 129). Paris, France: OECD.

Lussardi, A. (2012). Numeracy, financial literacy and financial decision-making. Numeracy, 5(1), 1–12.

Meng, R., & Finnie, R. (2006). The importance of functional literacy: Reading and math skills and labour market outcomes of high school dropouts. Statistics Canada Analytical Studies Branch Research Paper Series. Ottawa: Statistics Canada.

Organisation for Economic Cooperation and Development. (2013). OECD skills outlook 2013: First results from the survey of adult skills. Paris, France: OECD.

Organisation for Economic Cooperation and Development. (2016). Skills matter: Further results from the survey of adult skills. Paris, France: OECD.

Rivera-Batiz, F. (1992). Quantitative literacy and the likelihood of employment among young adults. Journal of Human Resources, 27(2), 313–328.

Rivera-Batiz, F. (1996). English language proficiency, quantitative skills, and the economic progress of immigrants. In H. Orcutt Duleep & P. V. Wunnava (Eds.), Immigrants and immigration policy: Individual skills, family ties, and group identities (pp. 57–77). Greenwich, CT: JAI Press.

United Nations Educational, Scientific and Cultural Organization. (1978, November 27). Revised recommendation concerning the International Standardization of Educational Statistics. Retrieved from http://portal.unesco.org/en/ev.php-URL_ID=13136&URL_DO=DO_TOPIC&URL_SECTION=201.html

Janet Tsin-yee Leung Janet Tsin-yee Leung Leung, Janet Tsin-yee

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Quantitative Research Methods Quantitative research methods

1348

1352

Quantitative Research Methods

Quantitative research methods primarily rely on the collection and analyses of numerical data in the study of social phenomena. This methodological approach has been extensively applied in educational research. Embedded in the paradigm of positivism, quantitative research methods emphasize empirical inquiry to understand social phenomena. Educational research employing these research methods is expected to demonstrate internal validity (i.e., accurate interpretability of research results), external validity (i.e., generalizability of research results), and reliability (i.e., consistency and replicability of the methods and results) of the findings.

The Positivist and Postpositivist Paradigms

Quantitative approach as a research method has its roots in positivism. According to Sotirios Sarantakos, positivism views “reality” as objective, fixed, and independent of human consciousness. It is governed by natural laws that are strict and unchangeable. The world is regarded as deterministic, with causes producing effects under predictable conditions. Human actions are guided by fixed patterns that are empirically observable. As a tool of knowledge acquisition, science is based on strict rules and procedures that are deductive and nomothetic in nature. Hence, social research is a tool to examine social phenomena by revealing general causal laws and making predictions of outcomes. In the positivist view, science is empirical rather than metaphysical. Any propositions that cannot be tested and verified are meaningless.

Originating from the paradigm of positivism, postpositivism shares some fundamental principles with positivism. Postpositivism also believes in objective reality (i.e., reality as an objective entity governed by fixed natural laws). However, postpositivism shows some deviation from positivism in the ontological, epistemological, and methodological dimensions. According to the ideas of Yvonna Lincoln and Egon Guba, postpositivism shares the ontology of critical realism, which regards the objective world as imperfectly known and measurable. Hence, claims about reality are subject to critical examination on estimation as closely as possible. Epistemologically, postpositivism shares the view of modified dualist/objectivist (i.e., it is impossible to remove entirely the influence of the subject from the object of analysis), and objectivity is regarded as regulatory ideal. Methodologically, postpositivism employs critical multiplism (i.e., multiple methods of inquiry that are employed in revealing reality). Postpositivism permits a researcher to use quantitative methods in combination with qualitative methods in examining social phenomena.

Features of Quantitative Research Methods

Sharing the characteristics of positivism, quantitative research design has several unique features. First, the objectivity of the research is emphasized. Objectivity refers to the quality assurance that bias and subjectivity are minimized in data collection and analyses. Value neutrality is also expected in the research. The researcher is a neutral, objective scientist. Second, empiricism is stressed in quantitative research design. Empiricism means that the research is guided by evidence obtained from systematic research rather than by authorities. Third, accuracy and precision of measurements are determined by ensuring the reliability and validity of research. Fourth, logical reasoning is fundamental in quantitative research design. It relies on empirical methods having strict rules and clear procedures. Deductive methods such as hypothesis testing are employed. Fifth, parsimonious explanation is emphasized. The purposes of research are to explain the relationships among studied variables and reduce the explanations to simple general rules. Quantification of the results is emphasized with the use of mathematical models and statistical procedures and presentations. Last but not the least, replication of research is stressed (i.e., the results should be confirmed in subsequent research). Representativeness and generalization of the findings in explaining social phenomena and predicting outcomes are essential. As the aim of the research is to test a theory, further testing of the theory with different groups and under different contexts would help to confirm the theory or revise it.

Types of Quantitative Research Designs

Broadly speaking, there are two main types of quantitative research designs: experimental and nonexperimental research design. Experimental research design utilizes the principle of manipulation of the independent variables and examines its cause-and-effect relationship on the dependent variables by controlling the effects of other variables. Usually, the experimenter assigns two or more groups with similar characteristics. Different interventions will be given to the groups. In case there are differences in the outcomes among the groups, the experimenter can conclude that the differences result from the interventions that the experimenter performed. An example of an experimental design is an examination of whether there is any difference in students' learning motivation between classroom learning and experiential learning using an experimental group with intervention and a control group without intervention.

There are different types of experimental designs. In true experimental design, the subjects are assigned randomly to different groups. The random assignment procedure of group formation helps to minimize the differences of subjects' characteristics between different groups before intervention. Quasi-experimental design is similar to true experimental design, except that there is no random assignment of subjects to different groups. Single-subject designs involve one or a few subjects, but the cause-and-effect conclusion is drawn through repeated measurements. Pre-experimental design does not have a comparison group. Hence, the experimenter only measures the posttest results of the subjects (i.e., posttest-only design), or the experimenter measures pretest and posttest scores and assesses the changes between the tests (i.e., pretest–posttest design).

In contrast to experimental designs, nonexperimental designs are research designs that examine social phenomena without direct manipulation of the conditions that the subjects experience. There is also no random assignment of subjects to different groups. As such, evidence that supports the cause-and-effect relationships is largely limited. There are two main types of nonexperimental designs: comparative design and correlational design. In comparative research, the researcher examines the differences between two or more groups on the phenomenon that is being studied. For example, studying gender difference in learning mathematics is a comparative research. The correlational design is a study of relationships between two or more constructs. A positive correlation means that high values of a variable are associated with high values of another variable. For instance, academic performance of students is positively related to their self-esteem. On the contrary, a negative correlation means that high values of a variable are associated with low values of the other variable. For example, teacher–student conflicts are negatively related to the students’ sense of belonging to the school.

Methods of Data Collection

To ensure that the instruments and tests are adequate to measure the constructs, validation tests to assess the reliability and validity of the measurements are necessary. Reliability refers to the consistency of the measurement (i.e., the extent to which the measures are free from errors). Thomas Black lists seven methods of reliability estimates: (1) test–retest reliability—a single instrument is administered by a group of respondents more than once to assess the temporal stability of the measurement, (2) parallel forms—reliability is estimated by assessing the equivalence between two different forms of a measure designed to measure the same domain, (3) split-half estimate—the single instrument is divided into two equivalent halves and is administered by a group of respondents at one time, (4) Cronbach’s α—the internal consistency approach to assessing consistency among items measuring the same constructs, (5) Kuder–Richardson reliability—estimation of internal consistency for dichotomous responses, (6) scorer reliability (interrater and intrarater reliability)—two or more raters rate the same instrument (interrater) or a single scorer rates the instrument over time (intrarater) to obtain the degree of agreement in the ratings, and (7) estimate of reliability in criterion-referenced tests—assessing reliability with the distributions of scores that are not normally distributed.

Validity refers to the extent to which the instrument adequately reflects what it is designed to measure. According to Allen Rubin and Earl Babbie, there are five approaches for assessing measurement validity: face validity, content validity, criterion-related validity, construct validity, and factorial validity. Face validity is the judgment that an operational definition appears in the measurement. Content validity refers to the assessment of content of the measurement based on the solicited opinions of the researchers and experts. Criterion-related validity is the extent to which the scores of the measurement can be accurately compared with some criteria external to the test. Typically, there are two types of criterion validation: concurrent validity and predictive validity. Concurrent validity is established when the measurement scores are closely related to the scores of a criterion measured at the same time. Predictive validity is the ability of a measure to predict scores of a criterion measured in the future. Construct validity refers to the extent to which a measure relates to other variables within a system of theoretical relationships. Factorial validity is the assessment of whether the items making up the factors are the components the researcher anticipates to measure and associate.

Standardized tests, scales, and questionnaires are extensively used in data collection of quantitative research in education. A standardized test is a standard set of structural questions used to assess the subjects’ attributes and proficiency. Standardized tests are administered and monitored by uniformed procedures, including the qualifications of the implementers, the conditions of the administration, and the time allowed to perform the tests. The scoring of the responses is objective, and the results are interpreted by trained qualified professionals and/or researchers.

In the interpretation of scores obtained by objective tests, two major categories, norm-referenced and criterion-referenced interpretations, are generally used. In a norm-referenced interpretation, the individual score is compared with the scores of a well-defined reference group (i.e., the norm group). As the interpretation of the results depends on how individual scores are compared with others in norm-referenced interpretation, the characteristics and formation of the norm group as well as the ability of an instrument to differentiate between individuals are critical. In contrast, criterion-referenced interpretation is based on a set of criteria or standards defined by professionals in making the assessment. Individual scores are compared with the standards of performance set by the professionals.

In education, achievement tests and aptitude tests are tests commonly performed in schools. Achievement tests assess what the students have learned in terms of academic aspects and skills. Examples of achievement tests are the Stanford Diagnostic Mathematics Test, the California Diagnostic Reading Test, and the Stanford Achievement Test Series. Aptitude tests emphasize the assessment of abilities of an individual for future performance. They provide evidence of the potentials of individuals in their performance. The Wechsler Intelligence Test and the SAT are well-known examples of aptitude tests.

Self-reported validated measurements are frequently used in educational research to assess different dispositions and traits of individuals, including personality, attitudes, values, affections, behaviors, and interests. The measurements are a set of items related to the concepts of interest, and the respondents are required to rate their agreements on a rating scale with indicators at different levels. Examples of validated measurements include the Rosenberg at different levels. Examples of validated measurements include the Rosenberg Self-Esteem Scale, the Minnesota Vocational Interest Inventory, the Positive Youth Development Scale, and the Omnibus Personality Inventory. For instance, the Chinese Positive Youth Development Scale contains 90 items measuring 15 aspects of adolescents' developmental assets, including bonding, resilience, social competence, emotional competence, cognitive competence, behavioral competence, moral competence, self-efficacy, self-determination, spirituality, beliefs in the future, prosocial involvement, prosocial norms, and recognition for positive behaviors.

In most cases, the data collection is performed by paper-and-pencil tests or questionnaires. As information technology has been developing rapidly, the respondents may answer the tests or questionnaires using computers.

Data Analyses

Statistical methods are used to analyze the quantitative data. Generally speaking, there are two types of statistical techniques: descriptive and inferential. Descriptive statistics are used to summarize and organize large numbers of observations to describe the data and make sense of them. By summarizing and reducing the data into some meaningful statistical results derived from mathematical formulas, the characteristics of the data can be interpreted. There are several measures of descriptive statistics. The frequency distribution shows the number of times each score occurs and is mostly represented by means of percentage and graphic representations. Measures of central tendency, including the mean (i.e., arithmetic average of the scores), the median (i.e., midpoint of the distribution of the scores), and the mode (i.e., the most frequent score), are locators of the distribution of the scores. Measures of variability present the dispersion and spread of the scores. These include the range (i.e., difference between the highest and lowest scores), the standard deviation (i.e., average dispersion of scores around the mean), and the variance (i.e., measure of dispersion of scores). Last but not least, measures of the relationship tell the relationship between two variables. A scatterplot (i.e., a graphic representation of the relationship between two variables) is commonly used to show the direction and shape of the relationship. Moreover, bivariate correlation is extensively used in measuring the relationship. A correlation coefficient is calculated to show the direction and strengths of the relationship. Among different correlation coefficients, the Pearson product-moment coefficient, mostly represented as *r*, is the most common technique in measuring bivariate relationship between the two variables.

Inferential statistics, in contrast, use the data of a subset (i.e., sample) and make inferences to the population. Generally speaking, there are two procedures: hypothesis testing and parameter estimation. The researcher tests the hypothesis to see whether it is consistent with the data of the sample. Parameter estimation can be conducted by means of point estimation (i.e., estimating the parameter by a single value) and interval estimation (i.e., defining the confident interval on the measurement scale that contains tenable values of a parameter). To test the mean difference between groups, *t* tests (e.g., paired *t* test), analysis of variance tests, analysis of covariance tests, and nonparametric tests (e.g., Mann-Whitney U test) are used. To assess the relationships between variables, correlational and regression analyses are performed.

Advanced statistical techniques, such as structural equation modeling, are used to assess the relationships between different variables. According to Barbara Tabachnick and Linda Fidell, structural equation modeling is a confirmatory technique to assess theory-derived causal hypotheses. It deals with the statistical estimation of relationships between factors of variables, tests complex structural models on the interrelationships between latent variables, and assesses the direct and indirect effect of independent variables on dependent variables.

Guiding Principles of Evidence-Based Inquiry

Evidence is important in educational research. According to the National Research Council in 2002, there are six guiding principles that address the evidence-based inquiry. First, the research questions should have an impact on knowledge and/or practice and can be investigated empirically. Second, the research should be linked to a theory or conceptual framework; that is, the theories and conceptual frameworks help to explain the results. Third, the research method provides empirical data for the investigation of the research questions. Fourth, there is a coherent and logical chain of reasoning throughout the research process, with detailed descriptions of procedures and analyses. Fifth, the results can be generated and replicated across studies. Sixth, the findings can be disseminated to peers and professionals for inquiries and critiques.

Application in Educational Research

Educational research can have four main functions: basic, applied, evaluation, and action. Basic research aims at testing specific theories to explain social phenomena related to human behaviors. Basic research is fundamental for knowledge building and development. An example of basic research is a longitudinal study on assessing the relationships between school performance and life satisfaction and hopelessness of secondary school students.

Applied research has its function to develop research-based knowledge concerning practice. It focuses on answering questions related to practice and seeks to find out solutions to improve practice. The topics are mainly related to current issues in education that concern educators and policy makers. An example of applied research is a survey conducted on the teaching styles of teachers in schools with different bandings.

Evaluation research determines how well a particular practice or program performs in a real-world setting and examines how it can be improved. In other words, evaluation research determines the quality, merits, and worth of a particular practice or program. There are five aspects in which a practice or program can be evaluated: need assessment, theory assessment, implementation assessment, impact assessment, and efficiency assessment. An example of an evaluation research is to conduct an evaluative study on assessing the effectiveness of a positive youth development program in secondary schools.

Action research focuses mainly on solving a specific problem or issue that local practitioners may face in schools and communities. It can be implemented in three levels: individual practitioner research, team research in a single school, and school-wide research. Action research is different from basic and applied research, as it emphasizes local practice and issues. It is mainly initiated and conducted by teachers, practitioners, administrators, and other educational professionals. Hence, it is more participatory in nature. An example of action research is to study the teachers’ strategies in handling the behavioral problems of students experiencing truancy.

Janet Tsin-yee Leung and Daniel Tan-lei Shek

See also Correlation; Experimental Designs; Inferential Statistics; Objectivity; Positivism; Postpositivism

Further Readings

Black, T. R. (1999). Doing quantitative research in the social sciences. London, UK: SAGE.

Burke, J. (2012). Educational research: Quantitative, qualitative, and mixed approaches. Thousand Oaks, CA: SAGE.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). Handbook of qualitative research (2nd ed.). Thousand Oaks, CA: SAGE.

Lincoln, Y. S., & Guba, E. G. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (2nd ed.), pp. 163–187). Thousand Oaks, CA: SAGE.

McMillan, J. H., & Schumacher, S. (2010). Research in education: Evidence-based inquiry (7th ed.). Boston, MA: Pearson Education.

National Research Council, Committee on Scientific Principles for Education Research. (2002). Scientific research in education. Washington, DC: National Academy Press.

Rubin, A., & Babbie, E. R. (2014). Research methods for social work. Belmont, CA: Thomson Brooks/Cole.

Sarantakos, S. (2013). Social research. Basingstoke, UK: Palgrave Macmillan.

Shek, D. T. L., & Li, X. (2016). Perceived school performance, life satisfaction, and hopelessness: A 4-year longitudinal study of adolescents in Hong Kong. Social Indicators Research, 126, 921–934. doi: 10.1007/s11205-015-0904-y

Shek, D. T. L., & Ma, C. M. S. (2014). Effectiveness of a Chinese positive youth development program: The Project P.A.T.H.S. In Hong Kong. International

Journal on Disability and Human Development, 13(4), 489–496. doi:10.1515/ijdhd-2014-0346

Shek, D. T. L., Siu, A. M. H., & Lee, T. Y. (2007). The Chinese Positive Youth Development Scale: A validation study. Research on Social Work Practice, 17, 380–391. doi:10.1177/1049731506296196

Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics. Boston, MA: Pearson Education.

Wiersma, W., & Jurs, S. G. (2005). Research methods in education (8th ed.). Boston, MA: Pearson.

S. Earl Irving S. Earl Irving Irving, S. Earl

Quartile

Quartile

1352

1353

Quartile

Quartile is a rank-order grouping. A quartile divides a distribution of data into four equally sized groups determined after ranking the data according to some measure or combination of measures.

There is some debate about who first used the term *quartile* (contenders include Carl F. Gauss, Donald McAlister, and Francis Galton), but it seems that Galton was the first to bring the terms *quartile, decile*, and *percentile* into common use.

The term *quartile* can have two meanings: a point on the distribution that divides the four groups, and the group to which a member of the distribution belongs. There are three decile values, which separate the data into the four groups. The third decile value (commonly known as the upper quartile) is the point in the distribution where three quarters of the data lies below that value. The second decile (commonly known as the median) is the point below which two quarters (one-half) of the distribution lie. One quarter of the distribution lies below the first quartile (known as the lower quartile). If a point on the distribution lies between the median and upper quartile, then it is a member of quartile 3. The interquartile range is the difference between the upper quartile and the lower quartile.

There is no precise definition for calculating the quartile values. Different software and different statisticians can use different values for the same distribution, but the differences are usually not important enough to impact the exploratory nature of this description or the interpretation of the data.

When an analyst breaks data into quartile groups, the purpose is to simplify the way in which they can describe and visualize the data, such as to look for patterns and trends, or to compare and contrast high-performing and low-performing groups.

A good example of this is the five-point summary, and the visual representation of this summary, the box plot. Box plots (or box-and-whisker plots) were devised by John Tukey in 1969 as an exploratory data tool to visualize some characteristics of a distribution. The box is plotted using the median and the upper and lower quartiles (called hinges by Tukey). Therefore, half of the data lies inside the box. There are variations in the statistic used to determine the length of the whiskers. Tukey used 1.5 × interquartile range to determine the end of the whiskers (he called these the inner fences), but the maximum and minimum value are also commonly used. There are other less common variations, so it is important to state what convention has been used. If values other than the maximum and minimum are used for the whiskers, outliers can be also shown.

S. Earl Irving

See also Box Plot; Decile; Descriptive Statistics; Interquartile Range; Percentile Rank

Further Readings

Galton, F. (1881). Report of the Anthropometric Committee. Report of the British Association for the Advancement of Science, 51, 225–272.

Hald, A. (1998). A history of mathematical statistics from 1750 to 1930. New York, NY: Wiley.

McAlister, D. (1879). The law of the geometric mean. Proceedings of the Royal Society of London, 29(196–199), 367–376. doi:10.1098/rspl.1879.0061

Rogers, T. B. (1995). The psychological testing enterprise: An introduction. Belmont, CA: Brooks.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Walker, H. M. (1929). Studies in the history of statistical method. Baltimore, MD: Williams and Wilkins.

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Jing Wu Jing Wu Wu, Jing

Quasi-Experimental Designs

Quasi-experimental designs

1353

1356

Quasi-Experimental Designs

Quasi-experimental designs share the same purpose with true experiments that attempt to test the causal impact of an independent variable such as an intervention by manipulating the intervention and observing the outcome. However, a quasi-experimental design does not utilize random selection and/or random assignment of participants. Rather, participants are assigned nonrandomly to the experimental and control groups.

There are three basic forms of quasi-experimental designs. The first type is *nonequivalent groups design* (NEGD), in which researchers often use intact groups (expected to be as similar as possible before the intervention) as experimental and control groups. The second form is *time series design*, in which multiple consecutive observations on an outcome are measured before and after the intervention over time. Discontinuity in behavior in the time series after the intervention is interpreted as the existence of its impact. The third type is *regression discontinuity design* (RDD), in which researchers use a cutoff score on a measured preintervention variable to decide the eligibility for an intervention. If the outcome variable for the treatment group is discontinuous from that of the control group, an effect of treatment is inferred. This entry describes the basic types of quasi-experimental designs, a brief development history, and an overview of the strengths and weaknesses of quasi-experimental designs.

The word *quasi* means “seemingly” or “almost” in Latin. Quasi-experimental designs are alternatives when it is impossible to meet the requirements for a true experiment. In true experimental designs, the participants are randomly selected and assigned to conditions to make sure the manipulated outcomes are not affected by systematic individual differences. However, as real-life educational field settings are complex, randomizing depends on many constraints. First, it is sometimes impossible to determine the population. Taking “at-risk students” (e.g., substance abuse) as an example, it is not possible to identify the population because of underreporting of risk behavior by the students. Second, random assignment of participants to the experimental group or the control group may be difficult in some intervention contexts. For example, when treating students with depressive symptoms, it would be unethical to randomly assign students to the control group. Random assignment of students to different intervention conditions may also be difficult because of practical time and class arrangement constraints. Third, educational evaluators may seek for more general results with higher external validity instead of findings based on artificial laboratory settings. Finally, it would be too expensive to conduct randomized controlled trials in large-scale educational research.

Classical Quasi-Experimental Designs

NEGD

To assess the effectiveness of an educational program or an intervention, groups with different attributes (such as classes, grades, or schools) are often used to compare with each other. This is the NEGD, which is widely used in educational studies. It works like a true experiment but without random assignment. In a common NEGD, researchers use a treatment group and a control group that are supposed to be similar and assess both groups at pretest. Then the intervention condition is implemented to one group, and posttests are given to both groups.

If the two groups are similar in their pretest scores prior to intervention but differ in the posttest scores afterward, researchers can be more confident to declare the effect of the intervention. However, the groups selected cannot always be guaranteed to be alike in all possible ways expected. Participants assigned to the experimental group may differ from the ones in control condition in many systematic (nonrandom) ways other than just the presence of the intervention. Thus, the outcome of the study may be affected by many other factors, such as history (some unanticipated event occurs while the intervention is in progress, which affect the dependent variable), maturation (changes in the dependent variable may be due to normal developmental processes instead of the intervention), testing (pretest may affect the scores on the posttest), contamination (participants in the control group may know about the intervention through a third party), and instrumentation (the examiners, instructions, and procedures may be different in different groups). These are the biggest threats to the internal validity of an NEGD. For this reason, researchers have to enumerate and rule out other alternative explanations for the observed effect as far as possible to get a more valid estimate of the intervention effect. For the external validity of an NEGD, replication of research findings with different time, population, and setting parameters is needed.

Time Series Designs

If a group received repeated measurements before and after the intervention rather than once at pretest and once at posttest, a time series design is established. In a typical time series design, multiple pretests are taken at equal intervals to establish a baseline, and the intervention is followed by several posttests sequentially. Multiple observations promise a more stable and accurate estimation of the effectiveness of an intervention.

A longitudinal time series design usually lacks a control group, which makes it different from the NEGD. The single-group time series design requires only one group and multiple assessments before and after the intervention. An overall trend of continuous positive changes observed across the multiple time points demonstrates a meaningful effect. Sometimes a control or a comparison group is added to refine a single-group time series design. It is then called a multiple time series design. The control group helps rule out alternative extraneous causes that would be expected to affect both groups. It also assists in controlling maturation and instrumentation effects. However, the addition of a control group also raises the practical question of group-selection procedures in some situations.

The major threats to internal validity in a simple time series design are history effects, changes in measurements (instrumentation), and experimental mortality (differential loss of participants across groups). As for external validity, interaction between selection of the population and the particular intervention needs to be considered, including the interaction effect of testing and the interaction effects of selection biases and the experimental treatment.

RDD

An RDD assigns participants who satisfy some criteria to the intervention group. A cutoff score on a measured continuous variable is used to determine the criteria. Accordingly, RDDs are especially appropriate when researchers intend to target an intervention to those who most need it. Regression lines of the outcome variable for the treatment group and control group are compared. Researchers expect a discontinuity in the regression line to appear exactly at the cutoff point for the treatment group. The discontinuity suggests the effectiveness and causality of an intervention.

Donald L. Thistlethwaite and Donald T. Campbell published the first article about RDD evaluating an application in education in the 1960s. Although not popularly adopted, it has still become an established effective tool in psychology and education since the 1990s. RDD was used in some of the large national evaluation studies at the Institute for Educational Sciences in the United States, such as the Reading First, Early Reading First, and the George W. Bush administration’s No Child Left Behind legislation.

The RDD is quite different from the NEGD and the time series design. First, the variables tested at preintervention and postintervention are not the same in RDDs. More importantly, in true experiments and most quasi-experiments, researchers try their best to achieve the equivalence between the experimental group and the control group. However, an RDD has a predetermined cutoff criterion, which deliberately avoids random assignment. Therefore, participants in the treatment group and the control group are maximally different on the pretested variable. Besides, RDD needs 1.75 times more participants than a randomized experiment to achieve the same statistical power. The attractiveness of an RDD is that it provides intervention for the ones most in need. However, there are also several threats to the internal validity and external validity of an RDD. First, it is not always easy to set up scientific criteria for the cutoff value. The cutoff criterion may be well correlated with the outcome, resulting in selection bias. Meanwhile, an RDD provides limited external validity. The effect estimate is generalizable around the cutoff provision. However, it may be different for those further from the cutoff point.

Role of Quasi-Experimental Designs in Educational Research

Most experiments conducted before the 1920s were actually quasi-experimental designs. For example, in 1918, Walter F. Dearborn and John M. Brewer conducted a nonequivalent control groups design to study college students' transfer learning ability. Donald T. Campbell and Julian C. Stanley are believed to be the first ones to use the term *quasi-experiments*, in their 1963 book *Experimental and Quasi-Experimental Design for Research*. In the next four decades, Campbell and his colleagues extended and practiced this class of designs in two ways. First, they explored and described more types of quasi-experimental designs. For example, quasi-experimental designs can be either inherently longitudinal (e.g., time series design) or made longitudinal by adding more observations before or after intervention. Second, they developed four validity types (statistical conclusion validity, internal validity, construct validity, and external validity) to evaluate the integrity and quality of causal inferences resulting from quasi-experimental designs. At the same time, they enumerated several methods to prevent the threats to validity when utilizing a quasi-experimental design. For example, maturation (normal development over time) is one of the most common threats to validity. To make sure whether the rate of change during the intervention is similar to the maturation rate before intervention, consecutive pretests before the intervention were introduced. Moreover, Campbell also developed statistical analysis methods to adjust the potential threats after they have already occurred, such as analysis of covariance. Several other scholars from different research fields have also contributed to the development of quasi-experimental designs, such as William G. Cochran in statistics and James J. Heckman in economics.

Untested and unevaluated educational interventions were very common prior to the late 1900s. The dominance of evidence-based education reform facilitated the development of research methods in education. Since the 1960s, almost all developed societies have sought to improve the performance of school systems. Taking the United States as an example, RDDs were adopted in the nationwide evaluation system for compensatory education programs funded under the Elementary and Secondary Education Act in 1965. The U.S. government enacted the Comprehensive School Reform Demonstration legislation of 1997 and the No Child Left Behind Act of 2001. Both laws and related policies emphasize the application of experimental and quasi-experimental research in education. The increasing international cooperation in educational planning and policy (such as Trends in International Mathematics and Science Study and Programme for International Student Assessment) also requires a closer coordination of experimental and quasi-experimental designs and practices across nations experimental and quasi-experimental designs and practices across nations.

Conclusions

The experimental group and control group in true experiments are treated identically in every aspect other than the manipulated presumed independent variable. Hence, experimental designs are well suited for causal inference. However, sometimes there are a number of ethical, practical, legal, or political reasons that experiments cannot be implemented, particularly in educational research. Utilizing quasi-experimental designs are beneficial in education-related studies. They are easier and more feasible to implement. Also, quasi-experimental designs minimize threats to ecological validity and they allow for more generalizations. In addition, quasi-experimental designs can be easily followed up in different environments without strict control. Thus, they may be more efficient in longitudinal research. However, the correlational nature and lack of randomization of quasi-experimental designs pose threats to internal validity to a great extent. Therefore, researchers using quasi-experimental designs try their best to rule out unrelated explanations through efforts such as matching participants and statistical analysis and thus show that the outcome can be attributed solely to the intervention.

Daniel Tan-lei Shek and Jing Wu

See also Causal Inference; Experimental Designs; Posttest-Only Control Group Design; Pretest–Posttest Designs

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental design for research. Chicago, IL: Rand McNally.

Cook, T. D. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics, 142(2), 636–654.

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

Karen D. Multon Karen D. Multon Multon, Karen D.

Quota Sampling Quota sampling

1356

1358

Quota Sampling

Sampling designs are typically separated into two fundamental types: probability and nonprobability. In probability-based sampling, each member of the population has a known, nonzero chance of being selected. Nonprobability samples are generated from more subjective criteria of the researcher, such as personal experience, convenience, and volunteers. Quota sampling is one form of a nonprobability or judgmental sampling design used to acquire data from population subgroups.

In quota sampling, participants or locations are selected nonrandomly according to a fixed quota or percentage of the population based on one or more characteristics. The quota selected may be proportional or nonproportional to the actual population distribution. In order to obtain a representative sample, a proportional quota sampling design requires a priori knowledge about the underlying population characteristics. Nonproportional quota sampling is less restrictive, as the researcher specifies the percentage of data elements to be sampled from each subgroup independent of the population.

The quota sampling procedure is outlined as follows: (a) divide the population into subgroups that are exhaustive and mutually exclusive (i.e., all data points occur in one and only one division), (b) stratify the population into classes based on one or more characteristics (e.g., gender) and determine the proportion in each class, (c) pick a sample size, (d) select a quota for each subgroup that may either be proportional or be nonproportional to the population, and (e) collect data points until the quotas are completed.

For example, a psychologist is interested in the attitudes of people toward therapy based on U.S. political party affiliation in their area. From voter registration data, the population is 60% Democrat, 35% Republican, and 5% Independent. A sample size of 200 adults is chosen with political affiliation as the criterion. For a proportional quota sample, the psychologist selects 200 individuals who represent the political party distribution of their area (i.e., 120 Democrats, 70 Republicans, and 10 Independents). If the psychologist is more interested in the attitudes of Republicans, the quotas could be altered to a nonproportional sample with a higher sample size for Republicans than is found in the general population (i.e., greater than 35%); however, the sample would be less representative of the population.

Quota sampling offers several advantages over more complex probability sampling methods. For primary data collection, quota sampling is relatively inexpensive, quick, and simple. Researchers can also ensure data are collected from all subgroups for a given set of characteristics, thus guaranteeing smaller groups are represented in the sample. Yet, a major disadvantage of quota sampling is that the selection process is nonrandom and subjective, especially for nonproportional samples. Consequently, difficulties arise in determining the sample error or making inferences about the general population from the sample.

The shortcomings of quota sampling are often reduced in stratified random sampling, the probability-based alternative whereby an element of randomness is introduced. Using stratified random sampling, each data member of the population has the same probability of being selected as any other member throughout the sampling process. In contrast, quota sampling reduces the chance of a data member being chosen as the quota is filled.

Jill S. M. Coleman and Karen D. Multon

See also Convenience Sampling; Stratified Random Sampling

Further Readings

Levy, P. S., & Lemeshow, S. (2008). Sampling of populations: Methods and applications (4th ed.). New York, NY: Wiley.

R

Alon Friedman Alon Friedman Friedman, Alon

R

R

1359

1361

R

R is a powerful language and software environment for statistical and graphics calculation. It is an open-source program, which means it is free to use, and its popularity reflects a shift in statistical computing and visualization output. The importance of R as a statistical language in today's data-rich environment is its utility in analyzing and visualizing large data sets with the strong backup of its community of users and developers. The popularity of R continues to grow and achieve a global reach not only with academic researchers and programmers but also with private companies and local and national governments. After reviewing the history of statistical computing and open-source software, this entry focuses on open-source R and the solutions it provides for analyzing large data sets.

The History of Statistical Computing

The field of statistical computing became a popular field even before the appearance of the first mechanical calculator machine. In the 1920s, universities and research labs began to acquire the early IBM mechanical punch card tabulators. They used these machines not only for tabulating and computing statistical summaries but also for the calculation of more complicated statistical models, such as analyses of variance and linear regressions. During the 1960s, the growing interest in statistical computing brought on the development of large mainframe computers. At that time, two additional developments in statistical computing applications occurred. The first occurred in 1968 at the University of Chicago, where Norman H. Nie led his team in the development of SPSS software. They aimed to develop targeted statistical software for the social sciences fields. At the same time, a different application was developed at North Carolina State University in its agricultural department. The application, called SAS, was originally spearheaded by Anthony J. Barr and was created to aid the business community. At Bell Laboratories, John Chambers and his team started to develop the S language as a statistical programming language. The S language is a computing language and an interpreter wrapped around compiled code for numerical analysis and probability. However, the S language never reached a wide audience as SPSS and SAS successfully did during that time. Overall, these three applications were developed for and offer different statistical solutions for different audiences.

The History of Spreadsheets

In the 1980s, another major development occurred when Daniel Bricklin together with Bob Frankston developed an interactive visible calculator, also known as a spreadsheet. Bricklin named this software program VisiCalc. However, the market during that time was experiencing rapid changes that VisiCalc could not adopt. The leading PC maker, at the time, IBM, started to develop its own spreadsheet built around its technology called Lotus 1-2-3. In 1985, Bill Gates introduced his Excel spreadsheet software application that later became one of the foundations of Microsoft Office. With the success of Microsoft Office in capturing the global market, the Microsoft business model promoted the idea that the end user is charged for utilizing its product line.

Open-Source Software

With all the development of statistical programming and spreadsheets during the 1970s and 1980s, another major change occurred that influenced the landscape of the computer industry: the appearance of the open-source movement. Open source refers to software for which the source code of the application is available to the general public for use and/or modification from its original design free of charge (i.e., open). Open-source code is typically created as a collaborative effort in which programmers improve upon the code and share the changes within the community. Open source emerged in the technological community as a response to expensive proprietary software owned by corporations.

In the statistical computing environment, the introduction of an open-source application occurred in 1995 when two faculty professors from the University of

Auckland in New Zealand, Ross Ihaka and Robert Gentleman, revised the S language and converted it to open-source code, also known as the GNU project. They called the revised language “R” based on their first names, Ross and Robert. Together with 17 other people, they established the R-code Foundation.

Open-Source R

The development of R allowed its community members to address specific problems that the traditional spreadsheet software was not able to capture. An important reason for R’s popularity is its packages. A package is essentially a library of prewritten code designed to accomplish a specific task or a collection of tasks. Today, there are more than 6,000 packages available on the Comprehensive R Archive Network each written by different individuals.

Recent interest in analyzing large sets of data has created new problems. This type of data usually appears in an unstructured form, making it hard to analyze through conventional techniques and applications. Providing a solution, R allows its community to conduct a number of tasks that are essential for the effective processing and analysis of big data. One of the advantages of R is that it facilitates data management processes, such as transformations, subsetting, and “cleaning,” and helps users to carry out exploratory data analysis and prepare the data for statistical testing. R also contains numerous ready-to-use machine learning and statistical modeling algorithms that allow users to analyze big sets of data. There are also “big data packages” in the Comprehensive R Archive Network library with built-in functions to help the analysis of these types of data. These packages include a variety of approaches to mitigate or minimize the memory choke point.

In any data analysis task, a majority of the time is dedicated to data cleaning and preprocessing. Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. For large sets of data, data cleaning is an essential procedure, and it is preferable to perform the operation within subgroups of a data set to speed up the process. In R, this type of data manipulation can be done with base functionality. One of the most popular R packages available to sort this type of data is “plyr.” This package consists of a set of tools for a common set of problems: the ability to split up a big data structure into homogeneous pieces, apply a function to each piece, and then combine all results back together. Plyr builds on the idea of built-in functions by giving the user control over the input and output formats and keeping the syntax consistent across all variations. Plyr also provides other niceties like error processing, parallel processing, and progress bars. Furthermore, the package allows the user to run parallel processing on different machines.

R allows the user to produce visualization as an essential part of the statistical analysis. The primary goal of visualization is to communicate information and data results clearly and efficiently to users via statistical graphics, plots, information graphics, and charts. There are many packages in R that allow the generation of visualization from data; however, “ggplot2” stands out. The ggplot2 package was created by Hadley Wickham in 2005. The ideology behind the package is that visual grammar can be applied to a statistics report. David Cox, known for his contribution to the field of statistics, reported that good grammar will allow users to gain insight into the composition of complicated graphics and reveal unexpected connections between seemingly different graphics. In 2005, Wilkinson and colleagues made the comparison between grammar and statistics to report that grammar can be used to describe and construct a wide range of math and statistical calculations. In the ggplot2 package, Wickham proposes the alternative parameters necessary for the visual grammar to be successful in the R environment. ggplot2 is built around the idea of adding layers to the visualization of the chart. In contrast to base R graphics, ggplot2 allows the user to add, remove, or alter components in a plot at a high level of abstraction. This abstraction comes at a cost, however, as ggplot2 is slower to produce than lattice graphs.

Initially, the R console consisted of a basic command line interface. Under the graphical user interface, the user had to type the code to make R functions or run the analysis. After the user had typed the commands, the user needed to press the Return key to see the answer. With all progress in R and its packages through the help of its supporting community, in 2010 yet another major development occurred: J. Allairem, the creator of the programming language ColdFusion, introduced RStudio software. RStudio provides its users a window interface that is very similar to development. In more recent development, RStudio continues to develop a server side and web-based application called Shiny.

With all the benefits of open-source R in statistical computer programming, one of the merits that stands out for R is its transparency. The benefits of transparency in the statistical calculation may include fostering trust with one’s customers and community, setting an example of business and scientific practices, and allowing creative problem solving out in the open. In R, the user practices, and allowing creative problem solving out in the open. In R, the user can share any statistical calculation in the code and data with the user's community to receive feedback and additional recommendations. This practice is not available in other standardized statistical applications.

R has a global community of more than 4 million users and developers who voluntarily contribute their time and technical expertise to maintain, support, and extend the R language and its environment, tools, and infrastructure. At the center of the R community is the R Core Group of approximately 20 developers who maintain R and guide its evolution. The "official" public structure for the R community is provided by the R Foundation. This foundation is a nonprofit organization that ensures the financial stability of the R project that holds and administers the copyright of R software and its documentation. The R community is home to more than 150 user groups throughout the world that discuss new R packages and functions, present applications, share codes, and best practices. LinkedIn, one of the largest professional network sites, is itself home to more than 100 user groups that share and guide their members on how to apply and build better R packages.

Worldwide, millions of statisticians and data scientists use R to solve their most challenging problems in fields ranging from computational biology to quantitative marketing. R and the R community aim to incorporate every data manipulation, statistical model, and visual pattern that the modern data scientist could ever need.

Alon Friedman

See also SAS; SPSS

Further Readings

Cox, D. R. (1978). Some remarks on the role in statistics of graphical methods. Applied Statistics, 27(1), 4–9.

Ihaka, R. R. (1996). Past and future history. In Proceedings of Interface 96. Retrieved from https://www.stat.auckland.ac.nz/~ihaka/downloads/Interface98.pdf

Maciej Taraday Maciej Taraday Taraday, Maciej

Anna Wieczorek-Taraday Anna Wieczorek-Taraday Wieczorek-Taraday, Anna

R^2

R^2

1361

1363

R^2

R^2 (pronounced as “R squared”), or the coefficient of determination, is a statistical measure that is interpreted as the proportion of variance of the dependent variable (DV, or regressand) that is explained by the independent variable (IV, or predictor or explanatory variable) or by the statistical model. In other words, the R^2 value gives the information on how well the IV explains the outcome variable (accounts for the variability of the DV).

The value of R^2 is standardized (ranges from 0 to 1), which makes it easy to interpret; therefore, it is commonly used in statistical models for research (e.g., in education, psychology, biology, or economy). In social sciences, the coefficient of determination might be used in studies aimed at predicting school grades or estimating the outcome theoretical model that takes into account various related measurements. The R^2 value indicates how well a model fits to the set of observations or the difference between observed and expected values. The higher the R^2, the better the model’s goodness of fit.

How to Interpret the R^2

Because R^2 is standardized, its value varies between 0 and 1. After multiplying that value by 100%, you obtain the percentage of variability explained by the IV or the statistical model. If the R^2 value is equal to 0, you know that the IV does not explain the variance of DV at all (0 × 100% = 0% variability explained), whereas a value of 1 would mean that the variance of outcome variable is fully explained by the IV (1 × 100% = 100% variability explained). However, models that explain 100% of DV variance rarely happen.

For example, let us say you would like to predict the educational success of college students in one of the courses they attended. To assess it, you perform a research. You measure each student's IQ, which accounts for intellectual abilities. As an outcome measure—DV—you measure each student's grade at the end of the school year. After collecting data from the group of participants, you run a regression analysis using a statistical software program, in which the student's grade is an outcome (dependent) variable and the IQ is an IV. The statistical program prints the output of the analysis, which provides you with information about your model. The R^2 value is .5. What does this mean?

The IV explains 50% (0.5 × 100% = 50%) of variance of the student's grade. But you might ask, where is the lacking 50% of explained variance? The 50% of the variance that this model did not account for are the variables that were not included in the model, such as student's commitment to studying, socioeconomic status, or interest in the course.

Low R^2 Versus High R^2

A question that often arises in the context of estimating goodness of fit of the model is how does one know if the value on the output is "high enough." The answer depends on the research area. Usually, social scientists obtain lower R^2 values than biologists do due to the fact that behavior depends on a complex set of variables and the fact that social scientists fail to control all of them.

On the other hand, low R^2 might also be the effect of choosing an inaccurate or unreliable research method. The value of R^2 depends on whether or not the researcher has transformed the data prior to running the model. The low R^2 indicates that the statistical model is not well fitted to the expected outcome. It might also mean that the relationship between variables is not linear. In some cases, it is obvious that the relationship between variables might be *U*-shaped. If the relationship between the model and the data is not linear, researchers usually tend to use derivatives of R^2, for example, nonlinear regression.

What R^2 Does Not Indicate

Although R^2 is an intuitive statistical measure, there are some caveats to keep in mind while planning the research. First of all, the R^2 value does not estimate whether the statistical analysis a researcher chooses is the right one to answer his or her research question or whether the researcher chose the correct type of regression analysis. Second, it does not indicate whether the IV is the cause of DV's fluctuations (it does not tell whether there is a cause-and-effect relationship between variables). Third, it does not indicate if the researcher chose the best set of IVs to answer the research question.

Maciej Taraday and Anna Wieczorek-Taraday

See also Correlation; Multiple Linear Regression

Further Readings

Andrew, G., & Jennifer, H. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press.

Blanden, J., Gregg, P., & Macmillan, L. (2007). Accounting for intergenerational income persistence: Noncognitive skills, ability and education. The Economic Journal, 117(519), C43–C60.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250.

Chatterjee, S., & Hadi, A. S. (2015). Regression analysis by example. New York, NY: Wiley.

Cohen, L., Manion, L., & Morrison, K. (2013). Research methods in education. New York, NY: Routledge.

Draper, N. R., & Smith, H. (2014). Applied regression analysis. New York, NY:

Wiley.

Faraway, J. J. (2016). Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models (Vol. 124). London, UK: CRC Press.

Lleras-Muney, A., & Lichtenberg, F. R. (2002). The effect of education on medical technology adoption: Are the more educated more likely to use new drugs (No. w9185). Cambridge, MA: National Bureau of Economic Research.

Miaou, S. P. (1996). Measuring the goodness-of-fit of accident prediction models (No. FHWA-RD-96-040). McLean, VA: Federal Highway Administration.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). Introduction to linear regression analysis. New York, NY: Wiley.

Morgan, S., Reichert, T., & Harrison, T. R. (2016). From numbers to words: Reporting statistical results for the social sciences. New York, NY: Routledge.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction. Wadsworth

Rights, J. D., & Sterba, S. K. (2017). A framework of *R*-squared measures for single-level and multilevel regression mixture models. Psychological Methods. (March 16).

Jeanette Joyce Jeanette Joyce Joyce, Jeanette

Kevin Crouse Kevin Crouse Crouse, Kevin

Race to the Top Race to the top

1363

1365

Race to the Top

Race to the Top (abbreviated as R2T, RTT, or RTTT) was a competitive grant from the U.S. Department of Education (ED) to states and districts willing to make substantial changes to their educational policy in an effort to improve outcomes. This entry describes the key elements of the initial state-level R2T program and also two subsequent programs involving school districts and early childhood education. It concludes by discussing the outcomes of R2T.

Over $4.35 billion were allocated for R2T from the American Recovery and Reinvestment Act in 2009. To date, R2T comprises three separate programs: the initial state competition, a district competition, and a subsequent state competition to improve early childhood education.

The R2T initiatives were unique in that they were competitive grants given to state and local educational systems to implement federal educational priorities. As such, they have also been quite controversial. The initiatives brought to the forefront discussions about how teachers should be evaluated and how student-level data should be incorporated into school systems as well as the role the federal government should play in outlining educational priorities for states and districts.

R2T State Competition

The largest allocation of R2T funding, about $4.15 billion, went to the primary state competition. Applications were accepted in three rounds between March 2010 and December 2011. All states, Washington, DC, and Puerto Rico were eligible to submit plans to implement four primary initiatives over the course of 4 years: (1) adopt standards and assessments to prepare students for college and career success, (2) build data systems to measure student outcomes, (3) design and implement new educator evaluation systems, and (4) create plans to turn around the lowest performing schools. Over the course of three rounds, 18 states and Washington, DC, were selected to receive funding. States were required to distribute 50% of the grant to districts that receive Title I funding; the remaining 50% could be distributed to other districts or could further supplement programs in Title I districts.

In addition to the four areas of focus, there were several other requirements that ranged from policies that must be in place at the state level to providing evidence that the states had already prioritized education by increasing the annual budget. The remainder of this section describes these requirements. States also needed to have policies to provide equitable funding to school districts with high poverty or other high needs and to allow and support charter school systems without capping their growth.

Common Standards and Assessments

States were required to participate in and adopt a set of standards developed by a consortium of states as well as develop a common set of assessments aligned to those standards. Although R2T did not explicitly require the implementation of the Common Core State Standards, its application was considered to be a key element in the selection of states to receive the awards. ED provided a separate application process and funding for state consortia to develop common assessments that met this description. Two consortia emerged from this process. The Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers. As of 2016, assessments developed by these two consortia were beginning to be used for state achievement testing in math and English-language arts in Grades 3–11.

Data System

R2T required states to implement longitudinal data systems that would provide significant information about all students: enrollment and transfer information, demographic data, assessment data, records of teachers and classes taken, courses completed and grades received, test scores on college readiness exams, and information on whether the student enrolled in a postsecondary institution. It and information on whether the student enrolled in a postsecondary institution. It was implied that having such a system would lead to widespread data-based decision making at each level of the administrative hierarchy.

Educator Evaluation System

The most often discussed R2T demand was the call for new teacher evaluation systems. The new evaluation systems were required to use multiple, research-based measures that included assessments of educator practice and also measures of student growth as calculated by the difference between assessments administered at two points in time. The state must classify educators into no fewer than three rating categories, although most R2T states have at least four. States had to require districts to use these ratings to make personnel decisions that included professional development, compensation, dismissal of low-performing educators, rewarding of high-performing educators, and decisions surrounding tenure.

Although terminology differs across the states, most R2T states developed similar progressions of rating levels: The bottom level often had significant consequences that included loss of tenure, pay freezes, or outright dismissal; this was followed by a level in which progress would be closely monitored; and similar consequences would ensue if improvement was not evident within a set period. The third level was typically the expected level of performance for teachers, and the fourth level was to be grounds for promotion, extra pay, or other forms of recognition. In practice, however, many states that adopted this model found most of their teachers rated at the expected or highest level year after year.

Additional educator effectiveness criteria were required. States needed to develop a plan to use measures of student growth to assess in-state teacher preparation programs and professional development providers. They also needed to allow alternative routes to teacher certification. Further, the program required states to develop plans to assess areas of teacher and principal shortages and ensure high-performing teachers were equally distributed to high-poverty and high-minority schools. Further, plans were required to turn around persistently low-achieving schools using one of the four intervention models specified.

R2T District Competition

The second program, the R2T District competition, provided awards in two rounds during 2012 and 2013 designed to support local agencies for 4 years. Local school districts, charter school administrative units, and consortia of school systems submitted plans and were assessed on their (a) vision for reform and improvement, (b) demonstration of past success, (c) plan to prepare students for college and careers, (d) school practices and infrastructure, (e) assessment of current achievement and improvement progress, and (f) the sustainability of their budget and project goals. In total, 21 districts, charter schools, and school consortia were awarded nearly $500 million in two rounds of funding.

R2T Early Learning Challenge

In the R2T Early Learning Challenge, states submitted 4-year plans to implement changes to improve the preparation of young children with high needs to enter the K–12 school system. Applications were evaluated on (a) assessing early learning programs in a tiered rating scale, (b) assessing children’s learning at the time of school entry, (c) creating approaches for sustaining early learning outcomes through early elementary school, and (d) addressing the needs of children in rural areas. Over the course of three rounds, 20 states were selected to receive over $1 billion to implement these plans within 3 years.

Conclusion of R2T

Funding for R2T concluded in summer of 2015. The results were considered mixed, with backlash to the increased testing that resulted from the R2T requirements for educator effectiveness systems and for longitudinal data systems to track progress toward educational goals. Another widely criticized element was the potential consequences for teachers and schools. ED officials stressed the positive outcomes of increased college enrollment and beneficial new professional development programs as evidence of success.

In December 2015, the Obama administration signed bipartisan legislation called the Every Student Succeeds Act that replaced the No Child Left Behind Act of 2001. Although Every Student Succeeds Act continued to require states to have testing in Grades 3 through 8 and some form of accountability system that must be approved by ED, the act was largely seen to return decision making to states and make it more difficult for ED to exert control over state educational policy. ED was explicitly prohibited from requiring states to adopt a common set of standards (such as the Common Core) or to incentivize its adoption through competitive grants such as R2T.

Jeanette Joyce and Kevin Crouse

See also Accountability; Common Core State Standards; Every Student Succeeds Act; No Child Left Behind Act; Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium; Standardized Tests

Further Readings

Boser, U. (2012, March 26). Race to the top: What have we learned from the states so far? A state-by-state evaluation of race to the top performance. Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/reports/2012/03/26/11220/ to-the-top-what-have-we-learned-from-the-states-so-far/

Hallgren, K., James-Burdumy, S., & Perez-Johnson, I. (2014, April). NCEE evaluation brief: State requirements for teacher evaluation policies promoted by race to the top. Retrieved from http://ies.ed.gov/ncee/pubs/20144016/pdf/20144016.pdf

U.S. Department of Education. (2010, January). Appendix B: Scoring rubric. Retrieved January 14, 2017, from http://www2.ed.gov/programs/racetothetop/scoringrubric.pdf

U.S. Department of Education. (2011, November 8). Race to the top fund. Retrieved January 14, 2017, from http://www2.ed.gov/programs/racetothetop/phase1-resources.html

Richard D. Harvey Richard D. Harvey Harvey, Richard D.

Jessica P. Harvey Jessica P. Harvey Harvey, Jessica P.

Random Assignment Random assignment

Random Assignment

Random assignment is a technique for assigning participants to experimental conditions and a prerequisite of true experimental designs. It requires the use of randomization methods to place participants of a particular study into experimental conditions (e.g., treatment vs. control). It ensures that each participant has an equal chance of being placed into either of the experimental groups.

Systemic differences at the outset of an experiment can hurt internal validity—the degree to which effects of the experiment can be attributed solely to the experimental treatment. The random assignment of study members into groups is a requirement for alleviating any initial systemic differences between experimental groups. However, random assignment alone is not a guarantee that there won’t be any initial differences between groups, but rather that any initial differences won’t be systemic.

Random assignment is commonly confused and used interchangeably with random selection. However, the terms denote different foci. Random assignment refers to the method by which study participants are randomly assigned to experimental conditions. By comparison, random selection refers to the method by which the sample is selected from the population for inclusion in a particular study.

Although random assignment is a necessary component of an experimental design, random selection can be used with any research design. However, for an experiment, random assignment would typically follow after random selection has occurred.

There are two distinct forms of random assignment: simple and matched. Simple random assignment ensures that the participants are independently assigned to an experimental condition. Although simple random assignment improves internal validity, the experiment may be vulnerable to extraneous variables (i.e., individual differences). Matched random assignment controls for individual differences by pairing participants in “sets” based on a shared characteristic and subsequently assigning them to different experimental conditions. If an experiment has multiple conditions or the sample size is relatively small, participants can be allocated in “blocks” to ensure equal sample size distribution (block randomization).

Simple random assignment could be achieved using a computerized randomizer or manual technique (e.g., flipping a coin). However, in the cases of small samples or possible confounding variables, a researcher would use the block or matched random assignment. For example, if a researcher wanted to examine the effect of a new curriculum on academic performance, participants could be paired into sets based on GPA and then randomly assigned to the experimental conditions. By pairing the sets based on GPA, the researchers can avoid an unequal distribution of skill in either condition.

Richard D. Harvey and Jessica P. Harvey

See also Experimental Designs; Generalizability; Random Selection; Threats to Research Validity; Validity; Validity Generalization

Further Readings

Christensen, L. (2012). Types of designs using random assignment. In H. M. Cooper, P. M. Camic, D. Long, A. T. Panter, D. Rindskopf, & K. Sher (Eds.), APA handbook of research methods in psychology. Quantitative, qualitative, neuropsychological, and biological (Vol. 2, pp. 469–489). Washington, DC: American Psychological Association.

Jackson, S. L. (2015). Research methods and statistics: A critical thinking approach (5th ed.). Boston, MA: Cengage Learning.

Random Error

Random error

1366

1366

Random Error

See Parameter Random Error; Residuals

Richard D. Harvey Richard D. Harvey Harvey, Richard D.

Falak Saffaf Falak Saffaf Saffaf, Falak

Random Selection

Random selection

1366

1367

Random Selection

Random selection (sometimes called random sampling) is the way in which a particular subset of a population (i.e., sample) is chosen. For selection to be random, two criteria must be met: (1) the members of a population have an equal statistical probability of being selected for the sample, and (2) the probability of being selected is independent of whether any other member has been selected. This entry compares random selection to nonrandom selection, distinguishes random selection from random assignment, describes several types of random selection, and finally provides basic examples of random selection.

Random selection is a crucial part of most research designs. It determines the participants of the study, which in turn provides the data the researcher uses to draw conclusions. Appropriate use of random selection lays the foundation for a strong research study.

Nonrandom sampling introduces potential biases by failing to ensure the equal statistical probability of being selected for a sample. Random selection is essential for ensuring that a sample adequately represents the population to which the researcher intends to infer (i.e., generalizability). Thus, all inferential statistics virtually assume that samples have been randomly selected from the target population.

Random selection is commonly confused and used interchangeably with random assignment. However, the terms denote different foci. Random selection refers to the method by which the sample is selected from the population for inclusion in a particular study. In comparison, random assignment refers to the method by which study participants are randomly assigned to experimental conditions (e.g., treatment vs. control). Thus, random selection would typically precede random assignment for an experimental study.

There are multiple types of random selection, such as simple random sampling (SRS), systematic sampling, stratified sampling, and cluster sampling. In SRS, each individual has an equal chance of being selected. In systematic sampling, every *k*th individual is chosen, with *k* being the population size (*N*) divided by the sample size (*n*). In stratified sampling, the population is divided into strata (i.e., groups) by the researcher. Within these strata, SRS is used to select the sample. In cluster sampling, the population is divided into natural groups which are preexisting. Within these natural groups, SRS is used to select the sample.

The type of random selection used is determined by the researcher, often based upon the researcher's level of access to the population of interest. For example, if a researcher has access to all members of a population (e.g., employees at company X), then the researcher may utilize SRS. However, if the researcher finds it important that desk workers and sales representatives at company X are represented proportionately, the researcher may employ stratified sampling.

Richard D. Harvey and Falak Saffaf

See also Cluster Sampling; Generalizability; Inferential Statistics; Random Assignment; Stratified Random Sampling; Survey Methods; Systematic Sampling; Validity; Validity Generalization

Further Readings

Christensen, L. (2012). Types of designs using random assignment. In H. M. Cooper, P. M. Camic, D. Long, A. T. Panter, D. Rindskopf, & K. Sher (Eds.), APA handbook of research methods in psychology. Quantitative, qualitative, neuropsychological, and biological (Vol. 2, pp. 469–489). Washington, DC: American Psychological Association.

Jackson, S. L. (2015). Research methods and statistics: A critical thinking approach (5th ed.). Boston, MA: Cengage Learning.

Ellen Hazelkorn Ellen Hazelkorn Hazelkorn, Ellen

Rankings

Rankings

1367

1370

Rankings

Rankings compare different items or activities. They are frequently applied to restaurants, to hospitals, and to sports. More recently, rankings are being applied to all facets of economic performance and innovation and to scientific-scholarly endeavor. Increasingly, and controversially, rankings are used to compare the educational performance and quality of primary and secondary schools as well as colleges and universities in many parts of the world.

Educational Rankings

The emergence and popularity of educational rankings reflects increasing public and political demands for greater transparency, accountability, and concerns about learning outcomes and future opportunities and employability. For students and parents, rankings purport to inform student choice and provide a signal of what to expect upon graduation in terms of future educational and/or career opportunities. Rankings are also used by governments, employers, and other societal stakeholders as a guide to quality and value-for-money and are used to shape policy, priorities, and resource allocation.

Increasingly, rankings are indicative of social and economic capital. Having highly ranked schools, colleges, and universities within a region or country is seen as a beacon of economic success and competitiveness. In turn, being highly ranked amplifies the elite status and attractiveness of such schools and colleges, thereby creating a virtuous circle of opportunity and benefit for the students, graduates, and faculty but with very different implications for those attending other institutions.

The multiplicity of different types of rankings and formats illustrate the extent to which they are part of a wider trend for increased public disclosure about and comparison of educational performance. Despite ongoing criticism about the appropriateness of their methodologies, rankings are now widely perceived and used as an acceptable measure of educational quality. Although many of the issues pertain to all education rankings, this entry focuses primarily on the rankings of colleges and universities, henceforth referred to as higher education institutions (HEIs).

History of Rankings

The history of educational rankings can be roughly divided into four main phases, each reflecting social and political characteristics of their time:

> *Phase 1* (1900–1950s)—Early rankings, influenced by the eugenics movement, sought to identify educational excellence according to the educational origins of distinguished men;
> *Phase 2* (1950s–)—Commercially based nationally oriented rankings (e.g., *U.S. News and World Report*, *Secondary Schools Reports and Rankings*) responded to growing massification and the significance of education as the mechanism for "human capital," in other words for personal and societal achievement;
> *Phase 3* (1990s–)—Supranational educational rankings (e.g., Programme for International Student Assessment, Trends in International Mathematics and Science Study, Programme for International Assessment of Adult Competencies, Assessment of Learning Outcomes in Higher Education, U-Multirank) reflect growing necessity to assess quality in the context of the internationalization of education and the intensification of globalization; and
> *Phase 4* (2003–)—Global university rankings (e.g., *Academic Ranking of World Universities, Times Higher Education World University Rankings, QS World University Rankings*) reflect increasing internationalization of the academic, student, and graduate labor markets, and the significance of higher education to the national competitiveness in the global knowledge economy.

International organizations, such as the U.N. Educational, Scientific and Cultural Organization and the Organisation of Economic Co-operation and Development, have been systematically compiling statistical information, and using educational indicators, to measure and assess education performance since the 1960s. *Education at a Glance* was first published by the Organisation of Economic Co-operation and Development in the 1990s followed by international assessments of student learning and adult skills beginning in 2000. Its recent attempts to develop an international measure of higher education learning outcomes has been stalled for the moment. Although these are not strictly rankings, the results are regularly presented in an ordinal format as a way to highlight characteristics of successful education systems and nations.

School rankings are another growing phenomenon in all parts of the world. Parents are keen to send their children to schools which (appear to) offer the best opportunities for their future. This reflects the importance of educational attainment for career, salary, and life chances. It also echoes supply and demand factors associated with entry into a relatively limited number of highly regarded schools and school districts. These rankings often measure entry criteria as well as exam success and admission to specific or highly ranked colleges and universities.

Higher education rankings have become a significant development over recent decades. Today, there are approximately 10 major global rankings and more than 150 national and specialist rankings. The latter includes rankings by field of science (e.g., natural science, mathematics, engineering, computer science, social sciences), discipline or profession (e.g., business, law, medicine, graduate schools), world region (e.g., Asia, Latin America, Arab region, India), institutional age (e.g., HEIs under 50), or specialist (e.g., “green” credentials, commitment to city or region).

There have also been attempts to evaluate the quality and performance of the educational systemas-a-whole using information about resources, policy environment, connectivity, and output. Most rankings are produced by private commercial organizations, with only a few developed by governments.

Methodology

Rankings compare educational institutions using indicators to measure different aspects. Each indicator is assigned a weight or percentage of the total score. The final score is aggregated to a single number and used to create an ordinal ranking or league table, wherein HEIs with the lowest score (e.g., first or second place) are considered the “best.” This methodology exaggerates differences between different scores even though the differences are statistically insignificant.

Of the world’s estimated 18,000 HEIs, rankings measure only about 500 or less than 3%. Use of the term *league table* for education rankings highlights the growing competitive and marketized environment, one in which hierarchy, status, and reputation are increasingly held in highest regard as differentiators of life chances and opportunity.

Different types of rankings—global, national, or specialist—use data from different data sources. Because of the difficulties associated with making comparisons, especially internationally, rankings tend to rely on what can be (easily) measured rather than that which might be most appropriate or meaningful.

National rankings have access to a wide range of data sources from across the national system; problems with data definition are less problematic in a national context. In contrast, global rankings are limited by what they can measure, and the way in which different societies define and measure students; especially international students or faculty can vary considerably. This is one of the reasons that global rankings focus predominantly on research. However, even here there are difficulties. For example, by using publication and citation data from *Web of Science* or *Scopus*, there can be an overemphasis on bio-and medical sciences research that publishes frequently. Student and faculty characteristics (e.g., student entry scores, student/faculty, and doctoral student/faculty ratio) and internationalization are used as proxies for educational quality. Rankings also survey peers, employers, and students to calculate reputation despite the fact that such methodologies are prone to rater bias, and the response rate is, geographically by world region, uneven. Weightings for research and research-related factors can constitute more than 70% of the major global rankings while reputation can be as high as 50%.

In contrast, rankings do not measure teaching and learning, the student experience, or “added value”—what an HEI contributes beyond the student’s preentry characteristics. They also benefit colleges and universities that recruit high-achieving students who progress in a timely fashion despite the fact that nontraditional and mature students represent a growing category of students, especially in developed countries. Arts, humanities, and social sciences research receive less coverage because they publish in a wider range of outlets, which are not well covered by international databases. Rankings also do not include not well covered by international databases. Rankings also do not include indicators for regional or civic engagement, which is an important mission of higher education.

Influence of Higher Education Rankings

A considerable body of evidence has emerged about the impact and influence that rankings are having on higher education. Because being highly ranked appears to bring benefits, rankings underpin decision making at the institutional and national level, and are often used to set explicit strategic goals and measure performance and reward success. There is evidence of a close correlation between highly ranked HEIs and their ability to attract more international students, more sponsorship, and more private giving. Likewise, their respective countries appear more attractive to international mobile capital, business, and talent.

In turn, HEIs are becoming more strategic, reorganizing organizational structures and procedures, allocating resources to fields of study and research that are internationally competitive, and reengineering student recruitment. High-achieving and international students use rankings to inform their choice of institution and/or program. Other HEIs use rankings to identify potential partners or membership of international networks—and vice versa. Employers and other stakeholders may use rankings for recruitment or publicity purposes.

Governments are increasingly influenced by rankings. Many governments are restructuring their national systems and priorities with the aim of creating “world class” or flagship universities that can score highly in global rankings. The emphasis on research has led to changes in research practice, privileging research over teaching, and postgraduate over undergraduate programs, with implications for the academic profession. Rankings are also used to classify universities and allocate funding and as criteria for collaborative agreements, scholarship programs, and immigration laws.

Advantages and Disadvantages

Rankings are influential because they provide a simple, quick, and easy way to compare performance and quality, nationally and internationally. They act as an accountability tool for societies, schools, and HEIs with a weak or immature quality assurance culture or practices. By focusing on quality and performance, rankings have become an important source of information for students and parents, and for strategic decision making, as well as a driver of global positioning and branding of HEIs as well as countries.

Rankings are controversial because there is no internationally agreed definition of quality. Educational quality is complex and not easily reduced to quantification. Ultimately, the choice of indicators and weightings reflect the priorities or value judgments of each ranking. There is no such thing as an objective ranking. Schools, colleges, and universities are complex organizations catering to an increasingly diverse set of students and meeting a wide range of societal needs in different parts of the world. However, rankings usually measure and compare “whole institutions” irrespective of context, student cohort, or institutional mission. Many of the indicators measure socioeconomic advantage and privilege the most resource-intensive schools or HEIs and their students. In doing so, prestige and reputation become dominant drivers of education rather than pursuance of quality, equality, or diversity. Such weaknesses can lead to simplistic comparisons that encourage perverse behavior and poor judgment and policy making.

Rankings are part of a trend for greater transparency, accountability, and comparability at national and global level. Alternative rankings and alternatives to rankings are being developed by governments, nongovernmental organizations, and commercial organizations in response to this demand and problems identified herein.

Ellen Hazelkorn

See also Ordinal-Level Measurement

Further Readings

Hazelkorn, E. (2015). Rankings and the reshaping of higher education: The battle for world class excellence (2nd ed.). Basingstoke, UK: Palgrave Macmillan.

Hazelkorn, E. (Ed.). (2016). Global rankings and the geopolitics of higher education: Understanding the influence and impact of rankings on higher education, policy and society. New York, NY: Routledge.

Norrie, K. (Ed.). (2013). Measuring the value of a postsecondary education. Queen’s Policy Studies Series. Montreal, Canada: McGill-Queen’s University Press.

Rauhvargers, A. (2011). Global university rankings and their impact. Reports I and II. Brussels, Belgium: European Universities Association. Retrieved from http://www.eua.be/Libraries/publications-homepage-list/Global_University_Rankings_and_Their_Impact.pdf?sfvrsn=4

Rauhvargers, A. (2013). Global university rankings and their impact. Reports I and II. Brussels, Belgium: European Universities Association. Retrieved from http://www.eua.be/Libraries/publications-homepage-list/EUA_Global_University_Rankings_and_Their_Impact_-_Report_II

Yudkevich, M., Altbach, P. G., & Rumbley, L. C. (Eds.). (2016). The global academic rankings game: Changing institutional policy, practice, and academic life. New York, NY: Routledge.

David Torres Irribarra David Torres Irribarra Irribarra, David Torres

Rasch Model

Rasch model

1370

1374

Rasch Model

The Rasch model is a psychometric model used in the social sciences to analyze categorical response data, usually collected using a content knowledge test or attitudinal questionnaire, in order to assess the extent to which a set of persons has a certain level of an attribute of interest (e.g., mathematical proficiency or level of anxiety) and the extent to which a positive answer to the questions or statements demands a certain level of that attribute (e.g., difficulty of a mathematical question or level of anxiety required in order to agree or endorse a statement). Originally developed in the 1950s by Danish mathematician Georg Rasch for the analysis of dichotomous responses to intelligence tests, the Rasch model in its most basic form states that the probability that a person can correctly answer a test question—or that he or she endorses a given statement—can be modeled as a function of the difference between an effect associated with the person (e.g., a student’s mathematical proficiency or potentially any person property) and a question effect (e.g., the item difficulty or level of agreement demanded by a statement), such that both effects are in the same scale and

$$\text{Probability of person } p \text{ correctly answering item } i = f\left(\text{Person } p \text{ proficiency } - \text{ item } i \text{ difficulty}\right).$$

In other words, the Rasch model considers the performance of a person on a question to be a product of the trade-off between a person effect (oftentimes interpreted as the level of person proficiency) and an item effect (commonly interpreted as the difficulty of the item). The Rasch model is widely applied in the social sciences, particularly in the context of psychometric analysis of educational testing, where it is often considered to be a special case within item response theory (IRT), and, more generally, as a special case of a generalized linear model (GLM) in statistics. Although it can formally be understood as a special case within item response theory or GLMs, the Rasch model was developed by Rasch under a specific set of theoretical commitments regarding the nature of measurement in the social sciences and the requirements that a model must fulfill in order to be used for measurement, which have historically set apart the use and development of the Rasch model and its extensions from the wider psychometric and statistical literature.

Three Mathematical Formulations of the Rasch Model

In its original form, the Rasch model expresses the expected relation between a set of observed dichotomous responses and a set of unobserved person and item effects. This relation can be expressed in multiple ways, casting the relations between persons and items in alternative, but mathematically equivalent, manners. However, it is important to remember that regardless of the formulation, under the Rasch model, persons are solely characterized in terms of their proficiency and items are fully characterized in terms of their difficulty, and that different formulations simply express the relations between these parameters in alternative forms. There are at least three common ways of formulating the Rasch model: by focusing on the probability of answering correctly, the odds of answering correctly, and the log odds or logit of answering correctly. For the remainder of this entry, “answering correctly” will be used as any positive answer to an item, even if in some cases (e.g., attitude questionnaires) there is no “correct” response.

Probability Formulation

Formally, under the Rasch model, the probability that a person *p* correctly answers a question *i* is governed by a logistic function such that

$$\Pr\left(x_{ip} = 1 \mid \theta\right) = \frac{\exp\left(\theta_p - \delta_i\right)}{1 + \exp\left(\theta_p - \delta_i\right)}.$$

The probability of answering correctly is equal to 0.5 when the person’s proficiency is equal to the item difficulty, and the probability will approach 1 as the level of proficiency of the person is increasingly higher; and conversely, the probability of answering correctly will approach 0, as the level of difficulty of the item is increasingly higher than the proficiency of the person.

Odds Formulation

The Rasch model can also be expressed as modeling the odds of correctly answering a question (i.e., the ratio between the probability of an event occurring and the probability of the event not occurring), as was done originally by Rasch, such that odds are a function of the ratio between the exponentiated person proficiency and exponentiated item difficulty:

$$\text{Odds}\left(x_{ip}=1\mid\theta\right)=\frac{\Pr\left(x_{ip}=1\right)}{\Pr\left(x_{ip}=0\right)}=\frac{e^{\theta_p}}{e^{\delta_i}}.$$

Under this formulation, when the magnitude of the person proficiency equals the difficulty of the item, the odds of answering correctly are 1.0, but as the proficiency surpasses the difficulty, the odds become greater than 1, and as the item difficulty surpasses the person proficiency, the odds become smaller than 1.

Log-Odds (Logit) Formulation

Finally, the Rasch model can be expressed as a modeling the log odds (or logit) of answering correctly:

$$\text{logit}\left(x_{ip}=1\mid\theta\right)=\log\left(\frac{\Pr\left(x_{ip}=1\right)}{\Pr\left(x_{ip}=0\right)}\right)=\theta_p-\delta_i.$$

This formulation directly expresses the log odds of answering correctly as a function of the difference between the person proficiency and item difficulty. This formulation makes it easy to see that formally, the Rasch model can be understood as a special case of a GLM. Under these formulations, we model the response outcomes (1 for correctly answering question or endorsing a statement) using a Bernoulli distribution, and we use a logit link to model the distribution parameter (e.g., the probability of a correct answer) as a function of a linear component, namely, the difference between the person proficiency and the item difficulty.

Properties of the Rasch Model

The Rasch model is often considered a restrictive model in comparison with more flexible alternatives in the family of item response models such as the two-parameter logistic and three-parameter logistic models; however, these additional restrictions are the basis for a number of properties unique to the Rasch model and its extensions. It is worth remembering at the outset that these properties hold if and only if the Rasch model holds for the data being analyzed.

Features

A central feature of the Rasch model is the property of *specific objectivity,* a concept introduced by Rasch to emphasize that when measuring the proficiency of a set of persons, the results of that measurement should be independent of the specific questions used to assess them (assuming the questions are all designed to assess the property of interest), and when we assess the difficulty of a set of questions, the results should be independent of the specific persons we use to judge such difficulty.

A second important feature of the Rasch model is the presence of *sufficient statistics* for its person and item parameters. Under the Rasch model, the total number of correct responses for persons on a set of items constitutes a sufficient statistic for the person proficiency parameters, and the total number of correct responses for items among a given set of persons is a sufficient statistic for the item difficulty parameters. Both totals are referred to as “sum scores.” In other words, under the Rasch model, once we know the total sum score for a given person, as well as the set of items to which the person responded, there is no additional information that can be obtained from the specific pattern of responses; it does not matter which questions were answered correctly, only how many of them.

A third important property, related to both the idea of specific objectivity and the presence of sufficient statistics, is *separability of parameters,* which allows for the estimation of person parameters without requiring the estimation of item difficulties and vice versa. The lack of interaction between the person and item parameters permits conditioning out the item parameters to estimate solely the person parameters and vice versa.

Model Fit

Again, it is important to keep in mind that these properties only hold to the extent that the Rasch model adequately fits the data, a case that must be made by empirically examining the correspondence between the observed data and the data patterns expected according to the fitted model. One of the central elements that can be evaluated in terms of fit is the extent to which the items share a common slope or common level of relation to the underlying attribute, which is usually assessed via the information-weighted fit (infit) or outlier-sensitive fit (outfit) statistics.

Assumptions of the Rasch Model

The use of the Rasch model relies on four key assumptions: (1) the generating process of response data has a random component, (2) the generating process can be reasonably modeled by assuming the presence of an underlying quantitative attribute, (3) the observed responses are attributable to a single underlying attribute, and (4) the observed responses are independent conditional on the underlying attribute. The first and second assumptions correspond to fundamental conceptual assumptions about the nature of the process that is being analyzed with the Rasch model.

The first assumption is that the process being modeled does have a random component. The Rasch model is a probabilistic—as opposed to deterministic—model, and the presence of this random component when modeling persons' responses to questions should be conceptually justified.

The second fundamental assumption is that the response process can be reasonably modeled as a function of an underlying quantitative attribute. This assumption is at the basis of the Rasch model's use of a trade-off between a person effect and an item effect (in the form of either a ratio in the odds formulation or a difference in the log-odds formulation), such that we assume that both the persons and the items possess a certain amount, a level, of this underlying quantitative attribute.

The third and fourth assumptions are of a more operational nature and can be potentially relaxed by the use of extensions to the basic Rasch model.

The third assumption, usually known as the assumption of unidimensionality, indicates that all the correlations that can be observed in the response data are the product of a single underlying quantitative attribute. In other words, the Rasch model assumes that we are measuring one, and only one, attribute. For example, in a mathematics test, we expect that correctly answering a question is solely governed by mathematical knowledge and not, for instance, by the reading proficiency of the respondents.

The fourth and last assumption is the assumption of conditional independence, which indicates that the observed data (e.g., the responses to the items in a test) will show independence conditional on the underlying attribute. Specifically, the assumption of conditional independence states that all the observed correlations in the data are attributable to the underlying attribute, such that conditioning on it will remove all correlations in the observed data.

The assumptions of unidimensionality and conditional independence are related, but they are not the same. If a test is indeed unidimensional, it will also fulfill conditional independence, but the reverse is not necessarily true. Both assumptions can potentially be relaxed, either by the use of multidimensional extensions to the Rasch model or by the explicit inclusion in the models of, for instance, additional dependences among items.

Extensions to the Rasch Model

Although in its most basic form the Rasch model only deals with dichotomous data for unidimensional tests, extensions have been developed that allow the application of the key ideas of the Rasch model to a wider range of data and assessment contexts. The first set of extensions that were developed included the development of models for dealing with more than two possible response categories, including most prominently Andrich’s rating scale model and Masters’s partial credit model. In addition, models have been developed that focus on the decomposition of the item difficulty in terms of item features, with Fischer’s linear logistic test model being most prominent among them. Yet another important area of extension to the Rasch model is that of multidimensional models, for instance, under the framework of Adams, Wilson, and Wang's multidimensional random coefficient multinomial logit model. Finally, an entirely different class of extensions that have been added to the Rasch family of models since the 1990s includes the combination of the Rasch model with mixture models, allowing the application of Rasch models to heterogeneous populations that can be analyzed by examining person membership in unobserved latent classes, as in Rost's latent class Rasch model or Mislevy and Verhelst's mixture linear logistic test model.

Estimating the Rasch Model

As of the 2010s, there is a wide variety of off-the-shelf dedicated software that can be used to estimate the Rasch model and its extensions, including jMetrik, Winsteps, and ConQuest. However, considering that the Rasch model and its extensions can be viewed within larger statistical modeling frameworks such as GLMs, it is nowadays possible to estimate the Rasch model and its extensions using general purpose statistical analysis software, such as Mplus and LatentGOLD, as well as general purpose statistical computing languages such as Stata, SAS, and R.

David Torres Irribarra

See also Generalized Linear Mixed Models; Item Response Theory

Further Readings

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1–23.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561–573.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. Journal of Applied Measurement, 4(1), 87–100.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359–374.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrika, 55(2), 195–215.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press. (Original work published 1980) Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Theory of Probability (Vol. IV, pp. 321–333). Berkeley: University of California Press.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271–282.

Rost, J. (2001). The growing family of Rasch models. In Essays on item response theory (pp. 25–42). New York, NY: Springer.

von Davier, M., Rost, J., & Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In Multivariate and mixture distribution Rasch models (pp. 1–12). New York, NY: Springer.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum.

Ian Katz Ian Katz Katz, Ian

Cort W. Rudolph Cort W. Rudolph Rudolph, Cort W.

Rating Scales Rating scales

1374

1377

Rating Scales

The term *rating scale* refers to a closed-ended response format in which individuals provide reactions to a set of statements or questions that are guided by predetermined anchors. Generally, rating scales are used in survey research to capture information from a sample drawn from a larger population. Such a sample of individuals may be asked, for example, to complete one or more rating scales to capture more-or-less subjective quantitative data (e.g., opinions, emotions, political affiliation). Rating scales are used in educational and psychological measurements to collect empirical data to aid in estimating population parameters likely corresponding to a well-defined research question or hypothesis. Instead of providing participants with an open-ended, qualitative format to evaluate a construct, rating scales allow for a more uniform evaluation of a construct across any number of respondents. This entry discusses various features and forms of rating scales.

Response Values and Dimensionality

Regardless of the nature of the statement or question being asked, all rating scales have a defined set of possible values (i.e., anchors) that can be selected from during the response process; this is to ensure a consistent unit of measurement across all responses. The range and form of possible responses are determined a priori by the information that is intended to be inferred from such scales. From a psychometric standpoint, the preceding notion calls upon concerns regarding sensitivity and level of measurement. At the lowest level of measurement, nominal data call for organizing stimuli into categories based on a common factor (e.g., scaling biologically determined sex as “male” or “female”). The next highest level of measurement is the ordinal level, referring to the ordering stimuli in a series based on increasing or decreasing magnitude. Finally, interval and ratio levels of measurement, which are most commonly called upon in educational and psychological measurement, typically assign a number to a response such that the difference between the items is equivalent. More specifically, the difference between 1 = *strongly disagree* and 2 = *disagree* on the rating scale is equivalent to the difference between 4 = *agree* and 5 = *strongly agree* on the same scale. The distinction between interval and ratio levels of measurement lies in whether there is a true 0 value on the rating scale. Furthermore, higher measurement levels (i.e., interval and ratio) contain more information than lower levels of measurement (i.e., nominal and ordinal).

Scales can be distinguished by their dimensionality. A unidimensional scale comprises one or more items designed to measure a single construct (e.g., a single behavior or attitude). For example, if a researcher were interested in understanding a student’s perceptions of the quality of their instructor, a unidimensional scale might comprise a single item probing about lecture clarity (lecture clarity being the dimension of the respondent’s perception of interest). Moreover, a multiitem unidimensional scale would prompt each respondent to rate different features of the instructor, all related to the perceived lecture clarity, creating a multiitem unidimensional scale.

Researchers may wish to capture multiple dimensions of larger, more complex constructs that require more than one scale, often individually referred to as *subscales*. In the original perceptions of instructor example, the researcher was only interested in perceptions of lecture clarity, meaning that a unidimensional scale may be sufficient to capture meaningful variance in this perception. However, if the researcher were interested in understanding more nuanced factors contributing to respondent’s perception of the instructor (e.g., lecture clarity, availability for feedback, and feedback quality), a multidimensional scale may be more appropriate. The decision to use a unidimensional or multidimensional scale should be based on theory and an understanding of the nature of the construct being assessed.

Forms

Rating scales are constructed in a number of forms. A *Likert-type scale*, one of the most common forms, is composed of one or more statements about a specific variable used to assess respondents' level of agreement or disagreement with each statement on a uniform and symmetric numeric scale. Likert-type scales are unique from other numeric rating scales, such that Likert items assess positive or negative attitudes; Likert-type scales do not contain items that explicitly assess a neutral attitude. For example, if a Likert item states "the instructor speaks clearly in class," the response choices could include 1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, and 5 = *strongly agree*. Respondents would mark which number and corresponding anchor associated with their attitude. Likert-type scales are easy to construct and can be used to measure a variety of phenomena (e.g., attitudes, personality characteristics, perceptions). One disadvantage to using a Likert-type scale is that it creates an opportunity for respondents to demonstrate social desirability or manipulate responses to appear in a positive light.

Another example of a multiitem rating scale is the *Thurstone Scale*. The primary distinction in contrast to Likert-type scales is that Thurstone items can represent an entire range of attitudes from positive, through neutral, to negative. In contrast, the Likert-type scale's items only represent the positive or negative attitudes. To create the Thurstone Scale, judges sort items into 11 categories (i.e., levels) representing the degree of favorability expressed by each item and then rank the items from least to most favorable. Once the scale is completed, respondents are presented with the list of items in random order and asked to mark the items they agree with. Unlike the Likert-type scale, respondents are not given a numeric scale to respond to the items. Instead, they are able to either "agree" by marking an item or "disagree" by leaving it blank. The final score for each respondent is the average of the favorability scores, as previously determined by the judges, of the items marked. The main advantage to using Thurstone Scales is the ease with which respondents understand and respond to the scale. However, given the difficult nature of creating a Thurstone Scale, they are not popular options and are often replaced by Likert-type scales.

Guttman Scales are the third example of a multiitem rating scale. The Guttman Scale is similar to the Likert-type and Thurstone Scales, in that it is developed to assess respondent's attitudes toward a stimulus. Respondents are presented with a set of ordered attitude items of increasing favorability. Again, this is similar to the items used in the Thurstone Scale; however, the items are not presented in a random order in a Guttman Scale. The respondent marks the items in order, starting with the least favorable item, until the respondent does not agree with the item, indicated by leaving blank the first item the respondent disagrees with. The respondent’s score on the scale is the most favorable item marked, indicated by the last item the respondent marks. For example, if there are 10 items presented and the respondent marks Items 1–4, the respondent’s score on the Guttman Scale is a 4. Similar to the Thurstone Scale, it is easy for respondents to complete the Guttman Scale. Yet, the information obtained from the scale is somewhat limited. For example, it is possible that a respondent may agree with Items 1–4 and Item 6 but not Item 5. This person would receive a score of “4”; however, the Guttman Scale is not well equipped to capture this possibility.

In addition, another popular rating scale for assessing attitudes is the *semantic differential*. Most of the discussion thus far has revolved around numbers with accompanying verbal anchors to respond to a series of items, capturing the respondent’s agreement with the items. Semantic differential scales provide a different approach. Participants are shown a concept and are given bipolar adjective pairs (e.g., good and bad, unpleasant and pleasant) with any number of points between the extremes. The respondent marks on the continuum where the concept falls between the two poles of the adjective pair. For example, each item could be presented on a 7-point scale, ranging from −3 to +3, and the scale is scored as the sum of the item scores. It is important to use adjectives that are relevant to the concept under consideration when using a semantic differential scale. Researchers could use a semantic differential scale to gauge students’ attitudes toward feedback they have received from their instructor. Example adjectives for this item could include helpful/unhelpful or clear/unclear.

The discussion thus far has revolved around respondents evaluating their own attitudes about a stimulus, yet an important application for rating scales is for evaluating others’ behavior. One example of such a rating scale is a *graphic rating scale*. Similar to the Likert-type scale, the respondent is prompted with one or more statements regarding an aspect behavior under evaluation; however, instead of items reflecting one’s own attitude, they are prompting evaluation of another individual’s behavior or performance. Graphic scales vary in the amount of information provided by the scale. For example, a graphic scale could be used to evaluate a student’s attention to detail. The respondent would be prompted to evaluate a student’s attention to detail on a continuum from *very poor* to *very good* and shown a line connecting the two extreme options. The respondent would mark the line between *very poor* and *very good* where appropriate, indicating the respondent’s perception of the quality of the student’s attention to detail. Note that where the respondent marks the line is relatively ambiguous and interpretation could vary greatly between different raters. Although graphic rating scales are praised for their simplicity, they tend to lack clear instructions for the respondent, possibly leading to limited interpretation of the responses. Some scholars have overcome this weakness by displaying numeric values at the extremes of the line (1 = *very poor*, 5 = *very good*) or provide descriptions at each number on the line (1 = *very poor*, 2 = *poor*, 3 = *average*, 4 = *good*, 5 = *very good*), increasing the interpretability of the rating scale.

Decreasing ambiguity, *behaviorally anchored rating scales* provide respondents with a similar rating scale as graphic rating scales; however, respondents are provided with specific example behaviors associated with some of the different numeric values on the scale. Continuing with the previous example, raters making evaluations about one's attention to detail would be provided a behaviorally anchored rating scale with numeric values from 1 (*very poor*) to 9 (*very good*). However, the behaviorally anchored rating scale provides example critical incidents or behaviors associated with different numeric values on the scale. Although these behaviors tend to be non specific to the behavior of interest, they serve the respondent as a framework for understanding the scales' values. For example, at Number 3 on the hypothetical scale measuring attention to detail, "keeps track of appointments in an organized manner" is shown, indicating that an example of a student who performs this behavior would score a 3 on the scale. As mentioned, behaviorally anchored rating scales were developed to provide raters with more detail than the graphic rating scale and elicit a more objective response; however, research has shown that this is not necessarily true. One main advantage of using a behaviorally anchored rating scale is that respondents generally accept and understand this response format.

Like behaviorally anchored rating scales, *mixed standard scales* employ behavioral examples but use a nonnumerical response format. The rater would be provided with a set of behaviors that could be displayed by the ratee. Some of the items would reflect positive behavior, neutral behavior, and negative behavior, all related to the dimension of performance under question. For example, "Instructors often forget to provide their students with instructions," "Instructors provide instructions to their students," and "Instructors provide a comprehensive rubric for their students" are examples of negative, neutral, and positive performance, respectively. Next to each provided example behavior, the respondent marks whether the instructor's performance is better than, equal to, or worse than the behavior described. To capture a more objective sense of behavior, a behavioral observation scale may use the same items as the mixed standard scale; however, they may be more specific to the particular dimension of performance intended to be evaluated.

Ian Katz and Cort W. Rudolph

See also Guttman Scaling; Levels of Measurement; Likert Scaling; Thurstone Scaling; Semantic Differential Scaling

Further Readings

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Mason, OH: Cengage Learning.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educational and Psychological Measurement, 31(3), 657–674.

Murphy, K. R., & Cleveland, J. (1995). Understanding performance appraisal: Social, organizational, and goal-based perspectives. Thousand Oaks, CA: SAGE.

Whitley, B. E., & Kite, M. E. (2013). Principles of research in behavioral science (3rd ed.). New York, NY: Routledge.

Michael Tang Michael Tang Tang, Michael

Arunprakash T. Karunanithi Arunprakash T. Karunanithi Karunanithi, Arunprakash T.

Vivian Shyu Vivian Shyu Shyu, Vivian

Raven’s Progressive Matrices Raven’s progressive matrices

1377

1379

Raven’s Progressive Matrices

Raven’s Progressive Matrices (RPM) is a multiple-choice group intelligence test that measures abstract reasoning and fluid intelligence. Intelligence tests can be classified according to whether they are a group of individual texts and whether they are verbal, nonverbal, or mixed, consisting of verbal, numerical, and nonverbal spatial questions. The main individual intelligence tests are the Stanford-Binet Intelligence Scales, the Wechsler Intelligence Scale for Children, and the Wechsler Adult Intelligence Scale, which consist of mixed content. Group-administered intelligence tests such as the Multidimensional Aptitude Battery and the Cognitive Abilities Test are generally used in mass testing situations such as the military and schools and both have a mixed format. RPM is designed for large groups and are completely nonverbal with only spatial questions.

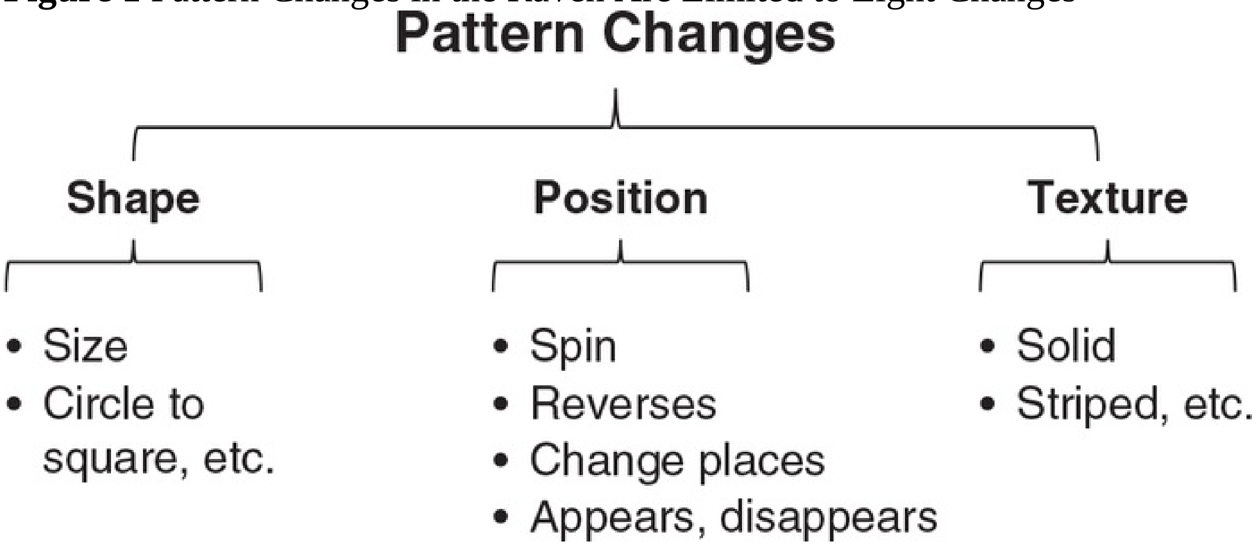
The tests have a long and significant history and are published by Pearson Education, a British-owned publishing and assessment service. Because of the nonverbal group format of RPM, the test has been highly useful and influential in cognitive science research and assessment across cultures globally. This entry describes the test, its structure and scoring, its use in developing the theory of general intelligence or *g* factor, and its impact.

Description

The original test, Raven’s Standard Progressive Matrices (SPM), consists of nonverbal spatial questions presented as a set of abstract black and white matrices with a blank shape or missing piece in each question. The test taker then is instructed to select the correct answer from several possible choices to fill in the blank space much like filling in the last piece to a jigsaw puzzle. Other versions of RPM are Raven’s Coloured Progressive Matrices and Raven’s Advanced Progressive Matrices.

Raven’s Colored Progressive Matrices (Figure 1), designed for children 5 through 11 years, elderly people, and people with mental and physical disabilities, consists of colored matrix questions and answers from RPM to make taking the test and answering its questions easier for the intended participants. Although most of the questions are in color, the last few items in a second set are presented as black and white patterns for further testing of subjects.

Figure 1 Pattern Changes in the Raven Are Limited to Eight Changes



Raven’s SPM, the original Raven test, consists of five sets (A–E) of 12 items each (e.g., A1 through A12), with each item within a set becoming increasingly difficult, hence the name “progressive matrices.” All items are presented in black ink on a white background and in a cyclical pattern, going from easier to the most difficult and then back again to less difficult questions.

Raven’s Advanced Progressive Matrices, also in black and white, contains 48 items in two sets for testing adults and adolescents of above-average intelligence. As with all Raven’s matrices tests, the questions become increasingly difficult as progress is made through each set.

Structure and Scoring

The algorithm behind all three tests involves answering the following two questions, as mapped in Figure 1:

1. What are the main parts of the question's visual structures?
2. How do these patterns change in terms of their shape, position number, and texture?

Figure 1 shows the eight spatial changes that can be detected to select the correct answer for any of three matrices tests. Pattern changes in the RPM are limited to eight changes, as shown in the figure. By considering changes that appear in each of the rows as changes in shape, position, or texture, the test taker can determine the correct answer.

After a test is completed, the total number of correct choices is calculated and scaled based on the number of correct answers divided by the age of the test taker ranging from 6.5 to 16.5 years. For example, if at age 9 a student gets 40 correct answers for a score of 95%, another student at age 13.5 must get 53 correct answers to earn the same 95% score.

γ Factor

RPM is originally designed in the 1930s as a result of John Carlyle Raven's collaboration with Charles Spearman on the *g* factor, which Spearman hypothesized after observing positive correlations in children's academic performance in what appeared to be unrelated subjects. Generally speaking, students who received high scores in math also got high scores in science and English.

From observation to theory, Spearman defined the *g* factor as the summation of all scores that positively correlated among different cognitive tasks found in different intelligence tests. For example, if a psychometrician administered an IQ test that consisted of 5 subtests, A, B, C, D, and E, and found that among the subtests only the scores of subtests B, D, and E correlated positively, then only those scores would be summarized to give the *g* factor score.

Raven, when working with Spearman on this matter, found the Stanford–Binet, an individual testing instrument, to be cumbersome and difficult to interpret. Consequently, Raven decided to design his own instrument to facilitate the research and ended up developing two separate tests. The first of these, RPM, was designed to measure what Spearman called eductive reasoning, which is also referred to as fluid intelligence and may be considered as the ability to make sense out of complexity. The second test, the Mill Hill Vocabulary Scales, was designed to measure what Spearman called reproductive ability, the ability to store and reproduce information. Although the two tests appeared to be measuring two completely different cognitive abilities, the correlation between scores on the Mill Hill Vocabulary Scales and RPM is approximately .75. Raven and Spearman believed this confirmed the existence of the *g* factor.

The measurement of fluid intelligence can be achieved by Raven’s nonverbal test questions that consist of patterns or matrices that are presented as mental algorithms of logical and sequential relations among parts and the wholes to which they belong. These mental processes involve detecting logical changes in the structure, spatial orientation, and the texture of the parts as subsets of the whole.

In addition, RPM was one of the first tests that used the interval scale where test items are sequenced in order of difficulty with the progressive sequence as a basis for normalizing scores as a function of the age. Moreover, to decide which items needed to be modified or rejected to construct such a scale, Raven used a statistical technique that later became known as item response theory.

Impact

RPM was used extensively during World War II, when psychologists and trainers needed to measure a huge number of military recruits including the illiterate and the semiliterate. After conducting validation exercises on Raven’s SPM, the most used test among the RPM suite, researchers produced an abbreviated version that was linearly instead of cyclically scaled. This version’s brevity and its nonverbal presentation, which eliminated literacy bias, became widely used around the world.

In addition to facilitating the measurement of intelligence, the administration of the shorter version of Raven’s SPM provided a huge database on intelligence that other researchers could use for further research. It was by analyzing and interpreting these data that James R. Flynn was able to demonstrate a continued increase in general intelligence scores. Flynn's publications on IQ indicated that both fluid (eductive) and crystallized (reproductive) intelligence significantly and continuously increased from 1930 to around 1980, a phenomenon referred to as the Flynn effect. In 1979, however, Flynn and others suggested that after 50 years of increasing global intelligence, a reverse Flynn effect may be taking place in the Western developed countries.

Flynn and his colleagues are also associated with the Raven tests because their experiments with the tests and the data associated with them have led psychometric researchers to come to different conclusions in regard to the *g* factor and its meaning. Flynn in particular has criticized the standard interpretation of the *g* factor, arguing that Spearman and his followers tend to overemphasize the genetic or inherent aspect of intelligence and underplay the role of the environment in determining mental ability.

Michael Tang, Arunprakash T. Karunanithi, and Vivian Shyu

See also Flynn Effect; *g* Theory of Intelligence; Item Response Theory; Stanford–Binet Intelligence Scales; Wechsler Intelligence Scales

Further Readings

Domino, G., & Domino, M. L. (2006). Psychological testing: An introduction. Cambridge, UK: Cambridge University Press.

Flynn, J. R. (2007). What is intelligence?: Beyond the Flynn effect. Cambridge, UK: Cambridge University Press.

Lewis, D., & Greene, J. (1982). Thinking better. New York, NY: Rawson, Wade.

Mackintosh, N. (2011). IQ and human intelligence. Oxford, UK: Oxford University Press.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. Cognitive Psychology, 41(1), 1–48.

Raven, J. (2008). The Raven progressive matrices tests: Their theoretical basis and measurement model. In Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics (pp. 17–68). Unionville, NY: Royal Fireworks Press.

Raven, J. (2016). John Carlyle Raven (1902–1970) papers: Biographical material. London, UK: Wellcome Library.

Spearman, C. E. (1904). General intelligence, objectively determined and measured. American Journal of Psychology, 15, 201–293.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. Personality and Individual Differences, 39(4), 837–843.

Jana Craig-Hare Jana Craig-Hare Craig-Hare, Jana

Reactive Arrangements Reactive arrangements

1379

1380

Reactive Arrangements

Reactive arrangements are considered a threat to validity within a research design. Elements within an experimental setting could cause subjects to react differently to the experimental arrangements rather than to the experimental variable alone. Reactive arrangements can be a difficult threat to compensate for, as they may not be decreased or eliminated by random assignment. Yet, if reactive arrangements are present, they may lead to confounded findings and opposing explanations.

In defining threats to validity, Thomas D. Cook and Donald T. Campbell, as well as other researchers, attribute these reactions to the use of measures in a study and/or other reactions to the fact that participants are aware that they are in a study. Reactive arrangements are present when these participant reactions become a functional part of the treatment or independent variable in the study. For example, research participants who receive a pretest might be more or less responsive to the experimental variable as a reactive response. This human reaction impacts the study treatment and may produce reactive results.

Reactive arrangements relate to changes in individuals’ responses that can occur as a direct result of participants being aware of their involvement in a research study. For example, the mere presence of observers in a classroom may cause students to behave differently than if the observer was not present, thereby altering the observation findings. Additionally, reactions to the study procedures may occur and cause reactivity. For example, reactivity may be present if participants respond favorably after receiving a nonactive drug within a study (the placebo effect). Increased motivation to please the researcher may cause a participant to perform higher or lower on a skill or achievement measure to accomplish the expected outcome.

One way researchers could protect against the threat of reactive arrangements would be for all control treatments to appear authentic without the subjects knowing the outcome measures or expectations. In this situation, it would also be important for the pretest measures to mask the expected outcomes. Oftentimes a no-treatment approach is utilized through a business as usual group, whereas the business as usual treatment would be considered to be weaker or ineffective to the intervention being studied. Assessing outcomes on a delayed basis would also be a less obvious method to protect against the threat of reactive arrangements.

Complete protection against the threat of reactive arrangements may not always be possible, as participants typically make their own hypotheses regarding the purpose and outcomes of a study. Ethically, researchers should provide ample information about the purpose of a proposed study for potential participants to give their informed consent to participate in the study. Reactive arrangements, however, are most often controlled for through carefully planned research designs and selective measurement variables.

Jana Craig-Hare

See also Ethical Issues in Educational Research; Hawthorne Effect; Placebo Effect

Further Readings

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Diana J. Arya Diana J. Arya Arya, Diana J.

Readability

Readability

1380

1382

Readability

Readability is generally defined by English-speaking literacy scholars as the level of knowledge and skill required to make full sense of a given printed text. This view of readability is most evident in formulas such as the Flesch Ease of Reading that was developed in 1948 based on the assumption that the fewer words in a sentence and the more familiar these words are in a given text, the less difficult it is for readers to comprehend this text. Word familiarity is an indirect yet stable indicator of students' ability to comprehend a word, which in turn has an effect on a reader's comprehension of a given text. Furthermore, the more simple and brief the sentence structures within a text, the greater the ease for readers to understand the intended messages carried within such structures.

The most currently ubiquitous readability indicator is the Flesch-Kincaid, which is essentially a revised version of the original Flesch formula, producing a grade level as its readability score. Originally developed by the U.S. Navy in 1975 to determine the relative difficulty of their various technical manuals, the Flesch-Kincaid formula has become an integral component of many widely used online reading programs and linguistic analytic tools.

Like other readability algorithms, the Flesch-Kincaid determines the word-level familiarity of a printed text by the average *frequency* of individual words (i.e., the likelihood that a reader would be exposed to a particular word based on the analysis of a corpus of books read by adults) and the average sentence length. The lower the average likelihood, or frequency value of words presented in the text, the more difficult the text is deemed for readers. Similarly, the longer the average sentence length, the more assumed difficulty in comprehending key points presented in embedded sentence structures. Simply put, the more frequently a word occurs in a language, the greater the likelihood that students will know its meaning. However, high-frequency words tend to denote more general concepts or categories such as *man* or *work*, rather than more specialized words like *radiologist* or *employment*. Thus, it may be argued that the more *frequent* a word, the greater the likelihood that while students will know its meaning, this meaning may be less precise than what was intended in the text.

Applications of readability or text analytic software for analysis, research, or text development purposes generally follow more qualitative efforts to achieve textual accuracy, coherence, and meaningfulness to readers. Even quantitative programs focused on determining textual cohesion can only do so at a lexical level; that is, the extent to which ideas presented in a text support one another can only be determined by a reader. Thus, while readability indicators offer a general idea about the difficulty of printed texts, such metrics should not be the sole guides for text development.

Readability and Text Quality

Much goes into the development of accessible and considerate texts for readers, particularly within the K–12 context. For instance, a text developer must be mindful of conceptual and linguistic parsimony. Readers should not be overwhelmed by the amount of conceptual information presented in a text, nor should there be too many unfamiliar words or phrases that would inhibit understanding, especially if those words or phrases are not providing critical information. Equally important is the presentation of concepts that foster accurate understandings and avoid potential misconceptions that may inadvertently develop from the use of everyday language to describe concepts. Thus, there is a tension between accuracy and familiarity for readers, which has a direct impact on the relative readability of a given text, and as such, school-based texts must have an optimal balance between these two qualities. Conceptual mapping of textual content can be helpful in clarifying ideas represented in a text, which in turn affects its general readability.

Any account of text difficulty that uses sentence length to establish the readability of texts assumes, at least implicitly, that unpacking the ideas within a single, complex sentence is more difficult for readers than making connections across related propositions stated in separate sentences. As such, a short sentence in itself may be easier to comprehend than a complex one. However, the challenge may come when the reader needs to integrate a cohesive meaning from a series of short sentences, which leads to greater demands on readers to make accurate inferences about how such short sentences connect and support one another in communicating larger ideas. Questions remain about whether the memory burden of complex sentences trumps the inference demands of integrating ideas across separate propositions.

The Multidimensionality of Readability

Readability formulas account for only a few variables that affect the level of difficulty of a text. Such formulas cannot take into account the inherent interest and motivation that readers may have when engaged in a printed text. The greater the interest in or desire for reading a particular text, the greater a reader's capacity for comprehending such a text. Stylistic qualities can also affect the relative ease of comprehending a text. For example, ideas inscribed in first person (i.e., using *I* and *you*, as if the author were having a personal conversation with a reader) tend to be easier for readers than if these ideas are written in passive voice. Narrative structures that follow a familiar pattern of conflict and resolution have been recently found to support comprehension of conceptual information compared to nonnarrative versions of the same content.

Determining the relative difficulty of texts requires both qualitative and quantitative approaches that include the considerations of genre, voice, and topic interest. A formula cannot account for the polysemy of words like *base*, which may at first seem like a generally familiar word (running to first *base*) but may actually be a specialized term (a *base* material use in chemistry).

Diana J. Arya

See also Literacy; Reading Comprehension; Reading Comprehension Assessments

Further Readings

Arya, D. J., Hiebert, E. H., & Pearson, P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. International Electronic Journal of Elementary Education, 4(1), 107.

Arya, D. J., & Maul, A. (2012). The role of the scientific discovery narrative in middle school science education: An experimental study. Journal of Educational Psychology, 104(4), 1022.

Bowey, J. A. (1986). Syntactic awareness in relation to reading skill and ongoing reading comprehension monitoring. Journal of Experimental Child Psychology, 41, 282–299.

Klare, G. R. (1984). Readability. In P. D. Pearson (Ed.), Handbook of reading research: Vol. 1 (pp. 681–744). New York, NY: Longman.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition & Instruction, 14(1), 1–43.

Pearson, P. D., & Camperell, K. (1981). Comprehension of text structures. In J. T. Guthrie (Ed.), Comprehension and teaching: Research reviews (pp. 448–468). Newark, DE: International Reading Association.

Diana J. Arya Diana J. Arya Arya, Diana J.

Jing Yu Jing Yu Yu, Jing

Reading Comprehension

Reading comprehension

1382

1384

Reading Comprehension

Reading comprehension generally refers to the intellectual, socioculturally embedded process of making meaning from printed texts. This meaning-making process involves three important factors: the texts to be interpreted; the readers who engage in interpreting; and the contexts of interpreting a particular text, including the historical background, purposes, cultural values, and the linguistic demands of a particular readership. Since the mid-1960s, each of these factors has been emphasized over others in terms of its relative importance to this meaning-making process. Currently, literacy educators and researchers adopt a more balanced model of reading comprehension, viewing all three factors as equally important for successfully comprehending texts. After providing a historical overview of reading comprehension, this entry discusses reading comprehension in the digital age, including implications for schools and classrooms.

Historical Overview of Reading Comprehension

Prior to the mid-1960s, the reading comprehension process was associated with the notion of digging, as if the meaning needed to be *extracted* from the text. As such, the reader uses various textual features (e.g., contextual clues and displayed images) to locate and *dig out* this meaning. The focus on instruction was on the accuracy and immediacy of recognizing words and their association with one another; if readers have the skills to immediately identify the intended meaning of words and the relationships of these words within and across sentences, then the readers will be successful in their excavation. Thus, the purpose of reading during this time was to gain an accurate interpretation of a given text.

During the 1970s and 1980s, notions of reading comprehension shifted away from the text-focused, digging out metaphor toward a model of a reader-centric process. Meaning is assumed to be constructed within an individual reader's mind, and as such, no two readers can interpret a text in the exact same way. Instructional practices during this time emphasized the importance of utilizing prior knowledge for making inferences from texts. That is, using what is explicitly present in texts, readers were encouraged to draw conclusions about a given topic or concept based on their own prior understanding. Furthermore, instruction began to take into account students' thoughts and feelings about what they are reading rather than solely focusing on learning new information from text.

During the late 1980s through the early 1990s, issues related to context and situation emerged as dominant foci of scholarship on reading comprehension. Literacy research was generally concerned with both cognitive and sociocultural perspectives, and thus reading became known as a social-or community-based practice. With his view that all forms of learning are socially constructed, Russian psychologist Lev Vygotsky was a major influence in this social view of reading comprehension. The champion instructional approach as inscribed in literacy research from this period was for teachers and students to engage in *negotiations* about the intended meaning of a given text. Vygotsky's theory emphasized that language in use is ideological or political in nature; people use language to persuade or affirm cultural values and principles. Relatedly, texts were viewed as nonneutral entities; thus, students were encouraged to problematize and critically interrogate texts.

During the mid-1990s, Allan Luke and Peter Freebody developed the four resources model that increased focus on ideological or critical approaches to reading texts. This practice-oriented model is based on the assumption that readers have different purposes for engaging with a text and thus take on different roles depending on their purpose. According to Luke and Freebody, readers will assume one or more of the four possible roles—the code breaker, meaning maker, text user, or text critic. Each role focuses the spotlight on a particular resource or aspect of the reading activity—the reader, the text, the immediate environment, or the general historical and sociocultural context. The code breakers focused the explicit print features of texts that include the alphabetical letters, the phonological representation of spellings, and the structural conventions and patterns of words within sentences. Code breakers are mainly interested in decoding for the purpose of accurately sounding out words as syntactically represented within a text. Meaning makers are focused on the intended meaning of the text in relation to their background knowledge and past experiences about the general ideas of the text. Text users focus on the pragmatics of reading, using intended messages from text for a variety of purposes; following a recipe, for example, requires a reader to *use* a text for the purpose of making a meal. Finally, the text critic views the text as a source for reflecting on and developing an argument about social, political, or economic issues represented in a given text. Text critics explore potential subtexts to examine the assumption and consequences of notions, events, arguments, and explications presented in text. An example of such critical reading might include a response to a political speech. Freebody and Luke emphasize that students can take more than one role in reading a given text, thus opening classroom discussions to multiple levels of understanding and various interpretations. The four resources model encourages teachers to pay more attention to the *quality* of students' reasoning. That is, producing a line of reasoning (i.e., an evaluative stance or argument based on the ideas represented in text) is preferred over merely offering an accurate summary of a text. Thus, all readers necessarily traverse all four roles, or resources, as each resource is the prerequisite to the ultimate goal of reading comprehension in the form of critical reasoning.

During the late 1990s, Walter Kintsch developed a cognitively oriented model of reading, the construction–integration (C-I) model, which emerged as a dominant paradigm for conceptualizing what is happening inside the mind as a reader engages in textual reading. This model attempts to account for the neurologically based processes and associated complexities of reading in action. According to the C-I model, a reader first decodes small portions of texts that contain embedded ideas or propositions. Almost simultaneously, the reader integrates these propositions to gain a general, text-based understanding, or key idea, of larger portions of the text. This text-based knowledge is then immediately integrated with the reader's prior knowledge about the indicated topic, issue, or concept. This process happens very quickly and is repeated as the reader proceeds through a text. The integrated process as explained by the C-I model assumes a balanced relationship between the text, reader, and various contextual factors. In the United States, the C-I model serves as a guiding reference for large-scale research, pedagogical practices, and policies for classroom-based reading comprehension activities. The Common Core State Standards, the RAND Corporation of educational research and analysis, and the National Assessment of Educational Progress are examples of influential educational entities in the United States that utilize the C-I model as a foundational framework in various empirical and assessment efforts. Moreover, literacy scholars such as P. David Pearson have suggested ways to integrate the cognitive C-I model with the practice-oriented four resources model in order to further clarify best approaches for supporting reading comprehension development for a variety of purposes and texts.

Reading Comprehension in the Digital Age

Recent studies on reading comprehension have explored the notion of text as more than words on a printed page. Images, diagrams, and even simulations have been recognized as forms of text that necessitate various levels of interpretation for a variety of purposes. Scientific tables presenting evidence in support of a particular argument, for example, constitute texts which readers interpret and critically evaluate, leading to various actions such as subsequent investigations. This ability to interpret and analyze scientific texts is a key standard within the practices dimension of the recently developed Next Generation Science Standards. As such, literacy researchers and scholars have expanded notions of text in recent years to include such multiple modes of communication.

Another recent development in reading comprehension research is the study of multiple-text comprehension. There is growing interest in learning more about the skills and strategies needed to triangulate information gathered or constructed from multiple sources of text. For example, a reading assessment task for middle school social studies students may involve triangulating key ideas from a primary source (e.g., the original text of a law written in 1882 to prohibit entry of Chinese immigrants), a secondary account of life during a particular time period (e.g., concerning life for Chinese immigrants during the 19th century), and perhaps key images from this time period (e.g., pictures of workers on the railroad) to construct complex understandings based on such sources of text. Such reading tasks have become more prevalent in district-and statewide assessments.

Implication for Schools and Classrooms

The increased complexities of reading comprehension as briefly described herein have significant implications for classroom practices. Teachers need to have more knowledge about the qualities and complexities of various forms of texts and about optimal ways to identify and use the most appropriate texts for various reading activities and assessments. Collaborative approaches to reading have received more attention in recent years as a way to maximize the value of the funds of knowledge and skills that each student brings to the classroom. The approach called Collaborative Strategic Reading, for example, has been found to be effective for such purposes.

Future research on reading comprehension will likely involve investigations into the particular contextual effects of collaborative reading and text qualities on the comprehension of discipline-and genre-specific texts. For example, Diana J. Arya and colleagues' recent studies suggest that genre and linguistic complexity have a direct impact on an individual's reading comprehension process. Future studies may focus on the mediation potential of collaborative reading on such text-specific effects.

Diana J. Arya and Jing Yu

See also Literacy; Readability; Reading Comprehension Assessments

Further Readings

Arya, D. J., Hiebert, E. H., & Pearson, P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. International Electronic Journal of Elementary Education, 4(1), 107.

Arya, D. J., & Maul, A. (2012). The role of the scientific discovery narrative in middle school science education: An experimental study. Journal of Educational Psychology, 104(4), 10–22.

Boardman, A. G., Klingner, J. K., Buckley, P., Annamma, S., & Lasser, C. J. (2015). The efficacy of collaborative strategic reading in middle school science and social studies classes. Reading and Writing, 28(9), 1257–1283. doi:10.1007/s11145-015-9570-3

Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J. F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. Educational Psychologist, 46(1), 48–70. Retrieved from http://dx.doi.org/10.1080/00461520.2011.538647

Gee, J. P. (2007). Learning to read as a cultural process. Toward defining and improving quality in adult basic education: Issues and challenges, 141–159. doi:10.4324/9780203936740

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge, UK: Cambridge University Press.

Luke, A., & Freebody, P. (1999). A map of possible practices: Further notes on the four resources model. Practically Primary, 4(2), 5–8.

Pearson, P. D., & Cervetti, G. N. (2015). Chapter 1: Fifty years of reading comprehension theory and practice. In P. D. Pearson & E. H. Hiebert (Eds.), Research-based practices for teaching common core literacy (pp. 1–40). New York, NY: Teachers College Press.

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Diana J. Arya Diana J. Arya Arya, Diana J.

Sruthi Swami Sruthi Swami Swami, Sruthi

Valerie Meier Valerie Meier Meier, Valerie

Reading Comprehension Assessments Reading comprehension assessments

1384

1387

Reading Comprehension Assessments

Reading comprehension assessments generally consist of texts with accompanied questions, tasks, or activities designed to inform educators about a student’s abilities, skills, or level of capacity to make meaning from, or comprehend, targeted texts. Reading comprehension assessments typically involve individual textual reading (either silently or aloud) with accompanying questions that are used to gauge a student’s ability to recall explicitly stated information and to understand implied ideas or arguments represented in a text. Such information could be used for making decisions about a student’s educational status and learning goals as well as for identifying best instructional supports. These assessments vary in their utility for providing such information, and some assessments are better than others for informing next steps in instructional practice.

Skills Assessed Within Reading Comprehension Assessments

Reading comprehension has long been viewed as an ability encompassing various subskills and abilities. As such, one or more of the following variables or abilities may be assessed on any given test of reading comprehension: phonological awareness (i.e., awareness of and access to sounds in oral language), graphophonemic knowledge (i.e., knowledge about the sound–print relationship that supports decoding text), lexical automaticity (i.e., ability to immediately read common sight words), reading fluency (i.e., automaticity and immediately read common sight words), reading fluency (i.e., automaticity and prosody of decoding texts), information comprehension (i.e., direct recall of and inference-making related to ideas represented in text), ability to use reading strategies (i.e., specific skills used to clarify meaning in text, such as the use of contextual clues to understand unfamiliar terms), and vocabulary knowledge (i.e., understanding of word meanings). Because of the multidimensional nature of reading comprehension, multiple reading assessments that target the various skills, some of which were just listed, are useful for gaining a composite understanding about students' abilities to make sense of and apply information from multiple kinds of texts.

The targeted skills and task design of reading comprehension assessments have historically aligned with the transformations in beliefs about what constituted reading comprehension. During the mid-1960s, for example, assessments of reading were designed to elicit a student's ability to identify the correct meaning from text. Reading assessments in recent years have focused more on a student's ability to critique and compare texts and provide a logical argument for a particular line of reasoning rather than a *correct* answer.

Classroom-Based Reading Assessments

Reading comprehension assessments are developed for both formative and summative purposes. For classroom teachers, formative reading assessments typically consist of informal reading tasks designed for a particular student or group of students. Such a task could include reading aloud a class-assigned text and summarizing (recalling key information) following the reading. Teachers observe the reading and subsequent summarization in order to gauge the student's fluency (i.e., demonstration of decoding and prosody while reading aloud the text) and sense making of key ideas presented in the text. This information is then used for selecting and/or adapting subsequent texts as well as planning future instruction.

These observational activities reflect the general practices of the Qualitative Reading Inventory, which is an assessment program widely used by literacy specialists and classroom teachers. Literacy expert Yetta Goodman demonstrated the instructional benefits of the retroactive miscue analysis that is often used in conjunction with the Qualitative Reading Inventory. Retroactive miscue analysis involves follow-up questions after the student has read aloud the selected text; these questions help to uncover the students' thinking processes as they paused, self-corrected, or misread a particular word or phrase. Retroactive miscue self-corrected, or misread a particular word or phrase. Retroactive miscue analysis sessions provide teachers the opportunity to learn more about the particular abilities and strategy that students use during reading.

Teachers also use summative assessments, particularly when determining achievement or growth in reading ability at the end of unit or grading period. Such summative assessments tend to be administered to all individuals within a classroom and involve silent reading of texts followed by a standard set of questions in the form of multiple-choice or constructed response (i.e., short answer or essay items). This type of assessment is designed to provide teachers with a general picture of their students’ overall learning of a concept or general level of comprehension of an informational text.

Large-Scale Reading Assessments

Within the United States, reading comprehension has long been identified as a key skill for determining academic achievement and, as such, assessing all students’ abilities to comprehend grade level, academic texts (i.e., informational texts that students are expected to comprehend according to their grade level) has been an explicit imperative in U.S. educational policies. Large-scale reading comprehension assessments are standardized (consistent) forms of texts with accompanying standardized questions that are systematically administered to large groups of students. Although the standardized reading assessments are used for summative purposes, some of these are used for diagnostic, formative purposes. One of the most widely used large-scale, diagnostic reading assessments is the STAR, which is a 20-minute, computer-administered test used by K–12 schools for making placement decisions at the beginning of a school year.

Similarly, summative large-scale instruments such as those used by the Smarter Balanced Assessment Consortium are administered for accountability purposes. Schools within districts for every state must demonstrate their ability to close the achievement gap for all students, and scores from assessments like Smarter Balanced provide information that is used to determine such growth. The high-stakes decision making and accountability efforts associated with such large-scale reading assessments have provoked a great deal of controversy over the usefulness and validity of such instruments for supporting reading growth and achievement. For example, many assessment specialists and scholars raised concerns related to cultural and linguistic bias of texts and associated items or tasks, contending that presently existing large-scale reading comprehension assessments fall short of providing adequate, culturally responsive measures of comprehension ability.

Reading Assessments for an Increasingly Diverse Student Population

The increasing use of high-stakes standardized reading assessments mandated by accountability efforts may be somewhat at odds with the proliferating linguistic and cultural diversity found in schools. These tests have potentially serious consequences for all students, but particularly those from nondominant cultures. Results from standardized reading comprehension assessments are useful to the extent that they provide stakeholders with reliable and valid information concerning students' reading abilities. However, high-stakes standardized tests have frequently been found to fail to provide such information. Such assessments often oversimplify what is now accepted as a complex, multidimensional construct by targeting only a selected subset of skills (e.g., the ability to answer questions about explicit and implicit information in text). Such tests frame reading events as decontextualized activities (i.e., reading comprehension with no contextual purpose) and that ignore the central role of linguistic and cultural variation in shaping readers' responses to texts. In other words, many large-scale, standardized reading assessments continue to resemble a simpler, earlier view of reading comprehension as a discrete task involving little more than identifying the correct answer from a text. Failing to take cultural variation into account potentially compromises the validity of the inferences made about individual students' abilities. Such reductive and culturally unresponsive conceptualizations of reading comprehension could then have negative effects on classroom instruction, leading to both a narrowing of the curriculum and a devaluing of the perspectives of culturally and linguistically diverse students whose interpretations do not conform to expected responses.

Another serious concern is the use of assessments in English for students who are still in the process of learning the English language. For example, if students who speak Spanish as their dominant language are assessed in English, the level of complexity and pragmatic style in the phrasing of comprehension questions may inhibit the students' ability to fully express their ability to make meaning from a text. Words such as *describe* or even phrases such as *look up* that are often displayed in assessment directions require a level of pragmatic knowledge that depends on more than one's ability to understand and respond to the general meanings of a text. Such instances of pragmatic complexity bring into question the validity of such standardized assessments for determining a language learner's reading comprehension. As such, cultural responsiveness rather than the *sameness* of equality is an important consideration when using comprehension assessments for making school-and policy-based decisions that can alter the course of a student's life.

Cognitive Versus Sociocultural Views of Reading Comprehension Assessments

Reading both in and out of school comprises a range of deeply contextualized, socially situated activities in which individuals make meaning from print in different contexts, for different purposes, and as part of different participation structures. This widely accepted view of comprehension has historically been absent from the bulk of reading assessments that are overwhelmingly oriented to an individual, cognitive view of reading and thus provide scant information about individuals in relation to these sociocultural dimensions of reading. There are few instances of assessing students in social contexts; one example is the Collaborative Strategic Reading, a formative, classroom-based assessment used in gauging collaborative comprehension within the program.

Within Collaborative Strategic Reading, students are organized into cooperative groups of four members, each of whom has a specific role in supporting collective understanding of a shared text. The Collaborative Strategic Reading assessment focuses on the identification of key ideas and agreed-upon definitions of unfamiliar words (*clunks*) based on strategies that include the use of morphological knowledge (understanding of the meaning of word parts) and contextual clues. All students take notes in their own *learning log*, which captures shared ideas within the group, summarizing key points, new vocabulary, and a review statement that, according to the group, captures the full idea as the well as the group's stance on this idea. Teachers generally use these completed logs to assess individuals' abilities as well as their respective correspondence to others within and across groups. Such information can be used to identify and adapt texts for future activities as well as crafting mini lessons for strengthening strategy use or summarization skills.

Future developments of reading comprehension assessments may have a greater

Future developments of reading comprehension assessments may have a greater emphasis on collaborative abilities; such developments would align with the importance of peer discussion around a variety of challenging academic texts as inscribed in the Common Core State Standards. The next decade may bring forth new forms of assessment that more closely reflect the sociocultural, multidimensional, and dialogic nature of reading comprehension.

Diana J. Arya, Sruthi Swami, and Valerie Meier

See also Literacy; Readability; Reading Comprehension

Further Readings

Boardman, A. G., Klingner, J. K., Buckley, P., Annamma, S., & Lasser, C. J. (2015). The efficacy of collaborative strategic reading in middle school science and social studies classes. Reading and Writing, 28(9), 1257–1283.

Caccamise, D., Friend, A., Littrell-Baez, M. K., & Kintsch, E. (2015). Constructivist theory as a framework for instruction and assessment of reading comprehension. In S. R. Perry & K. Headley (Eds.), Comprehension instruction: Research-based best practices (pp. 88–102). New York, NY: Guilford Press.

Flores, G. S. (2016). Assessing English language learners: Theory and practice. New York, NY: Routledge.

Goodman, Y. M. (1996). Revaluing readers while readers revalue themselves: Retrospective miscue analysis. The Reading Teacher, 49, 600–609.

Leslie, L., & Caldwell, J. S. (2011). Qualitative reading inventory. Hoboken, NJ: Pearson.

Pearson, P. D., & Cervetti, G. N. (2015). Fifty years of reading comprehension theory and practice. In P. D. Pearson & E. H. Hiebert (Eds.), Research-based practices for teaching common core literacy (pp. 1–40). New York, NY: Teachers College Press.

Mitchell Campbell Mitchell Campbell Campbell, Mitchell

Markus Brauer Markus Brauer Brauer, Markus

Regression Discontinuity Analysis Regression discontinuity analysis

1387

1392

Regression Discontinuity Analysis

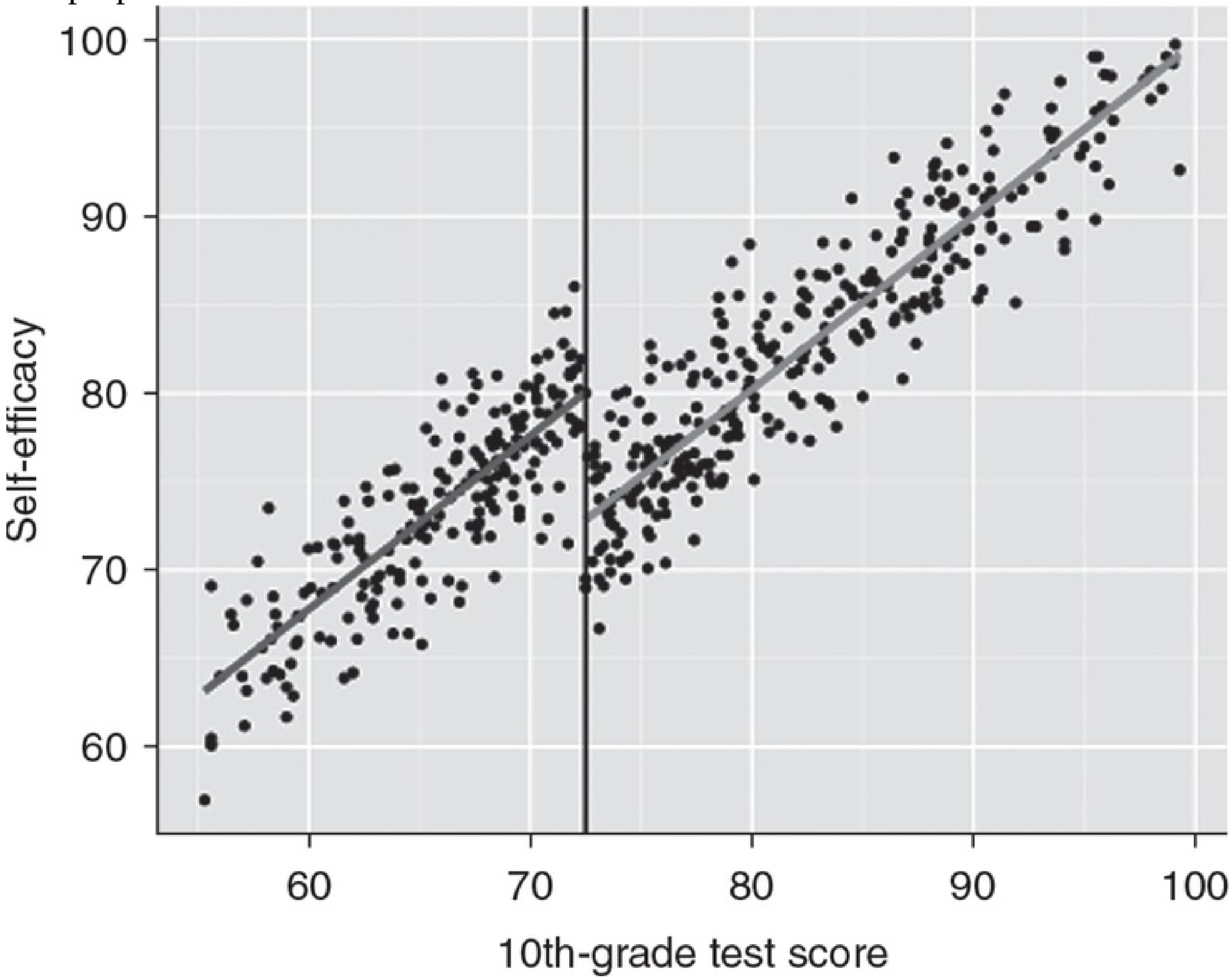
Regression discontinuity analysis is a statistical tool that allows researchers to examine the effectiveness of the treatment in such studies. Consider the following: One researcher wants to determine whether tutoring underachieving middle school students improves their math grades; another wonders whether providing financial aid to low-income students has the desired effects on student success and dropout rates; and a third hopes to assess the effectiveness of special support programs for promising high school athletes. These studies fit the definition of a regression discontinuity design, whereby participants who satisfy a chosen criterion are assigned to a certain treatment and some outcome variable is measured later.

Regression discontinuity analysis is used for studies in which participants are assigned to treatment conditions based on a known assignment rule rather than randomly being assigned to conditions. Researchers or practitioners define an a priori *cutoff point* (Z_0) for participants’ scores on an assignment variable (Z). Participants below the cutoff point receive the treatment, whereas those above the cutoff point do not (or vice versa). Participants are thus divided into groups defined by a dichotomous treatment variable (X). At a later point, the researchers measure the relevant outcome variable (Y). The goal of the regression discontinuity analysis is to determine whether the treatment has the desired effect on the outcome variable.

The Standard Model

To make these ideas more concrete, the following example will run through this text. In this hypothetical study, the assignment variable is a student's score on a standardized test taken in 10th grade, the treatment is whether the student is enrolled in a standardized test prep class, and the focal outcome measure is "self-efficacy," the student's belief in the student's ability to improve the student's standardized test performance. The school provides the test prep class to students scoring in the lowest 30% on the 10th-grade test. The data from this hypothetical example are displayed in Figure 1.

Figure 1 Students' self-efficacy scores increased significantly as a result of the test prep class



One key to understanding this type of analysis is noting that over and above the effect of the treatment variable, there is a relationship between the assignment variable and the outcome variable. In the given example, even if the test prep class is effective, it is quite possible that the students who attended the class will have lower self-efficacy scores on average than those who did not, simply because they started out at a lower level at the outset of the study. The question is whether the students who receive the class will have higher self-efficacy than would be predicted based on their 10th-grade test scores.

A regression discontinuity design is analyzed as follows: The outcome variable (*Y*, self-efficacy) is regressed on the treatment variable (*X*, attending the test prep class or not) and the assignment variable (*Z*, 10th-grade test score). One thus obtains the following regression equation:

$$Y = b_0 + b_1 X + b_2 Z + e.$$

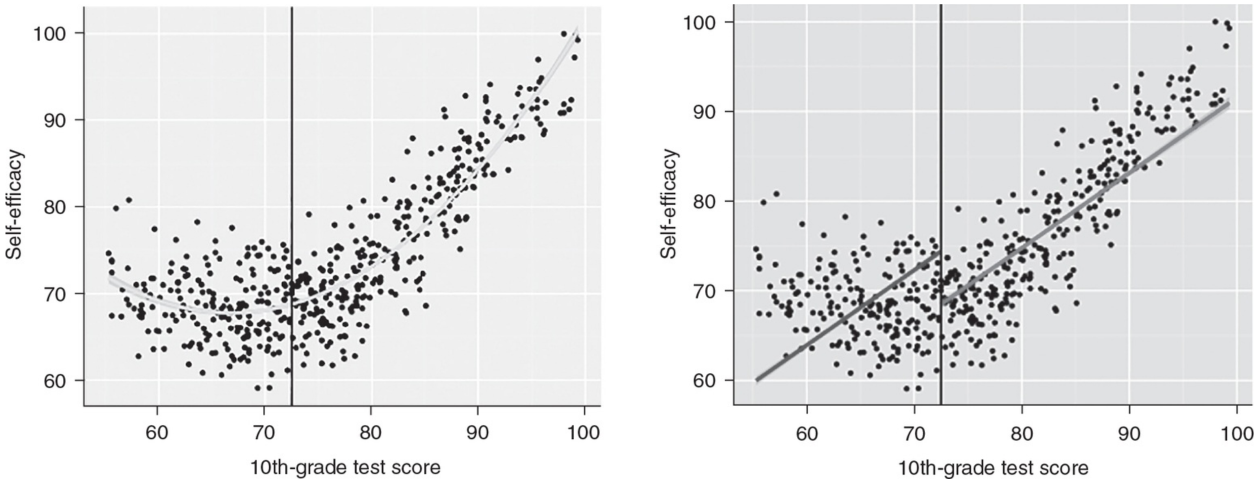
If the coefficient b_1 is statistically significant, the data suggest that the treatment has an effect on the outcome variable. On the graph, this treatment effect will manifest as a vertical discrepancy between the two parallel regression lines. In the given example, the self-efficacy scores of the students in the test prep class were higher than would be expected based on their 10th-grade test scores. Because the treatment is dichotomous, the treatment effect is exactly equal to the coefficient b_1. Students who attended the test prep class had self-efficacy scores 7.5 points higher as a result of taking the class.

Curvilinear Relationships

Aside from other core model assumptions (discussed elsewhere in this volume), linearity is exceptionally important in estimating the treatment effect without bias. If the relationship between the assignment and outcome variables is not linear, the b_1 coefficient will not represent the treatment effect accurately. For example, imagine a data set in which there is no treatment effect and a curvilinear relationship between the assignment variable and the outcome variable (see Figure 2a). The data could be accurately described with the following model:

$$Y = b_0 + b_1 Z + b_2 Z^2 + e.$$

Figure 2 (a) There is a curvilinear relationship between students’ test scores and their self-efficacy scores (b) Parallel lines fit to curvilinear relationship misestimate treatment effect



The treatment has no effect here, so the researchers should find a coefficient of 0 if they add treatment as a third predictor to this model. When the researchers ignore this curvilinear relationship and analyze the data with the model described in Equation 1, it is possible to obtain a significant treatment effect (see Figure 2b). However, this effect is due entirely to the fact that straight regression lines are being fitted to a curved data pattern.

When theory and prior studies suggest that there is a curvilinear relationship between assignment variable and outcome variable, it is advised to add a quadratic term to the model described in Equation 1. The full model would then be

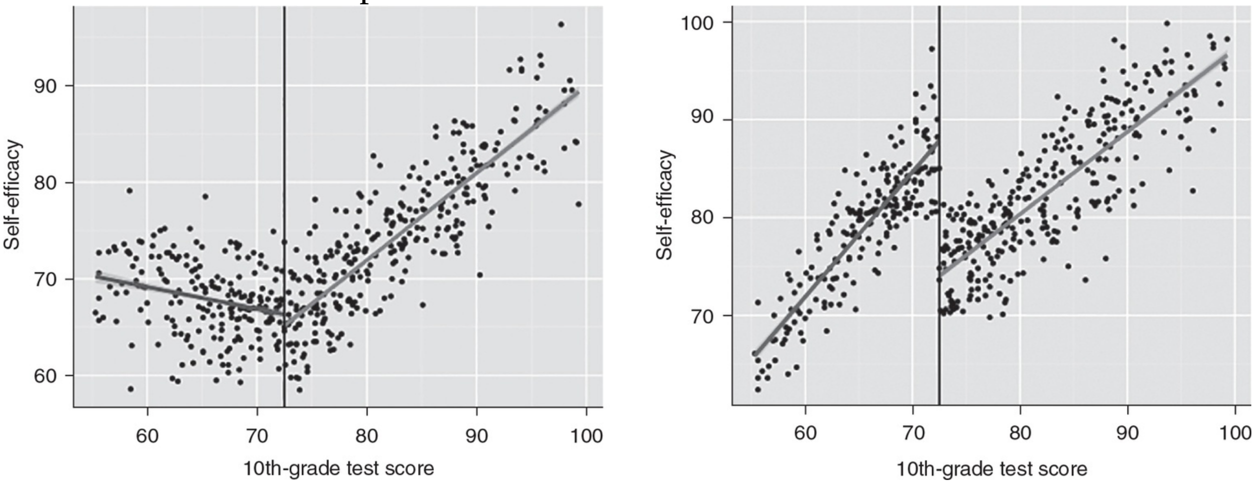
$$Y = b_0 + b_1X + b_2Z + b_3Z^2 + e.$$

Like before, the coefficient b_1 represents the treatment effect over and above the (linear and quadratic) effect of the assignment variable.

Interactions Between Treatment and Outcome

In certain cases, the treatment’s effectiveness depends on individuals’ scores on the assignment variable. Two cases are common: (1) individuals with scores on the assignment variable close to the cutoff point benefit *less* from the treatment (see Figure 3a) and (2) individuals with scores on the assignment variable close to the cutoff point benefit *more* from the treatment (see Figure 3b). Both cases are problematic for the classic regression discontinuity model, which forces the two regression lines representing the model predictions to be parallel. The model is thus misspecified. In addition, certain observations may have large residuals that decrease the statistical power to detect a treatment effect.

Figure 3 (a) Students scoring further from the cutoff point benefit more from the treatment than those who score close to the cutoff point. (b) Students scoring closer to the cutoff point benefit more from the treatment than those who score further from the cutoff point



The solution is to estimate an interactive model in which the *Y*–*Z* relationship is allowed to vary between the treated and the untreated groups. This can be achieved with the following model:

$$Y = b_0 + b_1 X + b_2 Z + b_3 XZ + e.$$

If the coefficient b_3 is statistically significant, the data suggest that the relationship between the assignment variable and the outcome variable is not the same in the two groups. Like in every interactive model, b_1 represents the treatment effect for a participant with a score of 0 on the assignment variable. If the assignment variable in this example were included in its raw form (i.e., uncentered), the coefficient b_1 would estimate the treatment effect for a student with a score of 0 on the 10th-grade test. Clearly, this coefficient, and its associated *F*- and *p* values, would be rather meaningless.

To address this issue, many texts on regression discontinuity analysis suggest centering the assignment variable around the cutoff point by subtracting the value of the cutoff point (here 71) from every student's score on the assignment variable. Now, b_1 represents the treatment effect for a student with a 10th-grade test score of 71. Note that this effect is *not* the average treatment effect, making this a suboptimal approach. This approach will lead researchers to underestimate the average treatment effect when individuals whose scores on the assignment variable are close to the cutoff point benefit comparatively less from the treatment (Figure 3a) and overestimate the average treatment effect when individuals whose scores on the assignment variable are close to the cutoff point benefit comparatively more from the treatment (Figure 3b).

A better data-analytic strategy is to center the assignment variable around the average score in the treatment group (here 65). With this form of centering, the coefficient b_1 will represent the treatment effect for the typical person within the treatment group, accurately reflecting the average treatment effect. Regardless of the type of centering that is done with the assignment variable, the coefficient b_3 indicates whether the relationship between the assignment and outcome variables is different in the treatment and no treatment conditions.

In practice, it is virtually impossible to distinguish the case in which there is a curvilinear relationship between assignment variable and outcome variable and no treatment effect (Figure 2a) and the case in which there is a linear relationship between assignment variable and outcome variable, an average treatment effect, and an assignment variable by treatment interaction caused by the fact that the treatment is less effective for participants with scores close to the cutoff point (Figure 3a). Both the Polynomial Model 2 and the Interactive Model 4 will fit the data quite well. Depending on the spread of the scores on the assignment variable, the researchers may be able to demonstrate that there is no evidence for curvilinearity among individuals in the untreated group. They may attempt to demonstrate that the interactive model fits the data better: showing that it has a smaller sum of squared errors, fewer outliers, and violates fewer model assumptions or by conducting a log-likelihood test showing that the observed results are more likely under the interactive hypothesis than under the curvilinear hypothesis. But researchers should be aware that the two interpretations are hard to distinguish empirically in a given data set, and ultimately they have to use theoretical arguments and refer to prior studies if they end up favoring one interpretation over the other. For example, in the hypothetical example, it makes little sense that individuals with very low scores on the 10th-grade test would score more highly on the self-efficacy measure if the test prep class had no effect. The researchers could argue that the interactive model is more logical than the quadratic model.

Statistical and Practical Considerations

The researchers might imagine that students' scores will also be affected by factors like parents' educational level or motivation to attend college. Covariates like these can simply be added to the regression equation. If a covariate coefficient is significant, it indicates that there is a relationship between the covariate and the outcome over and above the effect of the treatment and the assignment variable.

If the researchers are interested in exploring the effects of the treatment on individuals who are not part of the focal treatment group, they may choose to use a probabilistic assignment rule to decide who receives the treatment. This method contains elements of both a known assignment rule and random assignment: A certain proportion of participants on either side of the cutoff is given the treatment. For example, a researcher may decide that half of the students who scored in the bottom 40% on the test and one sixth of those in the upper 60% are randomly chosen for the test prep class.

Note that it is impossible to use a dichotomous variable as the assignment variable because such a variable will be perfectly confounded with the treatment variable: For example, in a study in which gender is the assignment variable, it would be impossible to say whether the observed differences are driven by treatment or gender.

The conclusion validity of studies with a regression discontinuity design is lower than that of randomized experiments, and a much larger sample size is required to achieve the same level of statistical power for two reasons. First, finding an effect of the treatment on the outcome variable over and above the effect of the assignment variable is difficult given that the treatment and assignment variables are, by definition, highly correlated. When the effects of multiple correlated predictors are estimated, the standard errors of regression coefficients are large, resulting in lower statistical power. Second, in practice, the treatment and comparison groups tend to be very different in size (e.g., people with an IQ over 150, families in the bottom 10% of household income). This imbalance between groups also decreases the power of the analysis.

The conclusions that can be drawn from results of studies with a regression discontinuity design are limited in scope because any treatment effects can only be assumed to hold for the treatment group and often not all alternative explanations can be ruled out. Including relevant covariates can help reveal the true effect of a treatment on the outcome of interest: If the relationship between the test prep class and self-efficacy persists when controlling for parents' the test prep class and self-efficacy persists when controlling for parents' education level, the researchers can have more confidence that the treatment is having the observed effect.

A number of ethical and practical concerns make research utilizing regression discontinuity designs rare or challenging. First, universal application of an assignment rule is difficult and, in some cases, unethical. By setting a cutoff point, a researcher ultimately decides who is deserving of a given treatment. Perhaps a student who scored just above the cutoff point for receiving the test prep class will lose motivation to go to college. Furthermore, when a set cutoff point is made public, it begins to lose its meaning (e.g., people lying on their taxes to qualify for government programs). As a result, the scores on the assignment variable may contain a lot of error, hampering one's ability to reach accurate conclusions. Finally, in many situations where a regression discontinuity design is being used, a randomized experiment would be more effective, depending on the questions the researchers are interested in answering. If the school is hoping to assess whether the test prep class should be made mandatory for all students, a randomized experiment would make more sense than a regression discontinuity design because conclusions from such a study can be assumed to hold for all individuals in the population of interest.

The regression discontinuity design belongs to the family of the so-called quasi-experimental designs. Other designs in this family are the nonequivalent control group design and the interrupted time series design. Like other quasi-experimental designs, the regression discontinuity design has less internal validity and less conclusion validity than a randomized experiment. However, it allows researchers to draw causal conclusions with greater confidence than a post-only correlational design or a simple pretest–posttest design. Yet, some researchers doing field research employ these latter designs when random assignment is not feasible, perhaps unaware that better alternatives, such as the regression discontinuity design, are available to them.

Mitchell Campbell and Markus Brauer

See also Correlation; Multicollinearity; Power Analysis; Quasi-Experimental Design

Further Readings

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design &

analysis issues for field settings. Boston, MA: Houghton Mifflin.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. Psychonomic Bulletin & Review, 11(5), 791–806.

Judd, C. M., & Kenny, D. A. (1981). Estimating the effects of social intervention. Cambridge, UK: Cambridge University Press.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. Annual Review of Psychology, 60, 607–629. doi:10.1146/annurev.psych.60.110707.163544

Maciej Taraday Maciej Taraday Taraday, Maciej

Anna Wieczorek-Taraday Anna Wieczorek-Taraday Wieczorek-Taraday, Anna

Regression Toward the Mean Regression toward the mean

1392

1395

Regression Toward the Mean

Regression toward the mean, or regression to the mean, is a statistical phenomenon that is often observed in student assessment and repeated measurements research in different branches of science. Regression toward the mean is present whenever a construct that is being measured is not accessible directly but is estimated by using methods that are not absolutely reliable. This is the case in the vast majority of measurements in social sciences, education, and students’ assessment. Observations with extreme values in the first measurement will tend to be closer to the mean in the second measurement, and extreme observations in the second measurement will tend to be closer to the mean in the first measurement, whenever two variables are not perfectly correlated.

Regression toward the mean must be seriously considered when designing scientific studies and data analysis to avoid making incorrect inferences. This phenomenon is observed both on a subject level and on a group level. It is caused by random fluctuations in the subjects and by nonrandom sampling of a group from the population.

The observed result of a measurement is the sum of unobserved real value and a random error of measurement. A random error influences single observations but does not affect the mean value of the whole set of observations (assuming that the sample was randomly drawn from the population). If the real value did not change between two measurements, it is expected that mean values for those two measurements stay the same. In each measurement, some observations are below the true value and others are above it. The observed value changes due to random error of measurement.

This entry presents basic information about regression toward the mean. After providing background information, it shows how this phenomenon could influence research at the subject level and at the group level and how to deal with this effect.

Background

The term *regression to the mean* was coined by 19th-century scientist Sir Francis Galton, probably best known for his works on eugenics. Galton observed that extreme height in parents is not passed completely to their offspring and he considered it a genetic phenomenon. Galton called it reversion to the mean or reversion to mediocrity. The difference in height between parents and their child is proportional to the parents' deviation from typical height in the population. Height of offspring shifts toward the mediocre point, which was identified as the mean value of height in the population.

Cognitive psychologist and 2002 Nobel Prize laureate in economics Daniel Kahneman uses regression toward the mean as an explanation of a common belief that rebukes seem to improve performance and praises seem to diminish it. Kahneman illustrates with an example of flight instructors: A flight school had adopted a policy of consistent positive reinforcement recommended by psychology experts, whereby each successfully executed flight maneuver of a cadet was verbally reinforced by flight instructors with a praise. After some experience with this policy, the instructors claimed that positive reinforcement is not optimal for cadets because they tended to make mistakes right after positive reinforcement took place. On the contrary, cadets punished after a bad execution tended to perform better next time. This claim is based on instructors' experience; it does not take into account regression toward the mean phenomenon. It is simply more viable that after a successful maneuver, the next execution will be less successful.

Group Level

Whenever a specific category of people takes part in research, regression toward the mean has to be controlled. Regression toward the mean is observed in groups which are nonrandomly picked from the population, especially in quasi-experimental studies, when individuals are assigned to groups based on their score on the baseline tests. For example, let us assume some testing efficiency of psychotherapy to remove the symptoms of arachnophobia—the fear of spiders and other arachnids. A sample included in such a research consists of people diagnosed with arachnophobia. Those with higher scores on a questionnaire measuring fear of spiders than the majority of the population are assigned to the research group based on the baseline measurement. It is expected that psychotherapy aimed at removing symptoms of the fear of spiders releases people from fear and allows them to function normally in the company of arachnids. Consequently, after the therapy, in the follow-up use of the questionnaire, scores of fear should be lower than in the scores in the first measurement.

The effect of regression toward the mean has to be considered in this situation, as it can be easily confused with reduction of fear caused by the therapy. The more extreme the score of the baseline measurement, the higher the probability that it is a result of a random error at most, so even if the therapy is not working at all, the score in the second measurement will drift toward the mean in population. This might be incorrectly interpreted as an effect of the therapy. To avoid false conclusions (the drop of fear score is an effect of a therapy), a control group—with the same mean level of fear—is introduced to the research. A decrease in the level of fear in the group that is not a subject of the therapy reflects the effect of regression toward the mean. If the drop of the fear score is significantly bigger in the therapy group in comparison to the control group, then the conclusion that therapy works is justified.

Unless regression toward the mean is taken into consideration, paradoxical effects might be observed. If a randomly chosen sample takes part in a research on the efficiency of therapy, those subjects who have the lowest fear scores will drift to the mean, even if the therapy is not affecting them at all. This change might be incorrectly interpreted as an effect of therapy. In this case, the conclusion that therapy increases the level of fear would be false.

Subject Level

The problem of regression toward the mean is not restricted to the group level but is also present on the subject level. In case of repeated measurements with the same subject, a pattern of extreme observations followed by less extreme ones is expected. For example, a student who scored a grade high above average in an initial test is expected to score lower in the follow-up test. The opposite is true for those who earn grades below average because a test result is an effect of knowledge and luck combined. The more extreme the grade, the higher the probability that it is mostly the result of luck. The greater the deviation from the group mean, the greater the regression to the mean effect.

Let us analyze a research on the effectiveness of a new educational program. A random sample of people is a subject of research on a new teaching method. The analysis of results fails to reveal any effect of the new educational program in comparison to the control group. But the researcher, in order to understand precisely who benefited the most from this method, might try to investigate whose gain of knowledge is the biggest on the basis of a baseline measurement. This kind of analysis is typically done by estimating Pearson correlation coefficient between the baseline measurement and the difference between the follow-up measurement and the baseline. Negative correlation coefficient in such an analysis is commonly interpreted as the effect of higher improvement for those whose results were the lowest in the baseline measurement. Unfortunately, this interpretation is false. The correlation coefficient between the baseline measurement and the gain is always negative due to regression toward the mean.

How to Deal With Regression Toward the Mean in Research

If the repeated measurements are not perfectly correlated, a regression toward the mean effect should be assumed. There are several ways to deal with this problem. The first three of them are generally applied when designing a research; the fourth can be introduced at the statistical analysis stage.

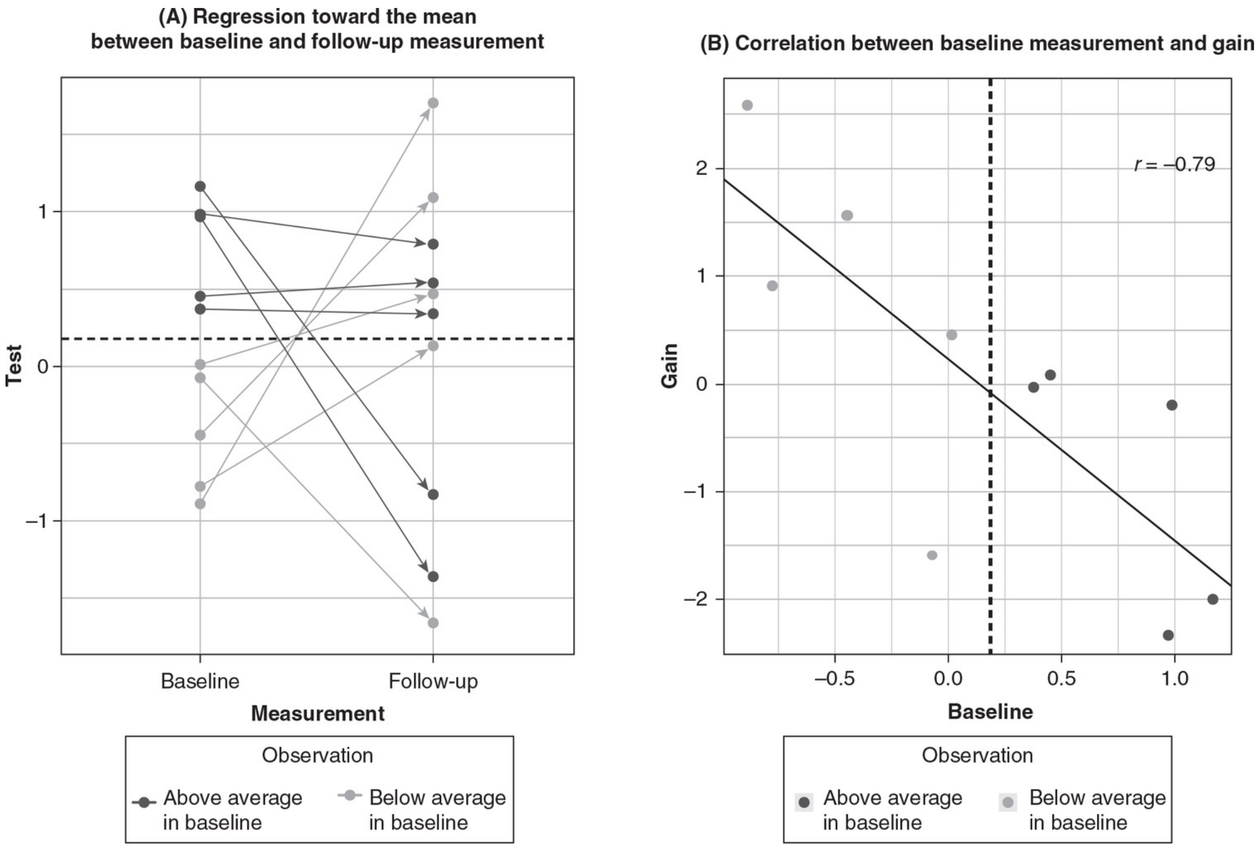
First, reliable psychometric tools must be used. The lower the reliability of a tool, the higher the regression toward the mean effect.

Second, to reduce regression toward the mean at the group level, participants must be randomly drawn from the population and assigned to groups at random. In this situation, the mean change caused by regression toward the mean is equal for each group. Only a change in the research group that is significantly bigger than a change in the control group can be interpreted as an effect of experimental manipulation and not simply a regression toward the mean.

Third, if assignment of participants to groups is not random but is based on some measurement, the variability of baseline measurement should be reduced. To do so, multiple baseline measures can be used and subjects can be incorporated into so, multiple baseline measures can be used and subjects can be incorporated into research groups on the basis of the mean value from those multiple measurements. This will significantly reduce effect of regression toward the mean.

Fourth, at the statistical analysis stage, regression toward the mean effect can be controlled by using analysis of covariance. This technique is a special case of general linear model and is available in most modern statistical packages (e.g., R, SAS, SPSS, STATA).

Figure 1 This graph depicts how regression toward the mean could be confused with the gain of knowledge. Panel A depicts regression toward the mean effect between baseline and follow-up measurement. Arrows show how each observation move due to the regression toward the mean effect. Dashed line shows mean value in the baseline measurement. Panel B depicts correlation between baseline measurement and gain (follow-up minus baseline) for this example. Observations which are above average at baseline measurement have lower gain than observations that are below average on the first measurement. Dashed line shows mean value in the baseline measurement



Maciej Taraday and Anna Wieczorek-Taraday

See also Analysis of Covariance; Generalized Linear Mixed Models; Parameter Random Error; Pearson Correlation Coefficient; Random Error of Measurement; Regression Reliability; Residuals; SAS; Scientific Method; SPSS; STATA

Further Readings

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. International Journal of Epidemiology, 34(1), 215–220.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Disadvantaged child (Vol. 3). New York, NY: Brunner/Mazel.

Campbell, D. T., & Stanley, J. E. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago, IL: Rand McNally.

Cronbach, L. J., & Furby, L. (1970). How we should measure “change”—Or should we? Psychological Bulletin, 74(1), 68–80.

Galton F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 15, 246–263.

Galton, F. (1887). Typical laws of heredity. Nature, 15, 492–495, 512–514, 532–533.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237–251.

McCambridge, J., Kypri, K., & McElduff, P. (2014). Regression to the mean and alcohol consumption: A cohort study exploring implications for the interpretation of change in control groups in brief intervention trials. Drug and Alcohol Dependence, 135, 156–159.

Mee, R., & Chua, T. (1991). Regression toward the mean and the paired sample *t* test. The American Statistician, 45(1), 39–42.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. Psychological Bulletin, 88(3), 622–637.

Stigler, S. M. (1997). Regression towards the mean, historically considered. Statistical Methods in Medical Research, 6, 103–114.

Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. British Medical Journal, 323(7321), 1123–1124.

Marilyn M. Ault Marilyn M. Ault Ault, Marilyn M.

Reinforcement

Reinforcement

1395

1397

Reinforcement

Reinforcement, or the use of a reinforcer, is a process identified and used within the framework of operant conditioning. In its simplest form, reinforcement occurs if the preceding behavior increases by some measure. It is exclusively identified by its effect on a preceding behavior. If the behavior increases in frequency or probability, duration, strength, rate, or a similar measure, then the condition following the behavior is considered to be a reinforcer. The process of administering, providing access to, encountering, or withdrawing something that results in the increase in frequency, strength, or improvement of a preceding behavior is considered reinforcement.

Reinforcement theory is generally considered to have been developed and refined by B. F. Skinner, during the 1930s–1980s, within the context of behaviorism. He interpreted reinforcement as resulting in an increase in response strength or response rate. During that period, and subsequently, many variations on types and conditions of reinforcement have been extensively developed and empirically validated to affect human behavior as well as that of other mammals and animals such as reptiles and mollusks.

The process of reinforcement has a significant role in the design, development, and analysis of interventions in general and special education, applied behavior analysis, and behavioral therapy. A tremendous body of work is available across educational and psychological research addressing the types, use, and effectiveness of reinforcement strategies. Whether dealing with behavior in the context of operant conditioning, or viewing behavior in a broader context such as developmental or social learning theory, the understanding and use of reinforcement has a place. For example, Albert Bandura, in his theory of social reinforcement has a place. For example, Albert Bandura, in his theory of social learning, included reinforcement as a component of learned behavior but listed other mediating factors, and Jean Piaget included reinforcement arranged by nature within cognitive learning theory.

Reinforcement Within Operant Conditioning

Within the context of operant conditioning, reinforcement can be further understood by different types and schedules used to affect the preceding behavior.

Positive and Negative Reinforcement

The terms *positive* and *negative* describe a type of reinforcement process. While the outcome for both types is identified by the effect of the reinforcement on a behavior, positive reinforcement achieves the effect by presenting something and negative reinforcement achieves the effect by removing something. Both result in increasing, strengthening, and so on, the preceding behavior. Positive reinforcement is the strengthening, and so on, of a preceding behavior by providing access to a stimulus, such as attention, food, shelter, or a preferred activity such as a game or being with friends. Negative reinforcement is the strengthening, and so on, of a preceding behavior by removing or allowing the avoidance of a negative or disliked stimulus, such as attention, pain, nagging, or a disliked activity such as cleaning, extra homework, a test, or isolation. Whether a stimulus or event will serve as a positive or negative reinforcer depends on its effect on the preceding behavior. For example, in some cases, teacher attention may be used as a positive reinforcement because giving attention increases the duration of correct responding. With other youth, however, teacher attention may be something the individual wants to avoid, and the teacher withdrawing proximity may increase the duration of appropriate behaviors.

Schedules of Reinforcement

The schedule or timing of the delivery of a reinforcer, whether positive or negative, has a significant impact on the strength of the preceding behavior. Strength can be determined by the persistence of a behavior after the reinforcement has been withdrawn. The two general categories of schedules of reinforcement are continuous and intermittent. Continuous reinforcement is access to the reinforcer every time the behavior occurs. For example, every time the teacher calls the name of a child, the child turns to look at the teacher and the teacher gives the child a hug or a toy. Assuming that the hug and toy are reinforcers, the child will quickly learn to respond to the child's name. Behavior built with continuous reinforcement is, however, weak and extinguishes quickly. Once the teacher attention is no longer present or the toy is not available, the response will quickly decline. Continuous reinforcement is, therefore, used to rapidly build a behavior, but the behavior is considered weak or easily extinguished. After a behavior is established, then the schedule of reinforcement is shifted to an intermittent strategy to strengthen and sustain the behavior.

An intermittent schedule of reinforcement, in its simplest form, is applied when the reinforcer is not accessed after the occurrence of each behavior. Access to the reinforcement varies, according to a certain type of schedule. This variation can be based on a fixed-ratio, variable ratio, fixed-intermittent, or variable-intermittent schedule. The schedules of reinforcement based on ratio address the number of times the behavior occurs before a reinforcement is accessed, and the schedules identified as intermittent are based on the interval of time between access to a reinforcer. The terms *fixed* or *variable* describe whether the number of occurrences or the interval of time is consistent or inconsistent. The variable schedules of reinforcement, particularly variable-ratio, produce the most persistent behavior. A large volume of research has identified the effectiveness of schedules of reinforcement and the use of these schedules across a variety of conditions in education, psychology, and management.

Primary and Secondary Reinforcers

The determination of what constitutes a reinforcer is based on the effect that access to it has on a preceding behavior. A distinction of a reinforcer that may assist in its identification and use is to determine whether it is considered primary or secondary. A primary reinforcer is one that is associated with physical need and survival, such as food, water, air, sleep, or sex. It is also referred to as unconditioned, indicating that the positive impact is not learned. By this definition, some chemicals may also be considered to be primary reinforcers in that the reaction to the stimulus is not learned.

A secondary reinforcer is an item or event that acquires its effectiveness through experience or learning and with its association, or pairing, with an established reinforcer, either primary or secondary. In this context, there are an endless number of potential reinforcers.

Additional Considerations

A large body of research has also explored various issues related to reinforcement. These include, for example, attention, contingencies of reinforcement, delayed reinforcement, internal versus external control of reinforcement, deprivation, and satiation.

Marilyn M. Ault

See also Behaviorism; Operant Conditioning; Premack Principle; Self-Regulation

Further Readings

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In Intrinsically motivated learning in natural and artificial systems (pp. 17–47). Berlin, Germany: Springer.

Beavers, G. A., Iwata, B. A., & Gregory, M. K. (2014). Parameters of reinforcement and response-class hierarchies. Journal of Applied Behavior Analysis, 47(1), 70–82. doi:10.1002/jaba.102

Kessen, W. (1971). Early cognitive development: Hot or cold? In T. Mischel (Ed.), Cognitive development and epistemology (pp. 287–442). New York, NY: Academic Press.

MacDonald, J. M., Ahearn, W. H., Parry-Cruwys, D., Bancroft, S., & Dube, W. V. (2013). Persistence during extinction: Examining the effects of continuous and intermittent reinforcement on problem behavior. Journal of Applied Behavior Analysis, 46(1), 333–338. doi:10.1002/jaba.3

Mace, F. C., Pratt, J. L., Zangrillo, A. N., & Steege, M. W. (2011). Schedules of reinforcement. Handbook of Applied Behavior Analysis, 55–75.

Skinner, B. F. (2014). Contingencies of reinforcement: A theoretical analysis (Vol. 3). Cambridge, MA: BF Skinner Foundation.

Fen Fan Fen Fan Fan, Fen

Jennifer Randall Jennifer Randall Randall, Jennifer

Reliability

Reliability

1397

1402

Reliability

According to the *Standards for Educational and Psychological Testing*, reliability (also referred to as measurement precision) refers to the consistency of assessment results over independent administrations of the testing procedure. The assessment results can be examinees’ scores or raters’ ratings of examinees’ performances on an assessment. Reliability is a central concept in measurement and a necessary condition when building a validity argument. Indeed, if an assessment fails to yield consistent results, it is imprudent to make any inferences about what a score signifies. Reliability is high if the scores or ratings for each examinee are consistent over replications of the testing procedure. Reliability coefficients range from 0 to 1, with 0 being extremely unreliable and 1 representing perfect reliability. There is no absolute critical value for acceptable reliability as the need for precision depends on the stakes of the assessment. Typically, high-stake assessments (e.g., college admission tests) necessitate higher reliability standards than low-stake assessments (e.g., classroom examinations). This entry describes the most popular methods for estimating reliability as well as factors impacting reliability from both the classical and modern test theory perspectives.

Methods to Estimate Reliability

In classical test theory, the consistency of test scores is evaluated mainly in terms of reliability coefficients, and defined in terms of the correlation between scores derived from replications of the test procedure on a sample of test takers.

There are four broad types/categories of reliability coefficients: stability coefficients, equivalence coefficients, internal consistency coefficients, and coefficients based on interrater agreement. Each type of coefficient reflects the variability associated with different data-collection designs and interpretations or uses of scores.

Stability Coefficients: The Test–Retest Method

The test–retest method, a measure of stability, is used to determine the consistency of the examinees' scores on a test over time. The test–retest coefficient is obtained by correlating the scores of identical tests administered to the same examinees twice under similar testing conditions. Carry-over effects and the interval of time between the two test administrations can influence the test–retest coefficient, so this method is most appropriate for tests measuring traits that are not susceptible to carry-over effects and that are stable across time intervals. In practice, the longer the time interval between administrations, the lower the estimated reliability.

Equivalence Coefficients: The Alternate Forms Method

The alternate forms method, a measure of equivalence, is used to examine the consistency of two sets of scores on two parallel forms of a test. The alternate form coefficient is obtained by correlating the scores of parallel (or equivalent) forms of a test to the same examinees under similar conditions in close succession. That is, one form is administered to a group of examinees followed (at a well-chosen close time point) by the administration of an alternate form. The quality or similarity of the parallel forms can influence the alternate form coefficient. In practice, if the forms are not parallel, the alternate form method produces low estimates of reliability.

Internal Consistency Coefficients: Split-Half, KR-20, and Coefficient α Methods

Both measures of stability and equivalence require two administrations of (or parallel forms of) a test, but the administration of two tests can be impractical or unnecessary in reality. Internal consistency coefficients, which require a single unnecessary in reality. Internal consistency coefficients, which require a single test administration, are used to assess the consistency of the examinees' responses to the items within a test. There are two broad classes of methods for estimating internal consistency coefficients. The first class is generally denoted as split-half procedures. The second class of methods requires an analysis of the variance–covariance structure of the item responses. With respect to the split-half methods, a test is administered to a group of examinees, then the test is split into two parallel halves, and the two sets of scores from the two split halves are correlated. This half-test reliability estimate is then used to calculate the full test reliability using the Spearman-Brown prophecy formula, which is written as follows:

$$\rho_{XX'_n} = \frac{2\rho_{AB}}{1+\rho_{AB}},$$

where is the reliability projected for the full-length test with *n* being the number of total items in a test, and ρ_{AB} is the correlation between the half-tests A and B.

When calculating reliability based on item covariance, the two most widely used procedures are KR-20 (Kuder Richardson-20) and coefficient α (often referred to as Cronbach's α). Coefficient is computed by

$$\hat{\alpha} = \frac{k}{k-1}\frac{\sum\hat{\sigma}_i^2}{(1-\hat{\sigma}_X^2)},$$

where *k* is the number of items on the test, is the variance of item *i*, and is the total test variance. KR-20, a special case of coefficient α for dichotomously scored items only, is also based on the proportion of persons passing each item and the standard deviation of the scores.

Coefficients Based on Interrater Agreement: Interrater Method

The interrater method, a measure of consistency of ratings, is used to examine the consistency of observed performances over different raters or observers. It is obtained by having two or more observers rate a performance of any kind and calculating the percentage of agreement between observations. The interrater approach is the preferred method when calculating the reliability of assessments/performances such as constructed responses, speeches, debates, or musical performances. Variation among raters and variability in the interpretation of assessment results are the two potential sources of error influencing interrater reliability.

Factors Affecting Reliability

In this section, the factors that impact the reliability of assessment results are discussed. Although individual characteristics (e.g., motivation, fatigue, health, and ability) as well as the quality of assessment itself (e.g., clarity of instructions and test difficulty) inevitably impact all reliability estimates, here, the focus is on the three most widely cited sources of error with respect to reliability.

Test Length

Generally speaking, the longer the measure is, the more reliable the measure is. As test length increases, the proportion of the student’s score that can likely be attributed to error decreases. For example, low ability students may answer a single item correctly, even if guessing; however, it is much less likely that low ability students will correctly answer all items on a 20-item test via guessing. The use of longer measures minimizes the impact of singular human error. Other test characteristics being equal (e.g., item quality), a measure with 40 items should have higher reliability than one with 20 items. The relationship between reliability and test length can be mathematically shown in the Spearman-Brown prophecy formula mentioned previously. The formula is based on the assumption that, when tests are shortened or lengthened, items of comparable content and statistics to those already in the test are deleted or added. For example, if the reliability of a 20-item test is determined to be 0.75, and the length of the test is doubled by adding items of comparable content and statistics, then the predicted reliability of the new test would be

$$0.86\left(\frac{2 \times 0.75}{(1+0.75)}\right).$$

Spread of Scores

Because reliability is sample dependent, all other factors being equal, the greater the spread of scores, the higher the reliability estimate. Indeed, larger reliability coefficients result when examinees remain in the same relative position in a group across multiple administrations of an assessment. To be sure, errors of measurement have less influence on the relative position of individuals when the differences among group members are large (when there is a large spread of scores). Consequently, anything that reduces the possibility of shifting positions in the group (e.g., a heterogeneous sample of examinees) also contributes to larger reliability coefficients.

Objectivity of Scoring

The objectivity of scoring influences reliability in the sense that the error introduced by the scoring procedure varies with respect to the extent that human judgment is required. With objective items such as multiple-choice or matching items, the scoring presents little opportunity for the introduction of human error. Constructed response items and performance assessments, however, often involve the subjective judgments of human raters or scorers. Consequently, they are subject to different degrees of scoring error, depending on the nature of the question and the scoring procedures. For example, short-answer constructed response items tend to be more objectively scoreable than longer, more complex student responses (e.g., essays) and products (e.g., projects).

Standard Error of Measurement (SEM)

Within a classical test theory framework, an examinee’s observed test score (*X*) is composed of two parts: the true score (*T*) and the error score (*E*):

$$X = T + E$$

The true score can be interpreted as the average of the observed scores obtained over an infinite number of repeated administrations with the same test or parallel forms of the test. The error score is the difference between the observed test score and the true score.

The SEM is an estimate of the extent to which an examinee’s scores vary across administrations. For example, for a group of examinees, each individual has a true score and several possible observed scores around the individual’s true score. Theoretically, each examinee’s personal distribution of possible observed scores around the examinee’s true score has a standard deviation. The SEM is the average of these individual error standard deviations for the group.

Another way of thinking about reliability is that it refers to the extent to which students’ scores are free from errors of measurement. Assuming errors are random and independent, the observed score variance can be further decomposed into the variance in true scores and the variance in the errors of measurement . The reliability coefficient (or the correlation between two measures of the same trait) can also be mathematically defined as the ratio of true score variance to observed score variance. SEM (σ_E) is a function of the standard deviation of observed scores (σ_X) and the reliability coefficient ($\rho_{\perp}XX'$):

$$\sigma_E = \sqrt[\sigma_X]{1-\rho'_{XX}}.$$

Note that as the reliability coefficient increases, the SEM decreases.

Classification Consistency and Accuracy

Decision consistency (DC) refers to the extent to which classifications of examinee decisions agree based on two independent administrations of the same exam or two parallel forms of an exam. Decision accuracy (DA) refers to the extent to which the actual classifications based on observed scores agree with the “true” classifications. The DC and DA are important for assessments with a purpose to classify examinees into performance categories (as is often the purpose of criterion-referenced tests). Similar to classical reliability with respect to the consistency of overall assessment results, consistency of students’ classifications is also a necessary condition when building a validity argument for criterion-referenced tests. Without certain confidence in the consistency of students’ classifications, any inferences based on the classifications would be dubious.

Methods to Estimate DC and DA

When calculating or determining DC and DA, the two most common indices are the agreement index *P* and Cohen’s κ. The agreement index *P* is defined as the proportion of times that the same decision would be made based on two parallel forms of a test. It can be expressed as

$$P = \sum_{j=1}^{J} P_{jj},$$

where J is the number of performance categories, and P_{jj} is the proportion of examinees consistently classified into the jth category across the two administrations or forms of a test. If Form 1 is one set of observed scores, and Form 2 is replaced with the true scores or another criterion measure, then P becomes the DA index. To get a more interpretable measure of decision-making consistency, Cohen’s κ can be computed as follows:

$$\kappa = \frac{P_0 - P_C}{1 - P_C},$$

$$P_0 = \sum_{j=1}^{J} P_{jj},$$

$$P_C = \sum^{J} P_{j}.P._{j},$$

where P_0 is the observed proportion of agreement, P_C is the expected proportion of agreement, P_{jj} is the proportion of examinees consistently classified into the jth category, and $P_j.$ and $P._j$ are the marginal proportions of examinees falling in the jth category across the two administrations of the test, respectively. P_C represents the DC expected by chance.

κ can be thought of as the proportion of agreement that exists above and beyond that which can be expected by chance alone. κ has a value between −1 and 1. A value of 0 and below indicates that the decisions are as consistent as the decisions based on two tests that are statistically independent. In other words, the decisions are very inconsistent and the reliability of classifications is extremely low. A value of 1 indicates that the decisions are as consistent as the decisions based on two tests that have perfect agreement.

Reliability From Item Response Theory (IRT) Perspective

Unlike classical reliability, which uses a single value to describe a measure's average reliability, in IRT, reliability is not uniform across the entire range of proficiency levels. Scores at both ends of the proficiency level generally have more errors associated with them than scores at the center of the proficiency distribution. IRT emphasizes the examination of item and test information in lieu of classical reliability. In mathematical statistics, the term (Fisher) *information* conveys a similar, but more technical, meaning. It is defined as the reciprocal of the precision with which a parameter could be estimated. For instance, in IRT, an interest is in estimating the value of the ability parameter (θ) of an examinee, which is denoted by . All ability estimates have a variance , which is a measure of the precision with which a given ability level can be estimated. The amount of information (*I*) at a given ability level is the reciprocal of this variance and can be shown as follows:

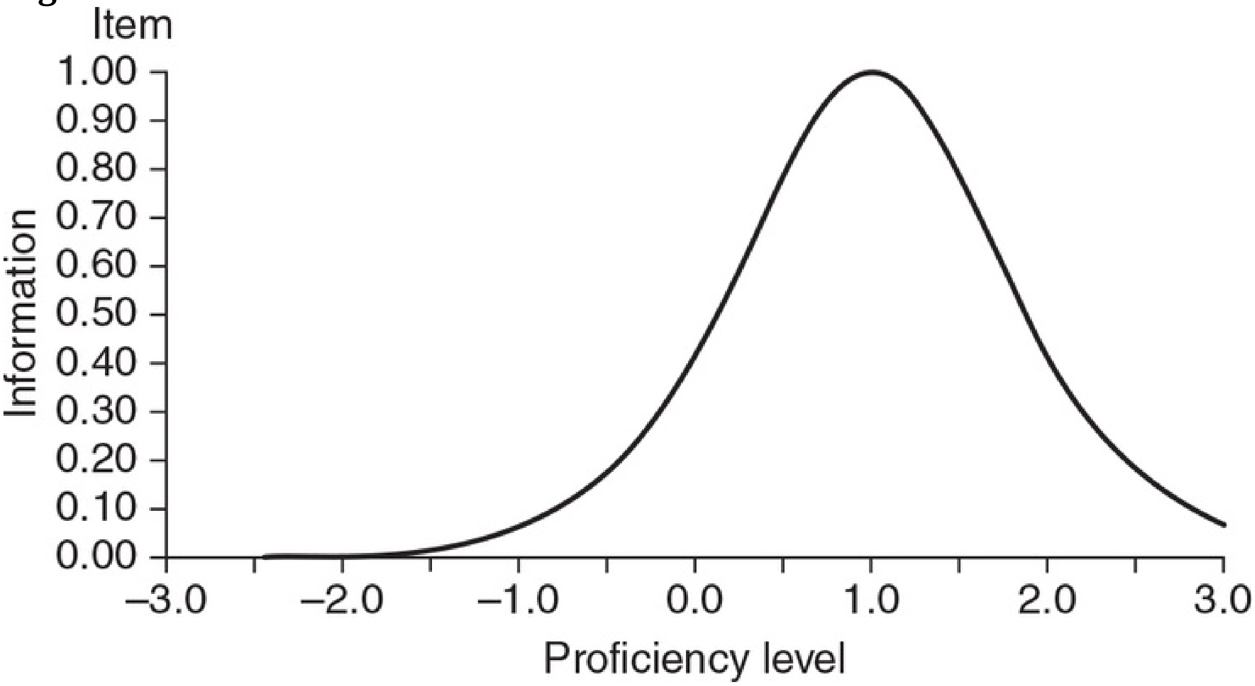
$$I \mid \theta = \sqrt{\left(\frac{1}{(\sigma^{\uparrow}2 \mid \hat{\theta})}\right)}.$$

The higher the information at a given ability level, the more precise the item parameter estimate tends to be than one with lower information.

Under IRT, each item on a test measures the proficiency level or ability of an examinee. Therefore, the amount of information for any single item can be computed at any ability level. The mathematical definition of the amount of item information depends upon the particular IRT model employed. For the one-parameter logistic and Rasch models, the item information is a function of the item difficulty parameter. For the two-parameter logistic model, the item information is a function of the item discrimination and item difficulty parameters, whereas for the three-parameter logistic model, the item information is a function of item discrimination, item difficulty, and pseudo-guessing parameters. Generally speaking, item information functions tend to have a bell shape. Highly discriminating items have tall, narrow information functions that provide considerable information but over a narrow range (Figure 1), whereas less discriminating items provide less information over a wider range (Figure 2).

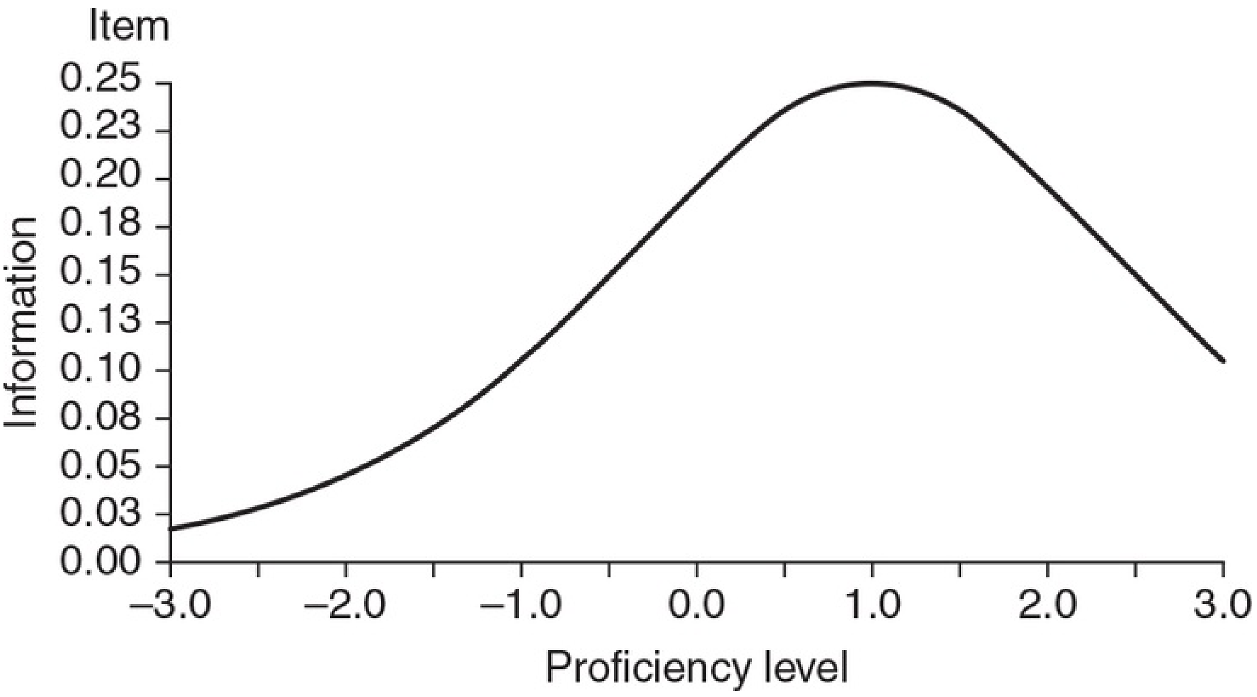
The highest item information of Item 1 is 1, whereas the highest item information of Item 2 is 0.25.

Figure 1 Item information function for Item 1



Note: This item is simulated using 2PL model with an item discrimination parameter of 2.0 and item difficulty parameter of 1.0 on the logistic scale.

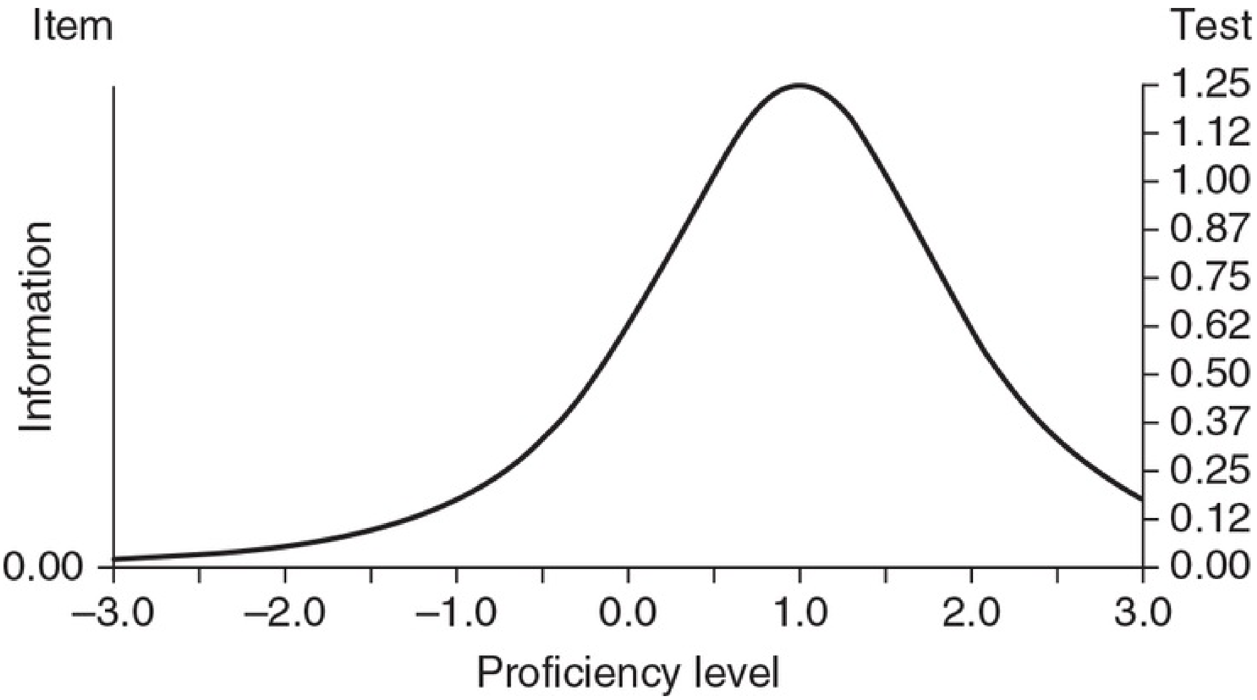
Figure 2 Item information function for Item 2



Note: This item is simulated using 2PL model with an item discrimination parameter of 1.0 and item difficulty parameter of 1.0 on the logistic scale.

Because items are conditionally independent of each other given an individual’s score, the test information function (TIF) is simply the sum of information of all items on a test. Assume that a test with the 2 items above, the TIF of the test looks like that shown in Figure 3.

Figure 3 Test information function



Note: The TIF of this test is composed of two items: one with item discrimination of 2.0 and item difficulty of 1.0, and the other one with item discrimination of 1.0 and item difficulty of 1.0.

The TIF is 1.25 (the sum of item information of Items 1 and 2) and it is modal around 1.0, which is the item difficulty of both items.

The conditional SEM, the reciprocal of the test information at a given trait level (θ), is obtained as follows:

$$\sigma_{\downarrow}E \mid \theta = \sqrt{\frac{1}{\mathrm{TIF}}}.$$

The aggregate SEM, which is analogous to the SEM from CTT perspective, is obtained as follows:

$$\sigma_{E} = \sqrt{\frac{1}{\mathrm{TIF}}}.$$

That is, the measurement error is equal to the square root of the reciprocal of the test information and it is interpreted in the same way as the traditional SEM. With a large item bank, TIFs can be manipulated to control measurement error very precisely because the TIF shows the degree of precision at each individual proficiency level.

Final Thoughts

The reliability—as it is a precursor to establishing test score validity—of a measure is a critical consideration. Reliability and the SEM can be obtained from both classical and IRT perspectives and they are conceptually the same. The choice of method for establishing an assessment's reliability should be determined in light of the data collection design (e.g., two test administrations or single test administration, the same test or parallel forms available) and the intended interpretation and/or use of scores (e.g., stability, equivalence, internal consistency, or classification consistency). The level of precision required depends on both the purpose and stakes of the assessment. To ensure reliable results when designing assessments, one should encourage test takers to perform their best, have scoring criteria that are readily available by test takers and raters (when appropriate), allow enough time, and have enough items. Ultimately, the purpose of any assessment is to provide meaningful feedback about what examinees know and are able to do. Well-developed assessments yielding consistent results are key to this goal.

Fen Fan and Jennifer Randall

See also Classical Test Theory; Internal Consistency; Item Response Theory; Split-Half Reliability; Test Information Function; Test–Retest Reliability; Validity

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart, and Winston.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, 32(2), 179–197.

Thorndike, R. M., & Thorndike-Christ, T. M. (2009). Measurement and evaluation in psychology and education (8th ed.). Boston, MA: Pearson.

Matthew Gordon Ray Courtney Matthew Gordon Ray Courtney Courtney, Matthew Gordon Ray

Repeated Measures Analysis of Variance

Repeated measures analysis of variance

1402

1407

Repeated Measures Analysis of Variance

The repeated measures analysis of variance (ANOVA) is an omnibus test that is an extension of the dependent samples *t* test. The test is used to determine whether there are any significant differences between the means of three or more variables (also called levels). The repeated measures ANOVA is used when the sampled observations are measured under a number of conditions (this is why sometimes the test is referred to as an ANOVA for correlated samples). Where this data condition exists, a standard ANOVA would not be appropriate as it would not take into account the natural correlation (relationship) between the repeated measures. In the context of educational assessment, if we were to test a group of students’ ability on a standardized math exam three times in a longitudinal study, we would expect a higher correlation between each of the three measured outcome variables. In a study in which three different groups of individuals were assessed, we would not expect such a strong relationship and therefore make use of the standard ANOVA. This entry reviews various aspects of the repeated measures ANOVA, including terminology, assumptions, and statistical procedures and calculations.

Terminology

There are several statistical terms commonly used to describe a repeated measures ANOVA. When investigations involve variables pertaining to studied participants, a sampled member is often referred to as a *subject*. In education, subjects are often students. When the same dependent variable (outcome) is measured repeatedly for all subjects across a set of conditions, the set of conditions is referred to as a *within-subjects factor*. For studies involving one group's standardized test scores on three occasions, the within-subjects factor would be the *Time* (e.g., Time 0, Time 1, and Time 2). The conditions that contextualize this factor is often referred to as *trials*. When the outcome of interest (dependent variable) is measured three or more times on different groups (such as control and intervention groups), the set of conditions is called the *between-subjects factor*. For educational studies involving the assessment of a control and an intervention group's standardized test scores on three occasions, the *between-subjects factor* would simply be *groups*. In this case, the research design would be a two-way repeated measures ANOVA.

When to Use Repeated Measures ANOVA

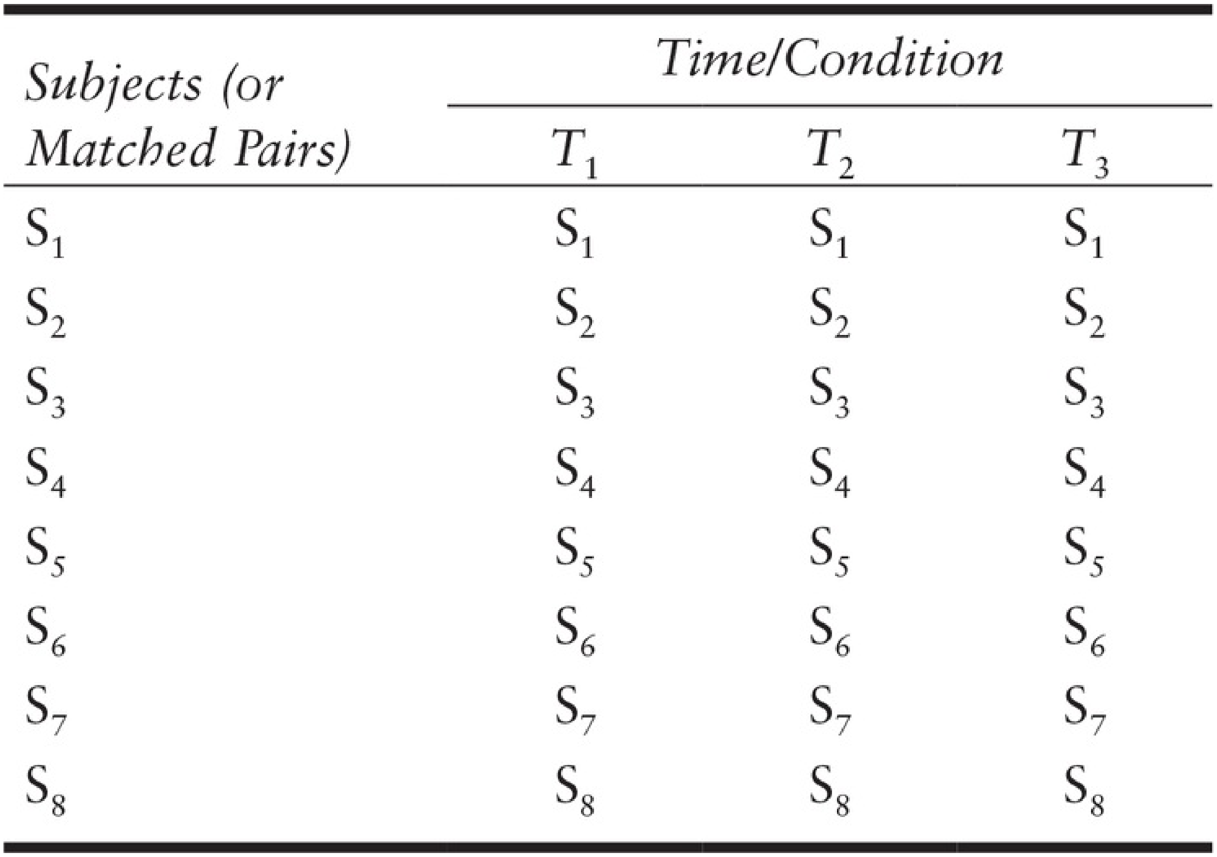
In the context of educational measurement, the repeated measures ANOVA is generally used in two different types of research conditions: studies that investigate (1) changes in means over three or more *time points* or (2) differences in means under three or more *conditions*.

For the first example, we may be investigating the effect of a new mathematics program on students' performance on a standardized test at three separate time points: Time 0, Time 1, and Time 2 (pre-, midway-, and postprogram intervention). This would enable us to develop an understanding of the possible timing and extent of improved mathematics ability. In this case, the within-subjects factor would be Time with three levels.

For the second example, we may be interested in the ability of students to recall historic events and associated dates and make use of three learning strategies. This might help us determine which strategy might best suit the students in the class. In this case, the within-subjects factor could be deemed the study condition.

The repeated measures ANOVA can also be applied when sample members have been matched. In this case (based on subject-level demographic attributes), subjects are matched and therefore measurements across conditions are treated like repeated measures. Basically, in this instance, each matched pair would be treated as a single observation or sampled member.

Table 1 expresses the basic repeated measures ANOVA data design in which eight subjects (or match pairs) are performing under three different time points (or conditions). Basically, the repeated measures ANOVA can be employed when subjects undergo repeated measurements at either different times or under different conditions.



Subjects (or Matched Pairs)	*Time/Condition*		
	T_1	T_2	T_3
S_1	S_1	S_1	S_1
S_2	S_2	S_2	S_2
S_3	S_3	S_3	S_3
S_4	S_4	S_4	S_4
S_5	S_5	S_5	S_5
S_6	S_6	S_6	S_6
S_7	S_7	S_7	S_7
S_8	S_8	S_8	S_8

Assumptions Necessary for Test

There are five general assumptions necessary to carry out a repeated measure ANOVA. The first assumption is that the outcome of interest (dependent variable) is at the continuous level (where ordinal variables are concerned, one would choose nonparametric tests). Second, the independent variable should be categorical and consist of at least three levels. This might constitute three time points, conditions, or matched entities. Third, there should be no significant outliers; and fourth, the distribution of the outcome of interest, across all repeated measures, should be approximately normally distributed. If outliers and/or nonnormality exist, transformational procedures may be employed to resolve these problems. The final assumption is that of sphericity, whereby the variances of the differences between all combinations of related groups must be equal. Mauchly's test of sphericity can be employed in many statistical software equal. Mauchly's test of sphericity can be employed in many statistical software programs, and statistically, nonsignificant test result would suggest that the data meet the assumption of sphericity.

Repeated Measures ANOVA and the Hypothesis Test

The repeated measures ANOVA assumes that the sample subjects were drawn from a related population. And, the test assesses whether there are likely to be any differences between related population means. Therefore, the null hypothesis (H0) states that the means are equal:

or

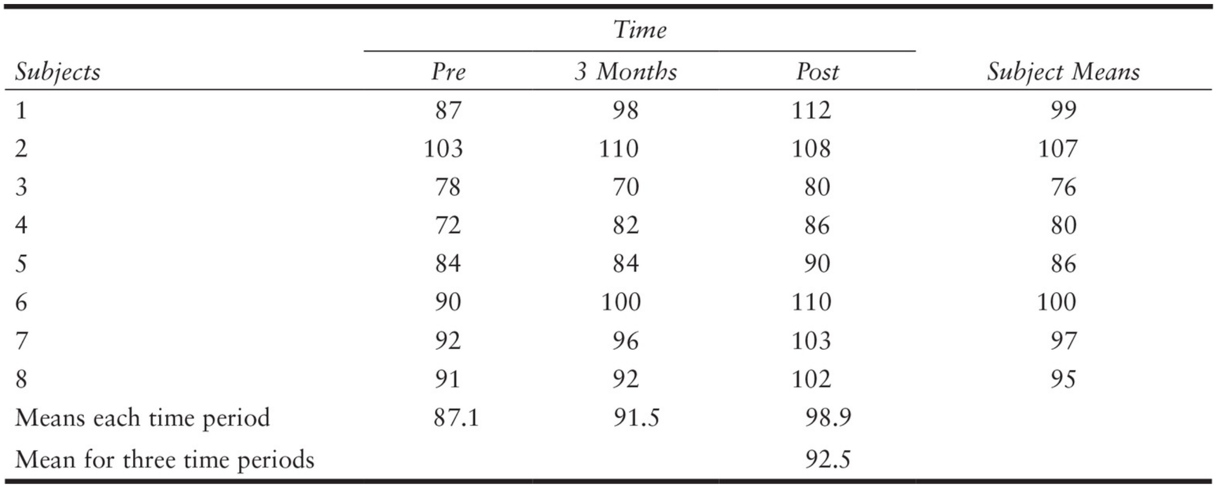
HA: at least two means are statistically significantly different,

where µ1 = population mean, *k* = number of related groups, HA = alternative hypothesis.

It is important to note that the repeated measure ANOVA will not inform the researcher which pairing of time or conditions constitutes a statistically significance difference in means. For example, does µ1 = µ2? How about µ1 = µ3? One would need to carry out post hoc dependent samples *t* tests to determine which pairings exhibit statistically significant difference in means.

Statistical Procedures Undertaken in the Repeated Measures ANOVA

For the purpose of illustrating a one-way repeated measures ANOVA, we make use of a fictional psychometric data set. Table 2 provides an example of eight participants' IQ scores over the course of a 6-month study.



Subjects	Time Pre	 3 Months	 Post	Subject Means
1	87	98	112	99
2	103	110	108	107
3	78	70	80	76
4	72	82	86	80
5	84	84	90	86
6	90	100	110	100
7	92	96	103	97
8	91	92	102	95
Means each time period	87.1	91.5	98.9	
Mean for three time periods			92.5	

Calculating Between-Time (Groups) Variability

The repeated measures ANOVA starts by calculating the variability associated with the different time points (VAR_{time}). Remember, in a different research design, this could also be the conditions. The between-time variability is calculated as follows:

$$VAR_{\text{time}} = i = 1kni(xi - xi)2,$$

where k is the number of conditions, ni is the number of subjects under each (ith) condition, x is the grand mean. So, based on the values in our table, we have

$$VAR_{\text{time}} = 8[(87.1 - 92.5)2 + (91.5 - 92.5)2 + (98.9 - 92.5)2]$$

$$VAR_{\text{time}} = 8[(-5.4)2 + (-1)2 + (6.4)2]$$

$$VAR_{\text{time}} = 8[29.2 + 1 + 41]$$

$$VAR_{\text{time}} = 8[71]$$

$$VAR_{\text{time}} = 568$$

Therefore, the VAR_{time} provides a metric of total variation between the three time points. In this case, VAR_{time} is 568.

Calculating Within-Subjects Variability

The repeated measures ANOVA then calculates the variability within the subjects (VAR_{time}). Remember, in a different research setting, this could also be the three conditions. The within-subjects variability is calculated as follows:

$$VAR_{\text{within}} = 1(xi1 - x1)2 + 2(xi2 - x2)2 + \ldots + k(xik - xk)2,$$

where *xi*1 is the score of the *i*th subject in Group 1, *xi*1 is the score of the *i*th in Group 2, and *xik* is the score of the *i*th subject in Group *k*. So, based on the values in our table, we have

$$VAR_{\text{within}} = \Big[(87 - 87.1)2 + (103 - 87.1)2 + (78 - 87.1)2 + (72 - 87.1)2 + (84 - 87.1)2 + (90 - 87.1)2 + (92 - 87.1)2 + (91 - 87.1)2\Big] + \Big[(98 - 91.5)2 + (110 - 91.5)2 + (70 - 91.5)2 + (82 - 91.5)2 + (84 - 91.5)2 + (100 - 91.5)2 + (96 - 91.5)2 + (92 - 91.5)2\Big] + \Big[(112 - 98.9)2 + (108 - 98.9)2 + (80 - 98.9)2 + (86 - 98.9)2 + (90 - 98.9)2 + (110 - 98.9)2 + (103 - 98.9)2 + (102 - 98.9)2\Big]$$

$$VAR_{\text{within}} = \Big[(-0.1)2 + (15.9)2 + (-9.1)2 + (-15.1)2 + (3.1)2 + (2.9)2 + (4.9)2 + (3.9)2\Big] + \Big[(6.5)2 + (18.5)2 + (-21.5)2 + (-9.5)2 + (-7.5)2 + (8.5)2 + (4.5)2 + (-0.5)2\Big] + \Big[(13.1)2 + (9.1)2 + (-18.9)2 + (-12.9)2 + (-8.9)2 + (11.1)2 + (4.1)2 + (3.1)2\Big],$$

$$VAR_{\text{within}} = \Big[0.01 + 252.8 + 82.8 + 228 + 9.6 + 8.4 + 24.0 + 15.2\Big] + \Big[42.3 + 342.3 + 462.3 + 90.3 + 56.3 + 72.3 + 20.3 + 0.3\Big] +, \Big[171.6 + 82.8 + 357.2 + 166.4 + 79.2 + 123.2 + 16.8 + 9.6\Big],$$

$$VAR_{\text{within}} = 620.81 + 1{,}086.4 + 1006.8,$$

$$VAR_{\text{within}} = 2{,}714.$$

Therefore, the VAR_{within} provides an overall metric of subject variation within each of the three time points. In this case, VAR_{within} is 2,714.

Calculating Subjects Variability

Thereafter, repeated measures ANOVA calculates the variability associated with each individual subject. This is calculated via the following formula:

$$VAR_{\text{subjects}} = \; k(xi - x)2,$$

where *k* is the number of conditions, *xi* is the mean of subject *i*, and *x* is the grand mean. So, based on the values in the table, we get:

$$\begin{aligned}VAR_{\text{within}} = 3\Big[&(99 - 92.52) + (107 - 92.52) + \\ &(76 - 92.52) + (80 - 92.52) + \\ &(86 - 92.52) + (100 - 92.52) + \\ &(97 - 92.52) + (95 - 92.52)\Big],\end{aligned}$$

$$\begin{aligned}VAR_{\text{subjects}} = 3\Big[&(6.52 + 14.52) + (-16.52 + 12.52) + \\ &(-6.52 + 7.52) + (4.52 + 2.52)\Big],\end{aligned}$$

$$\begin{aligned}VAR_{\text{subjects}} = 3\Big[&42.3 + 210.3 + 272.3 + 156.3 \\ &+ 42.3 + 56.3 + 20.3 + 6.3\Big],\end{aligned}$$

$$VAR_{\text{subjects}} = 3[806.4],$$

$$VAR_{\text{subjects}} = 2{,}419.2.$$

Therefore, the VAR_{subjects} provides an overall metric of individual variation across three time points. In this case, VAR_{subjects} is 2,419.2.

Calculating Error Variability

Calculating Error Variability

The repeated measure ANOVA procedure also calculates the error variance associated with the sample. We know that the within-subjects variability is equivalent to the subjects variability plus the error variability, as given by the following formula:

$$VAR_{\text{within}} = VAR_{\text{subjects}} + VAR_{\text{error}},$$

$$2,714 = 2,419.2 + VAR_{\text{error}}.$$

Therefore, via simple substitution:

$$VAR_{\text{error}} = 2,714 - 2,419.2,$$

$$VAR_{\text{error}} = 294.8.$$

Mean Sum of Squares for Time (MSS_{time}) and Error (MSS_{error})

To determine the MSS_{time}, we divide the variability associated with the different time points by its degrees of freedom. Because there are three time points in this example, there are two degrees of freedom ($df = k-1$):

$$MSS_{\text{time}} = VAR_{\text{time}}k - 1,$$

$$MSS_{\text{time}} = 5683 - 1,$$

$$MSS_{\text{time}} = 5682,$$

$$MSS_{\text{time}} = 284.$$

To calculate MSS_{error}, we divide the error variance by the $(n-1)(k-1)$ degrees of freedom, where n is the number of subjects and k is the number of time points. In this case,

$$MSS_{\text{error}} = \frac{294.8}{(8-1)(3-1)},$$

$$MSS_{\text{error}} = \frac{294.8}{(7)(2)},$$

$$MSS_{\text{error}} = \frac{294.8}{14},$$

$$MSS_{\text{error}} = \frac{294.8}{14},$$

$$MSS_{\text{error}} = 21.1.$$

Thereafter, the *F* statistic can be obtained.

The *F* Statistic

Finally, the *F* statistic is obtained by dividing the MSS_{time} by the MSS_{error}:

$$F = \frac{MSS_{\text{time}}}{MSS_{\text{error}}},$$

$$F = \frac{284}{21.1},$$

$$F = 13.5.$$

Reporting of ANOVA and Post Hoc Tests

Results generated from the fictitious data set may be presented the following way:

> A one-way repeated measures ANOVA was conducted to compare the effect of time on the sample subjects' IQ assessed at pre-, mid-and poststudy time points. Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated with $\chi^2 = 1.8$, *p* = .4. There was a significant effect of time on subjects' IQ, *F*(2, 14) = 13.5, *p* = .001, partial $\eta^2 = .66$.

The partial η^2 value of .66 means that 66% of the variability in IQ scores is accounted for by the time period that it was measured. As explained, one needs to carry out post hoc dependent samples *t* tests to determine which pairings exhibit statistically significant difference in means. The results of these tests could be presented as follows:

> Three dependent sample *t* tests were conducted to make post hoc comparisons between the three time points. A first *t* test indicated that there was a statistically significant difference between Time 1 (*M* = 87.1, standard deviation [*SD*] = 9.4) and Time 2 (*M* = 91.5, *SD* = 12.5); *t*(7) =

−1.9, *p* = .006. A second *t* test indicated that there was a statistically significant difference between Time 2 (*M* = 91.5, *SD* = 12.5) and Time 3 (*M* = 98.9, *SD* = 12.0; *t*(7) = −4.3, *p* = .004. Finally, a third *t* test indicated that there was a statistically significant difference between Time 1 (*M* = 87.1, *SD* = 9.4) and Time 3 (*M* = 98.9, *SD* = 12.0); *t*(7) = −4.3, *p* = .004.

Summary

The repeated measures ANOVA provides a simple way of assessing change in an outcome of interest over three or more time periods (or conditions). In studies involving educational measurement, the outcome of interest is often some repeated measure of social or educational engagement. With the advent of statistical software programs such as IBM SPSS Statistics, the procedure can be carried out very easily.

Matthew Gordon Ray Courtney

See also Analysis of Variance; Mixed Model Analysis of Variance

Further Readings

Lund Research Group. (2013). ANOVA with repeated measures using SPSS Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/one-way-anova-repeatedmeasures-using-spss-statistics.php

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. The Annals of Mathematical Statistics, 11(2), 204–209.

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research & Evaluation, 15(12). Retrieved from http://pareonline.net/pdf/v15n12.pdf

Randolph, J. J., Falbe, K., Manuel, A. K., & Balloun, J. L. (2014). A step-by-step guide to propensity score matching in R. Practical Assessment, Research & Evaluation, 19(18), 2.

Wessa, P. (2013). Box-cox normality plot—Free statistics software (Version 1.1.23-r7). Office for Research Development and Education. Retrieved from http://www.wessa.net/rwasp_boxcoxnorm.wasp

Markus Brauer Markus Brauer Brauer, Markus

Repeated Measures Designs Repeated measures designs

1407

1409

Repeated Measures Designs

The defining characteristic of repeated measures designs is the fact that independent units—usually participants—are “crossed with” at least one of the independent variables; that is, each unit provides at least one data point for each level of one or more independent variables. In other words, in repeated measures designs, at least one of the independent variables varies “within units” and is thus referred to as a within-unit variable (e.g., within-subjects variable). In the most general sense, repeated measures designs are characterized by data that are clustered by participants (or other units) and are thus nonindependent. Repeated measures designs are different from purely between-subjects designs, in which participants are said to be “nested under” one or more independent variables.

In the simplest repeated measures design, each participant provides one data point for each of the two levels of a dichotomous independent variable. Common repeated measures designs are studies in which participants’ responses are collected twice (e.g., at the beginning and at the end of the school year) or in which each participant is exposed to multiple types of stimuli (e.g., each student evaluates one structured and one unstructured task). In more complex repeated measures designs, independent units are crossed with more than one independent variable or are crossed with some independent variables and nested under others. It is also possible for the within-subjects variable to have more than two levels (e.g., students’ performance is measured 5 times during the academic year).

Statistical Power and Internal Validity

Compared to purely between-subjects designs, repeated measures designs usually have greater statistical power. This is due to the fact that more data points are obtained with the same number of participants and that individual differences are accounted for and therefore do not contribute to the error term of the inferential test. Repeated measures designs frequently have lower internal validity, in that there might be alternative explanations for the observed differences between experimental conditions. Many of the threats to internal validity can be eliminated; however, one can include practice trials before the actual study to avoid learning effects. One can keep the task short or maintain a high level of motivation to do well on the task (e.g., by rewarding participants for good performance) to avoid fatigue effects. Finally, one can space out the measurement moments or include a distractor task between them to avoid carry-over effects from the first experimental condition to the second.

The best way to increase internal validity in a repeated measures design is to counterbalance the order of conditions. Half of the (randomly chosen) participants first do Condition 1 and then do Condition 2 of the independent within-subjects variable, whereas the other half of the participants proceeds in the inverse order. Statistical power is generally increased if order is subsequently included as a predictor in the statistical analyses. The analysis is then a mixed-models analysis of variance with one within-subjects variable (treatment) and one between-subjects variable (order). Depending on the data analysis software the researcher is using, it may be necessary to “center” the order variable (i.e., to recode it into −.5 and +.5 or into −1 and +1) to obtain the treatment effect averaged across order conditions.

In certain pretest–posttest designs, statistical power can be increased by treating the pretest as a covariate (sometimes called analysis of covariance approach or regression adjustment) rather than treating pretest and posttest as two levels of a within-subjects variable (sometimes called repeated measures approach or change score analysis). As noted by G. J. P. van Breukelen, the more powerful pretest-as-covariate approach can be used only if certain conditions are satisfied: (a) There is one (and only one) dichotomous within-subject variable, and one of the two levels is clearly a pretest or a baseline measure, (b) there is at least one between-subjects variable, and (c) the assignment to the levels of the between-subjects variables is either random or determined by participants’ pretest score. The pretest-as-covariate approach consists of regressing the posttest on both the pretest and the between-subjects variables.

Advanced Techniques for Complex Designs

It is possible to statistically control for covariates in repeated measures designs. When the covariate varies between subjects (e.g., an individual difference measure), it suffices to add it to the regression model (like order). When the covariate varies within subjects (e.g., mood assessed at each measurement moment), it is sometimes called a time-varying covariate. According to Charles M. Judd, David A. Kenny, and Gary H. McClelland, the appropriate regression model is then $(Y_2 - Y_1) = b_0 + b_1 (Z_2 - Z_1) + b_2 ((Z_1 + Z_z)/2) - C$, where $(Y_2 - Y_1)$ is the outcome difference, $(Z_2 - Z_1)$ is the covariate difference, and $((Z_1 + Z_2)/2) – C$ is the mean-centered covariate average. The inclusion of the last term is not absolutely necessary, but without it, one makes the (often unreasonable) assumption that the covariate-outcome relationship is the same in both experimental conditions. In the aforementioned equation, the coefficient b_0 tests the (within-subject) treatment effect, statistically controlling for the time-varying covariate.

It is also possible to examine mediation in repeated measures designs. By definition, the mediator has to vary within subjects. Mediation is tested with the same regression equation as shown earlier, the only difference being that Z_1 and Z_2 now refer to the two mediator scores (one per experimental condition). The coefficient b_0 tests for the (within-subject) treatment effect, statistically controlling for the mediator (this effect is referred to as “Path *a*” in many relevant texts on mediation). The coefficient b_1 tests for the effect of the mediator on the outcome variable (usually referred to as “Path *b*”).

In certain repeated measures designs, participants provide multiple responses for each level of the independent within-subjects variable. Sometimes all participants provide responses to the same targets or materials (e.g., there are 10 structured and 10 unstructured tasks, and all students evaluate the same set of 20 tasks) and sometimes each participant reacts to the participant’s own unique set of targets or materials (e.g., each student is asked to nominate and judge 10 same-sex and 10 different-sex friends; each student evaluates a different set of 20 individuals). These types of studies are best analyzed with linear mixed-effects models. Note that these two designs require both a by-subject random intercept and a by-subject random slope, but that the former design—all students evaluate the same set of items—requires in addition a by-item random intercept, according to Charles M. Judd, Jacob Westfall, and David A. Kenny.

Markus Brauer

See also Generalized Linear Mixed Models; Mediation Analysis; Mixed Model Analysis of Variance; Random Assignment; Regression Discontinuity Analysis

Further Readings

Brauer, M., & Curtin, J. J. (in press). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within subjects and/or within items. Psychological Methods.

Judd, C. M., Kenny, D., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. Psychological Methods, 6, 115–134.

Judd, C. M., Westfall, J., & Kenny, D. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. Annual Review of Psychology. doi:10.1146/annurev-psych-122414-033702

Montoya, A. K., & Hayes, A. F. (2016). Two-condition within-participant statistical mediation analysis: A path-analytic framework. Psychological Methods, 22, 6–27. doi:10.1037/met0000086

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. Journal of Clinical Epidemiology, 59, 920–925.

Jonathan A. Plucker Jonathan A. Plucker Plucker, Jonathan A.

Matthew C. Makel Matthew C. Makel Makel, Matthew C.

Replication

Replication

1409

1412

Replication

Replication is the purposeful repetition of previous research to confirm or disconfirm the previous results. Replications also comprise the research used to compose meta-analyses. However, it is important to note that meta-analyses are not the same as replications. Replication is necessary for meta-analysis, but meta-analyses can be based on studies with quite varied purposes. For example, a meta-analysis on the effects of academic acceleration could rely on studies investigating grade skipping and early entrance into kindergarten even though the individual studies would not be considered replications of each other. Thus, studies may come from the same meta-analytic pool but may not serve the same purpose. Meta-analyses synthesize previous research, whereas replications seek to verify whether previous research findings are reproducible and, therefore, accurate. This entry reviews various conceptions of replication, discusses replication in education research and other fields, and explores the implications of replications with regard to scientific rigor.

Conceptions of Replication

Replication is typically broken into two primary categories: direct and conceptual. Direct replications follow the original study’s methods (e.g., similar participants, measures, and procedures) as closely as possible in order to test whether the original research results can be obtained again. Conceptual replications purposefully alter at least one component of the original study in order to test the underlying construct of interest. Replications serve as a critical part of the scientific enterprise by helping control for issues caused by sampling error, artifacts, fraud, generalizability, testing the general underlying hypothesis of a previous study, or some combination of those issues.

Conceptual replications can help identify the degree to which a particular finding is broad or narrow. A reanalysis of previously collected data is not a replication but is still a vital part of the scientific process. Reanalysis of data can reveal things ranging from simple calculation errors to more fundamental problems such as data manipulation and fraud, and reanalysis can provide new information by reanalyzing seminal data sets using modern statistical techniques.

There is not universal agreement about the necessary and sufficient features of a replication, despite its being one of the basic building blocks of science. A 2009 review by Stefan Schmidt connects replication theory with replication in practice. Schmidt lists five-function replications that serve to control for sampling error, to control for artifacts, to control for fraud, to generalize to different or larger populations, or to assess the general hypothesis of a previous study. Rather than deliberately avoiding the original methods, Schmidt suggests systematically changing individual facets of the original study to better understand its nature.

The relative importance of direct and conceptual replications has been debated. Some scholars argue that conceptual replication should be emphasized, whereas others support direct replications. The importance of each depends on the goal of the investigation, with direct replication typically seeking to verify or corroborate the original findings using the same methods as the original researchers; conceptual replications test more general models and theories. However, it is important to note that only direct replications can disconfirm or corroborate previous claims. This is because a failed conceptual replication does not automatically identify a flaw in the original study but instead has the potential to identify the generalizability (or lack thereof) of the original finding. Direct replication can help identify potential biases in the original study or confirm that the original finding was not an anomaly. Because of this, some scholars argue that direct replication should always precede conceptual replication attempts.

Replication research can help identify, diagnose, and minimize many methodological biases. Despite the benefits that replication brings to the research table, conducting replications is largely viewed in the social sciences research community as lacking prestige, originality, or excitement, a bias that is not community as lacking prestige, originality, or excitement, a bias that is not always shared in the natural sciences. Several publications have begun to discuss the hurdles and disincentives to conduct replications that appear to be endemic to the social science research infrastructure:

> *Submission bias*—Conducting research and submitting for publication is time-consuming, and investigators may purposefully remove replications from the publication process to focus on other projects or because they believe replications cannot be published.
>
> *Funding bias*—Research, especially an experimental study, requires resources, making replications difficult to conduct in the absence of external funding. Yet the major research funding agencies rarely fund replication studies.
>
> *Editor/reviewer bias*—Journal editors and reviewers may be more likely to reject replications, driven by a belief that replications are not as important or prestigious as nonreplication articles.
>
> *Journal publication policy bias*—Journals may have explicit policies against publishing replications.
>
> *Hiring bias*—Institutions may not hire researchers who conduct replications, with funding and editor/reviewer biases possibly playing a role in these decisions.
>
> *Promotion bias*—Organizations may not value replication research to the same extent as research perceived to be "new" within promotion and tenure processes.
>
> *Journal-analyzed bias*—Previous research analyzing replication rates may have selected journals that publish few replications. Because each journal has its own editorial policies, it may be that some journals are more likely to accept replications than others.
>
> *Novelty equals creativity bias*—Editors, reviewers, and researchers value creative contributions, but novelty and creativity are not synonymous. Most definitions of creativity and innovation propose criteria of novelty and utility; a novel result that cannot be replicated is by definition not useful and, therefore, not creative.

These biases may not uniformly deny publication of replications, but they are widely understood to impede the process, thereby discouraging replications before they are even initiated. Oddly, these biases exist even though the call for replications has existed for generations. Nonetheless, these limitations and impediments are being more widely discussed by social scientists, including education researchers, indicating that change may be on the horizon.

Replication in Other Fields

Although concern over replication exists in many fields, including biology, medicine, and marketing, it has received the most attention of late in psychology. This is due to a number of factors, including high-profile cases of fraud within psychological research and seminal studies failing to replicate. Regardless of the causes, psychologists have been discussing the need for more frequent replication of key research findings and debating how such replications should be conducted.

Data on replication rates are available for a few scientific fields. Research provides evidence that just over 1% of publications in the top 100 psychology journals were replications, although the rate after the turn of the 21st century has doubled to roughly 2%. Less than 10% of psychology replications failed to replicate previous findings, a better success-of-replication rate than in other fields, such as medicine.

One possibility for the dearth of replications is that many research studies include some form of replication, but as previously discussed, the perceived bias against such articles during the journal review process could encourage scholars to mask the true nature of their studies. There is evidence to support this concern: A survey of social science editors found that over half reported that being a replication contributes to being rejected for publication.

Replications in Education Research

A study of the publication histories of leading education journals found that less than one-quarter of 1% of articles were labeled replications, substantially lower than the replication rates of other domains. Contrary to previous findings in medical fields, but similar to psychology research, two-thirds of education replications successfully replicated the original studies. However, replications were significantly less likely to be successful when there was no overlap in authorship between the original and replicating articles. This difference raises questions regarding potential biases in replicating one’s own work and may be related to previous findings of questionable research practices in the social sciences. However, same-author replications could merely be benefiting from the wisdom or experience of the author having done the study previously and thus may be able to more closely replicate the original methods. In special education, the rate of replications in the top journals was half of 1%, with over education, the rate of replications in the top journals was half of 1%, with over 80% of these studies successfully replicating previous findings. But again, replications where there was at least one author overlapping with the original article were statistically significantly more likely to find successful results.

Replication and Rigor of Research

Given such low replication rates, the need to increase replications is apparent and permeates all levels of education research. Researchers and practitioners cannot know with sufficient confidence that an intervention works or that an effect exists until it has been directly replicated, preferably by independent researchers.

Replications are merely one of numerous solutions (e.g., reanalysis of data, preregistering hypotheses, and publishing studies of interest rather than results of interest) for increasing the rigor of research. But replication is critical for the development of the field because although fraud and error may eventually be revealed in an absence of replication, in the intervening time, research projects and educational policies could be based on faulty results. Replication can help identify errors more quickly while also potentially helping verify and clarify previous results, leading to a more effective and more respected body of education research.

Replication will not solve all problems concerning rigor, reliability, precision, and validity of education research. However, if education research is to be relied upon to develop sound policy and practice, then conducting replications on important findings is essential to move toward a more reliable and trustworthy understanding of educational environments. Although potentially beneficial for the individual researcher, an overreliance on large effects from single studies can weaken the field as well as the likelihood of effective, evidence-based policy. Direct replication of important educational findings can lead to stronger policy recommendations while also making such recommendations more likely to improve education practice and, ultimately, the lives of children.

Jonathan A. Plucker and Matthew C. Makel

See also Ethical Issues in Educational Research; Falsified Data *in Large-Scale Surveys*; File Drawer Problem; Meta-Analysis; Representativeness; Scientific Method; Threats to Research Validity; Trustworthiness

Further Readings

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 43(6), 304–316.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? Perspectives on Psychological Science, 7(6), 537–542. doi:10.1177/1745691612460688.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology, 13(2), 90–100.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Justin Jager Justin Jager Jager, Justin

Representativeness

Representativeness

1412

1413

Representativeness

This entry discusses the concept of representativeness and its importance in social and behavioral science research. In conducting such research, it is desirable but usually not possible to study an entire population; in consequence, researchers sample from the population and then generalize findings of the research based on the sample to the population. The analytic sample is thus supposed to *represent* the population. It is important to know whom a study sample represents to evaluate implications of the study and findings. Who is represented in a study's sample frames the study.

Two Main Types of Representativeness

The best samples are called probability samples; these incorporate some form of random selection into their sampling procedure, so it is straightforward to discern whom the sample represents. By design, a probability sample is a random sample of some specified target population, and researchers (as well as consumers of the research) can be confident that findings generated by a probability sample represent or generalize to that specified target population. By contrast, convenience samples entail collecting data from samples in an ad hoc or "first-come, first-served" basis. Such samples usually do not represent some larger population, and so it is not possible to apply a convenience sample's findings to a broader population beyond the specific individuals sampled.

Depending on their target populations, probability samples vary in the scope or degree of representativeness. Consider two probability samples: one of adults aged 55 or older who reside in New York City and another aged 55 or older who reside in the United States. Because both studies are based on probability samples, determining whom each study’s findings represent is clear and straightforward. However, the probability sample of New York City represents a narrower and more restricted population, whereas the probability sample of the United States clearly represents a broader and more diverse population.

Threats to Representativeness

There are different kinds of threats to representativeness. Convenience sampling is one. Attrition is another. Attrition occurs when respondents drop out of a study and/or are lost over time. Attrition is common with longitudinal designs and is a threat to a study’s representativeness because attrition is typically nonrandom (certain types of study participants, such as those from lower socioeconomic backgrounds, are more likely to drop out of a study than others).

To adjust for the effects of attrition, researchers can take advantage of “missing data” approaches.

Sample Weights

To ensure that a sample represents a diverse population and that sufficient numbers are included in a sample for the purpose of statistical power, researchers sometimes purposefully oversample a subpopulation. Then “sample weights” are applied to the data that result in parameter estimates (i.e., estimates of totals, proportions, and associations) that render the data representative of the population from which the sample was recruited.

Marc H. Bornstein and Justin Jager

See also Convenience Sampling; Longitudinal Data Analysis; Validity; Weighting

Further Readings

Bornstein, M. H., Jager J., & Putnick, D. L. (2013). Sampling in developmental science: Shortcomings and solutions. Developmental Review, 33(4), 357–370.

Davis-Kean, P. E., & Jager, J. (2012). The use of large-scale data sets for the study of developmental science. In B. Laursen, T. Little, & A. Card (Eds.), Handbook of developmental research methods. New York, NY: Guilford.

Davis-Kean, P. E., Jager, J., & Maslowsky, J. (2015). Answering developmental questions using secondary data. Child Development Perspectives, 9, 256–261.

Enders, C. K. (2013). Dealing with missing data in developmental research. Child Development Perspectives, 7(1), 27–31.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual Review of Psychology, 60, 549–576.

Jana Craig-Hare Jana Craig-Hare Craig-Hare, Jana

Research Proposals

Research proposals

1413

1414

Research Proposals

Research proposals are written to propose a research project and oftentimes request funding, or sponsorship, for that research. The research proposal is used to assess the originality and quality of ideas and the feasibility of a proposed project. The goal of the research proposal is to convince others that the investigator has (a) an important idea; (b) the skills, knowledge, and resources to carry out the project; and (c) a plan to implement the project on time and within budget. This entry discusses the process of developing a research proposal and the elements of an effective proposal.

For a graduate student, a research proposal may be required to begin the dissertation process. This serves to communicate the research focus to others, such as members of the student's dissertation committee. It also indicates the investigator's plan of action, including a level of thoroughness and sufficient detail to replicate the study. The research proposal could also be considered as a contract, once members of the committee agree to the execution of the project.

To be considered for funding, research proposals may be solicited or unsolicited. Solicited proposals respond to a request with specified requirements from a request for proposals (RFP). Unsolicited proposals are submitted without a prior request. A letter of intent may be required or requested so that potential funders can review a brief abstract prior to the submission of a full proposal. A letter of intent assists the funding agency with gauging interest in the RFP and helps to identify the number of reviewers they may need to review the research proposals.

Requirements may vary for research proposals. Educational research proposals generally follow the same format as a research paper or journal article, including an abstract, introduction, literature review, method section, and conclusion. Investigators need to review the requirements for the targeted RFP carefully and follow the outline provided in the sponsor’s guidelines. A research proposal has to clearly and concisely identify the proposed research and its importance. The background literature should support the need for the research and the potential impact of the findings.

The method section proposes a comprehensive explanation of the research design, including subjects, timeline, and data analysis. Research questions should be identified as well as measurement instruments and methods to answer the research questions. Proposals for research involving human subjects identify how the investigators will protect participants throughout their research project. A proposed budget identifies and justifies the financial support needed for the proposal, including personnel, supplies, and indirect expenses, as well as institutional resources to demonstrate the capacity to successfully accomplish the proposed research.

Proposals often require engaging in an external review either by an external evaluator or advisory board consisting of expert consultants in the field. References are included to provide documentation about the supporting literature identified in the proposal. Appendixes and supplemental materials may also be included, following the sponsoring organization’s guidelines. As a general rule, educational research proposals follow the American Psychological Association formatting guidelines and publishing standards. If funding is being requested, it is important for the proposal to identify how the research will benefit the sponsoring organization and its constituents.

Funding for educational research proposals is often provided by federal agencies such as the Institute of Education Sciences and the National Science Foundation. Educational foundations also support research proposals for areas in which they are interested in promoting or furthering their research agenda.

Funding organizations and foundations typically provide a proposal writing guide and/or detailed RFP. In addition, often webinars are provided to review the RFP requirements and ask questions of the project officer and/or sponsoring organization. Although these webinars may be archived for later review, project officers will often be available for additional questions and clarification. Investigators should collect as much information as possible about the funding organization or foundation, including their funding priorities and current RFPs organization or foundation, including their funding priorities and current RFPs, before submitting a research proposal for consideration.

The success of a research proposal depends on both the quality of the project and its presentation. A proposal may have specific goals, but if they are neither realistic nor desirable, the probability of obtaining funding is reduced. Similar to manuscripts being considered for journal articles, reviewers evaluate each research proposal to identify strengths and criticisms based on a general framework and scoring rubric determined by the sponsoring organization. Research proposals that meet the scoring criteria are considered for funding opportunities. If a proposal does not meet the scoring criteria, revisions may be necessary before resubmitting the proposal to the same or a different sponsoring organization.

Common mistakes and pitfalls can often be avoided in research proposal writing through awareness and careful planning. In an effective research proposal, the research idea is clearly stated as a problem and there is an explanation of how the proposed research addresses a demonstrable gap in the current literature. In addition, an effective proposal is well structured, frames the research question(s) within sufficient context supported by the literature, and has a timeline that is appropriate to address the focus and scope of the research project. All requirements of the sponsoring organization, including required project elements and document formatting, need to be met within the research proposal. Finally, an effective proposal is engaging and demonstrates the researcher’s passion and commitment to the research addressed.

Jana Craig-Hare

See also Abstracts; APA Format; External Evaluation; Human Subjects Protections; Institute of Education Sciences; Journal Articles; Literature Review; Methods Section; National Science Foundation; Rubrics

Further Readings

Creswell, J. W. (2013). Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.). Thousand Oaks, CA: SAGE.

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (2007). Proposals that work. Thousand Oaks, CA: SAGE.

Pzreworski, A., & Salomon, F. (1995). On the art of writing proposals. Brooklyn, NY: Social Science Research Council.

Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2012). When to use what research design. New York, NY: Guilford Press.

Maciej Taraday Maciej Taraday Taraday, Maciej

Anna Wieczorek-Taraday Anna Wieczorek-Taraday Wieczorek-Taraday, Anna

Residuals

Residuals

1414

1417

Residuals

Residual is a term with several closely related meanings that occur in the fields of business, finance, optimization, and statistics. Here, it will be only considered as a statistical term. In statistics, residuals are differences between observed values and values predicted on the basis of a statistical model.

This entry presents what residuals are in the examples of continuous (regression analysis) and categorical (contingency tables analysis) data. It is important to understand what residuals stand for, and what their properties are, as they are part of every statistical model used in educational research. Moreover, a wide repertoire of statistical models has specific assumptions about distribution of residuals that has to be met, so that the model represents unbiased relationships between variables. Besides the diagnostic of a statistical model, residuals can be used to identify unusual observations (anomalies) or to indicate which category occurs less often or more often than expected.

Residuals should not be confused with statistical error, which is the amount by which observations are different from their expected value based on the whole population (this quantity cannot be observed directly). Residuals refer to the amount by which observations are different from the sample mean. Therefore, residuals usually are treated as the estimates of statistical error.

Common Types of Residuals

Ordinary residuals are expressed on the scale of the variable for which they are being computed. Let us assume that a person's height (195 cm) residual is to be computed. The expected value of the height for each person in the sample is equal to the mean height of the sample (170 cm, standard deviation equals 8). That means that the residual is equal to 25 cm (195 – 170 = 25). Frequently, it is more convenient to use another type of residual, depending on the purpose of analysis.

Standardized residuals are raw residuals transformed to the so-called standard score (also called *z* score) and are useful in identifying observations that are not typical. Whenever a value of the variable for which residuals are being computed comes from normal distribution, standardized residuals inform whether the observation is usual or not. If the value of a standardized residual is below −2.58 or above 2.58, such an observation is treated as an anomaly. It means that observations with such a residual represent not more than 1% of the population. A standardized residual of a person's height is equal to 3.12 (25 cm divided by 8 cm—the length of a standard deviation). A person with the height of 195 cm appears rarely in the population (assuming that values of height come from a normal distribution).

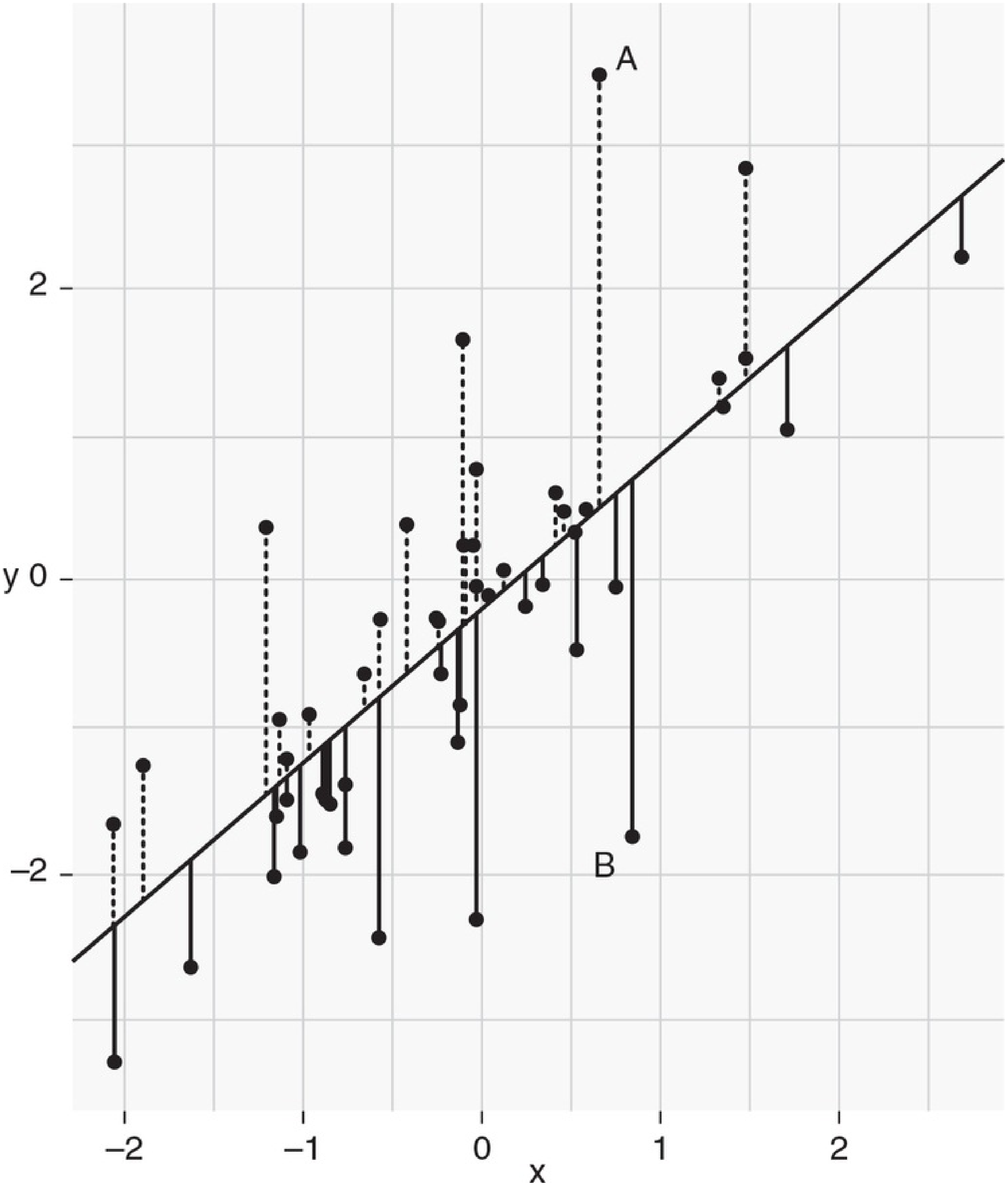
Studentized residuals are especially useful in cases of multiple linear regression. In contrast to standardized residuals, they are more robust for anomalies.

Residuals in Linear Regression

Let us see what residuals look like in the case of a simple linear regression.

Figure 1 is a typical example of a positive correlation between two variables. Values of variable *Y* can be expressed in a regression model as a linear combination of intercept (which in this case is equal to zero) and slope times variable *X* and residuals (*Y* = intercept + slope × *X* + *E*, residuals are usually marked by *E*).

Figure 1 This scatterplot depicts the relationship between variable *X* and variable *Y* (scatterplot of variables *X* against *Y*). The higher the value of observation on the horizontal axis (variable *X* called regressor), the higher the value on the vertical axis (variable *Y* called regressand)



Predictions of a model (predicted values) are depicted by the solid gray line (regression line). For each value on the *x* axis, the model predicts specific value on the *y* axis. In reality, there is always a discrepancy between predictions of a model and the observed values. These discrepancies are denoted by a solid or dashed line perpendicular to the horizontal axis. These lines represent residuals.

About half of the observations have underestimated predicted value—they are located over the regression line—and their residuals have positive values (dashed lines). Others have overestimated predicted value—those observations are located below the regression line—and their residuals are negative (solid lines). Residuals are interpreted as the part of variance of the regressand that cannot be explained or predicted on the basis of the regressor. As a matter of fact, it can be seen in Figure 1 that residuals are perpendicular to the regressor variable. Because of that, it is impossible to tell anything about the residuals' value by changing the value of the regressor. It is equally likely that observations with the same value on the *x* axis have positive or negative residuals. It should be noted that observations "A" and "B" have almost the same value on the *x* axis and the opposite value of residuals.

Residuals in linear regression are used for model diagnostics. Linear regression has important assumptions about the error in the model. In practice, these assumptions are tested by the analysis of residuals.

First, linear regression assumes linearity between the regressand and regressors. It means that the mean value of the regressand is a linear combination of slopes and regressors. This assumption can be tested by visual inspection of the scatterplot depicting predicted values against residuals. A random pattern of points indicates that a linear model decently fits the data. When the linearity assumption is violated, it is possible to observe a nonrandom pattern of points.

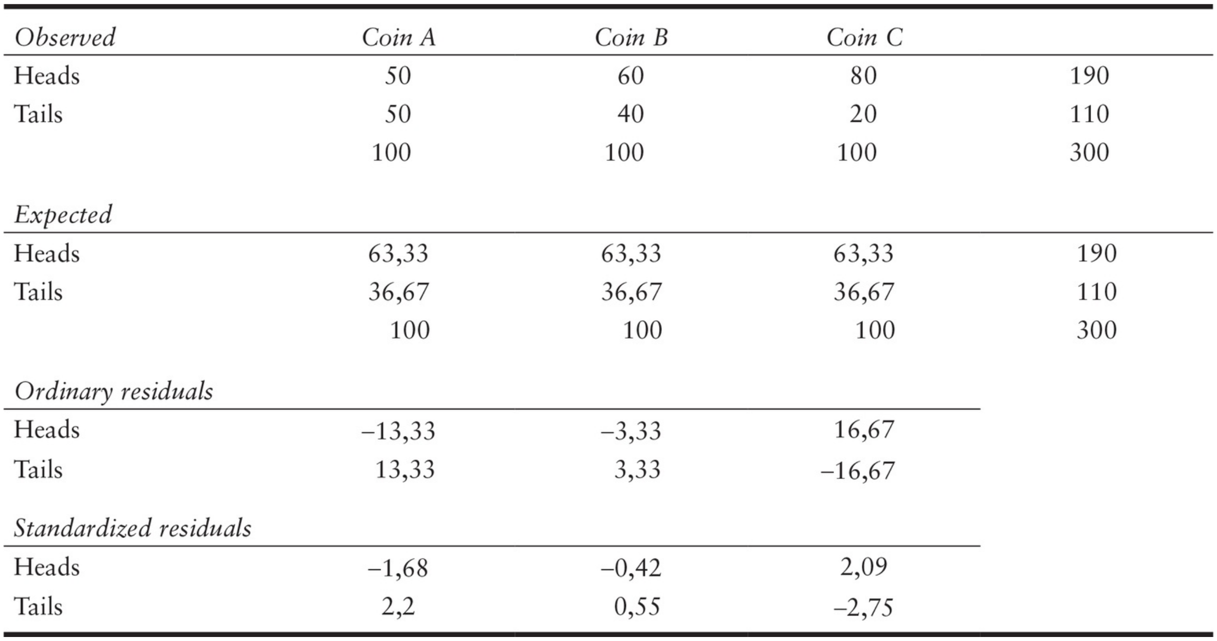
Second, it is assumed that error in the model is random, which means that residuals come from a standardized normal distribution. The mean value of residuals is close to 0 and the distribution of residuals is symmetrical over the mean value. This assumption can be tested by the visual inspection of a histogram or by using a statistical test that detects deviations from normality (e.g., Shapiro–Wilk test of normality or Kolmogorov–Smirnov test). However, linear regression is resistant to minor deviations from normality.

Third, the linear regression model assumes constant variance of error. Regardless of the value of the regressor, errors have constant variance. This assumption is violated whenever any correlation between predicted values and residuals is observed. Heteroscedasticity (i.e., inhomogeneous variance of residuals) can be observed on the scatterplot of predicted values against residuals. The pattern of points usually takes a triangular shape when an assumption of constant variance of error is violated.

Fourth, the linear regression model assumes independence of errors, which in terms of residuals means lack of autocorrelation between residuals. Autocorrelation of residuals is a situation encountered in repeated measurements when each person is tested several times. In such a situation, the previous measurement from the same person provides some information about the result of the next measurement. Then, the value of the residual for each person can be predicted on the basis of another measurement from the same person. Autocorrelation of residuals results in low goodness of fit of the regression model.

Residuals in Contingency Table Analysis

Residuals are also useful to identify categories that appear more or less frequently than expected in case of categorical data analysis. Let us assume that the goal of the analysis is to identify an “unfair” coin through tossing. Each coin has been tossed 100 times. To answer the question whether the result of tossing is independent from the coin, a Pearson’s χ^2 test can be applied. On the basis of the probability statistical value ($\chi^2 = 13.396$, $df = 2$, $p < .01$), null hypothesis—stating that the result of tossing is independent of the coin—should be rejected. After rejecting the null hypothesis, we are aware that one or more than one coin is “unfair.” Standardized residuals help to identify which coin or coins are unfair. Table 1 presents the results of the coin tossing.



Observed	*Coin A*	*Coin B*	*Coin C*	
Heads	50	60	80	190
Tails	50	40	20	110
	100	100	100	300
Expected				
Heads	63,33	63,33	63,33	190
Tails	36,67	36,67	36,67	110
	100	100	100	300
Ordinary residuals				
Heads	–13,33	–3,33	16,67	
Tails	13,33	3,33	–16,67	
Standardized residuals				
Heads	–1,68	–0,42	2,09	
Tails	2,2	0,55	–2,75	

In the case of Coin A, exactly the same number of heads and tails have appeared (50/50). There were slightly more heads than tails in Coin B (60/40), and also in the case of Coin C prevalence of heads has been observed (80/20). To identify the “unfair” coin, the values of the standardized residuals need to be checked. To do so, the predicted value must be computed. Predicted values in contingency tables are computed for every bracket. A predicted value for a bracket is equal to the sum column multiplied by the sum row and divided by the total number of observations (for A heads, it is 100 × 190/300 = 63.33). Ordinary residuals are the difference between the observed and predicted value (for A heads, it is $50 - 63.33 = -13.33$), and standardized residuals are ordinary residuals divided by the square root of the predicted value (for A heads, it is $(50 - 63.33)/\sqrt{63.33}$), which is −1.68). Brackets with a value of the standardized residuals lower than −2.58 or higher than 2.58 are identified as categories with frequency significantly deviating from randomness. In this example, Coin C has standardized residuals for tails lower than −2.58 (it is equal to −2.75), which leads to the conclusion that in Table 1, the result of tossing is not independent from the coins because of Coin C. This simply means that Coin C is different from the other two and can be identified as unfair.

Maciej Taraday and Anna Wieczorek-Taraday

See also Multiple Linear Regression; Variance; **Z** Scores

Further Readings

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press.

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York, NY: Chapman & Hall.

Ellenberg, J. H. (1973). The joint distribution of the standardized least squares residuals from a general linear regression. Journal of the American Statistical Association, 68(344), 941–943.

Haberman, S. (1973). The analysis of residuals in cross-classified tables. Biometrics, 29(1), 205–220.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction. Boston, MA: Wadsworth.

Andrea M. Garcia Andrea M. Garcia Garcia, Andrea M.

Resilience

Resilience

1417

1420

Resilience

Resilience is a process that allows people to adapt in a constructive way to threats and adverse experiences. The construct of resilience is complex and multidimensional. Research in this area has focused on the developmental process of resilience. However, debates remain concerning how to define resilience and how to best measure adaptive outcomes after exposure to hardship. Although resilience can be understood through a variety of theoretical models, the underlying theme among them is that resilience is a process and not a personal attribute that some people have and others don't.

This entry describes major developments of resilience from the field of clinical child and adolescent psychology. It begins by discussing the competing definitions of resilience and common problems with operationalizing and measuring resilience. It then discusses the developmental process of resilience in order to provide the unique properties of resilience and the protective and risk factors related to resilience.

Difficulties Operationalizing Resilience

Since the 1970s, resilience has been an area of research that moved away from the traditional medical model, which focused on clusters of symptoms, to focusing on holistic perspectives. Early pioneers focused on children's personality characteristics (e.g., charisma and autonomy); however, as the work in the area evolved, there have been variations in definitions and use of terminology. Differences in the conceptualization of resilience revolve around the contextual circumstances of adversity and the process of positive adaptation.

Because of these differences, various resilience models (e.g., social–ecological) and study designs (longitudinal vs. cross-sectional and person vs. variable based) have emerged.

The major differences between definitions of resilience include judgments concerning the operationalizing of adverse condition(s), which may range from a single event to chronic adverse events. Similarly, positive adaptation requires judgments about children's and adolescents' competence in particular domains (social, problem solving, and autonomy) and other proximal predictors such as family and social environment (school, peers, and community) characteristics. Given that resilience is a dynamic process that involves fluid predictive factors, a unique perspective to help examine such fluctuations is a unique developmental perspective, a developmental psychopathology process. As such, it is important to understand that resilience is a continuous variable and not a dichotomous attribute that is or is not present but a dynamic process for which no one factor promotes or inhibits resilience.

Developmental Psychopathology Perspective of Resilience

An important consideration is that children and adolescents and their families are influenced by different systems; therefore, it is important to examine resilience through a developmental psychopathology perspective. Developmental psychopathology is a developmental perspective on the etiology of mental disorders that begin during childhood and adolescence and uses a multidisciplinary conceptual approach. Developmental psychopathology is not just the study of disorders but also helps describe the developmental process between maladaptive and adaptive processes and the extent to which they influence developmental outcomes.

Knowledge of divergent pathways can inform researchers about particular adaptive or maladaptive processes. For example, an individual may be resilient in a particular context or experience but not in others. Furthermore, developmentally, characteristics of resilience may change across the life span. Therefore, resilience is best understood as a process.

Measuring Resilience

In terms of measuring resilience, the core objective of resilience research is to identify risk and protective factors that may ameliorate the detrimental effects of adverse conditions and to learn the extent to which mechanisms (protective or risk factors) of interest facilitate that relationship. Given that risk factors inhibit while protective factors promote resilience, it is important when measuring and evaluating resilience that researchers clearly define what aspects of the resilience researchers are evaluating.

Two common quantitative data analysis approaches to evaluate protective and risk processes in resilience are variable and person based. The variable-based approach evaluates the extent to which individual characteristics such as high IQ, problem-solving skills, and coping mechanisms are protective or vulnerability factors in the process of resilience. Although individual characteristics such as high IQ are unique to that individual, they should not be thought of as independent of the context of the individual.

The person-based approach focuses on individual and environmental differences that would suggest high-versus low-risk conditions of experiencing adversity and diverse adaptive profiles. Further, studies using variable-based or person-based approaches have used various methods to evaluate resilience, such as checklists, surveys, scales, and interviews; however, factors of interest have been operationalized a priori and evaluated through standardized measures. A few considerations to be mindful about are that each approach (i.e., person and variable) yields different perspectives and insights. Interpretation of results in variable-based or the person-based approaches is best understood as part of the process and not representing a personal attribute. Additional considerations include accounting for time of assessment, developmental systems assessed (family, individual, and community), and individual variation in responses based on contexts (e.g., child placement with caregiver).

Protective and Risk Factors of Resilience

Research on resilience in children and adolescents has frequently focused on populations that have experienced child maltreatment and pediatric conditions (e.g., children with cancer, diabetes, and chronic pain) and their families. Several protective and risk factors have emerged and include unique adaptive and maladaptive processes in individual, family, and social dynamics. Common individual dynamics include effective coping strategies, intelligence, self-efficacy, and optimism. Family dynamics include family cohesion, family stability, family connectedness, and responsive parenting. Social dynamics include peer and social support from friends, family, teachers, community, or church members.

Within schools, dynamics that foster resilience include having positive relationships with adult mentors, positive peer influence, and student engagement. For students, a positive relationship with teachers that involves appropriate expectations of the students and an inclusive, stable, and stimulating environment are helpful in promoting resilience.

Resilience is also promoted through school engagement, which refers to the student’s emotional attachment and beliefs related to the value of education. Promoting student engagement involves understanding the contextual factors that influence resilience. For example, understanding a student’s cultural values and community connections has been found to influence school engagement, with increased school engagement in turn promoting increased resilience. Similarly, engaging in extracurricular activities helps to promote a greater sense of belongingness, which promotes retention in the school and greater opportunities for positive adult mentors. Further, having positive peer influences such as peers who take initiative to pursue higher education or work opportunities can promote resilience, whereas having negative experiences such as bullying or engaging with antisocial peers can influence greater risk-taking behaviors and potentially limit positive educational outcomes. Given that youth may experience a variety of adverse events throughout their lives, the school can provide supports that promote resilience within each student.

Overall, resilience is a balance of protective and risk factors in which the outcomes for an individual depend on the unique maladaptive or adaptive process in the presence of adversity. The field of clinical child and adolescent psychology conceptualizes resilience as a continuous construct that is best understood through a multidisciplinary approach that considers the developmental process and context of the etiology of resilience among children and adolescents.

Andrea M. Garcia

See also Adolescence; Childhood; Emotional Intelligence; Erikson’s Stages of Psychosocial Development; Puberty

Early Readi

Further Readings

Cassen, R., Feinstein, L., & Graham, P. (2008). Educational outcomes: Adversity and resilience. Social Policy & Society, 81(1), 73–85.

Cicchetti, D. (2016). Social emotional, personality, and biological development: Illustrations from a multilevel developmental psychopathology perspective on child maltreatment. Annual Review of Psychology, 67, 187–211.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59–109.

Luthar, S. S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: A critical evaluation and guidelines for future work. Child Development, 71, 543–562.

Luthar, S. S., Crossman, E. J., & Small, P. J. (2015). Resilience and adversity. In R. M. Lerner & M. E. Lamb (Eds.), Handbook of child psychology and developmental science (7th ed., Vol. 3, pp. 247–286). New York, NY: Wiley.

Masten, A. S., & Cicchetti, D. (2016). Resilience in development: Progress and transformation. In D. Cicchetti (Ed.), Developmental psychopathology, volume 4, risk, resilience, and intervention (3rd ed.). New York, NY: Wiley.

Southwick, S. M., Bonanno, G. A., Masten, A. S., Panter-Brick, C., & Yehuda, R. (2014). Resilience definitions, theory, and challenges: Interdisciplinary perspectives. European Journal of Psychotraumatology, 5. doi:10.3402/ejpt.v5.25338

Ungar, M., Ghazinour, M., & Richter, J. (2013). Annual research review: What is resilience within the social ecology of human development? Journal of Clinical Psychology and Psychiatry, 54(4), 348–366.

Zolkoski, S., & Bullock, L. (2012). Resilience in children and youth: A review. Children and Youth Services Review, 34, 2295–2303.

Parvati Krishnamurty Parvati Krishnamurty Krishnamurty, Parvati

Response Rate

Response rate

1420

1422

Response Rate

A response rate is the ratio of the number of participants in a study to the number of participants who were asked to participate. Several formulas have been developed to calculate response rates, which are based on different definitions of what it means to have fully participated and how to count eligible units. Response rates are commonly used to measure data quality, and low response rates could result in nonresponse bias. Response rates are therefore an important measure for education surveys, and low response rates could potentially impact the validity of estimates, analysis, and inference in education research. Although response rates can be calculated for designs involving a variety of methodologies, the term usually refers to the level of participation in survey or interview research, so that context is the focus of this entry.

Reasons for Nonresponse

There are many possible reasons for nonresponse. People tend to refuse to participate in surveys or interviews, for example, due to lack of time, lack of interest in the topic, competing demands from many requests for their attention, suspicion that the survey request is actually a marketing pitch, or the sensitive nature of the questions asked.

Each mode of collecting survey data has its own challenges. For in-person surveys, interviewers are often unable to reach respondents who live in gated communities or high security apartment buildings. The major challenge for telephone surveys is the difficulty of finding people at home and devices like answering machines, caller ID, and cell phones. Mail surveys, which have some of the lowest response rates, face several obstacles, including requests being ignored due to an increasing volume of junk mail, the lack of personal contact with an interviewer, and the length of the survey on paper. Web surveys are easy to decline, and most types of web surveys (with the exception of web panels) do not have probability-based samples, which limit their usefulness for research.

Calculating Response Rates

The prevailing standards for response rate calculations are brought out and updated by the American Association of Public Opinion Research (AAPOR). AAPOR recognizes six different methods of calculating response rates. Before discussing the details, it is helpful to understand the basic terminology used.

- *Completed interviews* are cases in which the sample unit (e.g., household, person, business) was contacted and the interview was completed.
- *Partial interviews* are interviews terminated by respondents and left incomplete.
- *Noninterviews* occur when a respondent was located but did not complete the interview. This category includes refusals, noncontacts, and other types of noninterviews. Refusals happen when the eligible respondent refuses to participate. Noncontacts include other situations in which contact cannot be made, such as the respondent never being available, answering machines, inability to gain access to a building, or completed questionnaire not returned (for mail surveys). Other includes other kinds of noninterviews whereby contact was made and there was no refusal, but the survey could not be administered, for example, if there is a deceased, mentally or physically challenged respondent, or language problems.
- *Unknown* cases are those in which the researcher is not certain whether the sample element is eligible for the survey. Unknown cases estimated to be eligible are the sum of two categories—unknown households and unknown other. Unknown households include cases in which the researcher does not know if the sample element is a housing unit, for instance, when the phone was always busy, there was no answer or an answering machine, or if the address was not locatable or unsafe to reach. Unknown other includes various categories of returned mail from the postal service (for mail surveys) and in screening studies (i.e., when there is a screener before the actual interview is conducted), the screener was not completed.

The AAPOR response rates are as follows:

RR1 = Completes / ((completes + partials) + (refusals + noncontacts + other) + (all unknown cases)).

RR2 = (Completes + partials)/((completes + partials) + (refusals + noncontacts + other) + (all unknown cases)).

RR3 = Completes/((completes + partials) + (refusals + noncontacts + other) + (unknown cases estimated to be eligible)).

RR4 = (Completes + partials)/((completes + partials) + (refusals + noncontacts + other) + (unknown cases estimated to be eligible)).

RR5 = Completes/((completes + partials) + (refusals + noncontacts + other)).

RR6 = (Completes + partials)/((completes + partials) + (refusals + noncontacts + other)).

Of the six AAPOR response rates, RR1 through RR6, the odd-numbered rates count only completed interviews in the numerator, while the even-numbered ones include both completed and partial interviews in the numerator. RR1 and RR2 are the most conservative or lower bound response rates based on assuming that all the unknown cases were eligible. RR5 and RR6 are the upper bound or least conservative response rates based on the assumption that none of the unknown cases were eligible. RR3 and RR4, which are more commonly used in surveys, are based on the assumption that some proportion of the unknown cases was eligible. For RR3 and RR4, it is important to justify that assumptions made about the proportion of the unknown cases are eligible for the survey.

Screening studies have a slightly different calculation for response rates because they can have two levels of eligibility. For instance, in a survey of respondents under age 18 years in a household, there has to be a household in the location and then there also has to be an eligible child in the household.

Response Rates for Different Modes

Different study designs have slightly different issues to consider while calculating response rates. Mail surveys and in-person surveys have very straightforward calculations. However, for dual frame random digit dialing designs that include both landline and cell phones, response rate calculations are similar to those for screening studies. For web surveys, which are based on opt-in, nonprobability, or river samples, it is not appropriate to calculate a response rate. Calculations of response rates for Internet surveys of listed persons or probability-based Internet panels are more complex. In multimode surveys such as surveys conducted by phone and in person, the response rate calculation is always more complicated because there has to be a hierarchy of codes to determine the final disposition of a case.

How to Improve Response Rates

In practice, several techniques are used in designing the survey to improve or maximize response rates. Incentives paid to respondents, particularly prepaid monetary incentives, have been found to be effective in increasing response rates. Other techniques include sending advance letters notifying the household or individual that they are invited to participate, training interviewers and staff to be more persuasive, designing survey materials to be clear and visually attractive, making multiple contact attempts, and using various techniques for converting initial refusals. There is no consensus in the survey methodology literature on what is a minimum acceptable response rate, and response rates vary greatly by mode and type of survey. Response rates below 80% usually require some analysis of nonresponse to check for nonresponse bias.

Parvati Krishnamurty

See also Nonresponse Bias; Sample Size

Further Readings

The American Association for Public Opinion Research. (2016). Response rate calculator v4.0. Author. Retrieved from http://www.aapor.org

The American Association for Public Opinion Research. (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys (9th ed.). Author.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). Survey nonresponse. New York, NY: Wiley.

Daryl F. Mellard Daryl F. Mellard Mellard, Daryl F.

Response to Intervention

Response to intervention

1422

1426

Response to Intervention

Response to intervention (RtI) is an organizational framework for guiding instructional and curricular decisions to prevent students' academic and behavioral difficulties. RtI is not a curriculum and is not a program (like a reading program or a dropout prevention program). RtI models the public health approach for preventing and treating conditions and has increasingly specific and intensive interventions for students who are encountering academic and behavioral difficulties in school. The foundation for RtI was built from early reading research that demonstrated that specific components of academic domains were predictive of future achievement in that domain (e.g., reading, mathematics, and behavior). Research in early intervention addressed students' component deficits and improved those students' learning trajectories and achievement. Thus, predicted failure was averted and students progressed with their peers. This entry identifies intended outcomes of high-quality implementation, a review of RtI's four components, cautions of implementation, and challenges for school staff in bringing RtI to scale.

Intended Outcomes

Proponents of RtI propose four valued qualities. First, RtI incorporates a predictive approach. RtI's procedures can inoculate students against encountering academic and behavioral difficulties. This outcome is achieved by using screening measures in a predictive manner and identifying students who are predicted as at-risk for academic and behavioral difficulties. Those students predicted as at-risk are provided an intensive intervention that addresses their skill and ability deficits before they lag their peers.

The second-valued quality is that appropriate interventions are provided to students in a timely, data-based, and targeted manner. Rather than students failing, singling them out from their peers, and then developing interventions for them, the results of targeted assessments (i.e., screening and progress monitoring) can pinpoint specific skill deficits (e.g., phonemic awareness, morphological awareness, number sense, and number operations). Interventions in these specific deficits will support the students' learning and achievement.

The third quality is that RtI invites a systems approach to understanding students' learning, achievement, and behavior. Rather than positing that student difficulties are inherent to the student, the RtI framework emphasizes that classroom teachers and schools' administrative decisions have a greater impact. That is, the decisions made regarding curricular materials, instructional practices, intended outcomes, and behavior management are not the students' decisions. The students are responding to others' decisions. Examining the results of the universal screening and progress monitoring can help determine how well the school is achieving its intended outcomes.

The fourth quality is that RtI provides an alternative model to identifying students with specific learning disabilities. A generally accepted characteristic of specific learning disabilities is that students who one might expect to achieve do not respond to instruction and thus demonstrate an unexpected deficit in learning and performance. The RtI framework assesses a child's learning rate and the level of performance relative to peers and rules out that poor instruction is the causal agent for not responding. The conclusion is that the difficulties are intrinsic to the student as opposed to an external factor (e.g., poor instruction).

Essential Components

The RtI framework includes four essential components: universal screening, progress monitoring, levels of prevention or intervention, and data-based decision making.

Universal Screening

Brief, reliable, and valid assessments are administered to all of the students to predict who is at-risk for academic failure or behavioral difficulties. In practice, academic screening is typically in the reading (e.g., phonemic awareness and academic screening is typically in the reading (e.g., phonemic awareness and reading word identification) and mathematics components (e.g., number sense and operations) that are predictive of achievement. Screening is conducted at least annually (e.g., at the beginning of a school year) but also may be administered midyear and at the end of the year. In practice, schools use multiple screening instruments. When multiple screening instruments are used, the school district's RtI framework specifies how the different scores are integrated (i.e., judging a student's at-risk status). One of the decisions for school district implementation teams is whether to use national norms or develop local norms for comparing students' performance.

Progress Monitoring

Progress monitoring is a formative assessment approach intended to inform the school's data team about the student's responsiveness to intervention. As a formative measure, assessments are administered frequently enough to judge the student's progress, and such data are necessary for judging the student's responsiveness to the intervention. The progress monitoring data will provide two data elements: the student's learning rate and level of performance or achievement. The frequency and measures in progress monitoring vary with the level of prevention or intervention (e.g., primary, secondary, or tertiary levels). In the general education classroom (which is referred to as the primary level of prevention or Tier 1), periodic classroom monitoring is completed every 3–4 weeks. In the more intense intervention levels, secondary and tertiary levels, the questions are also, How well is the student responding to the intervention? Is the student's learning rate accelerating enough so that the student will catch up to classroom peers? Is the level of achievement corresponding to classroom peers? In the secondary prevention level, Tier 2, and the tertiary prevention level, Tier 3, the progress monitoring could be completed daily (e.g., such as monitoring self-injurious behaviors) or weekly and biweekly for academic skills.

Several considerations are important in progress monitoring. First, the assessment must be sensitive to assessing change; and thus, the items must be aligned to the instructional intervention. Second, the assessment must be administered after sufficient instruction to expect change. Last, an ambitious goal must be clearly stated, so that a student's progress can be assessed toward that goal.

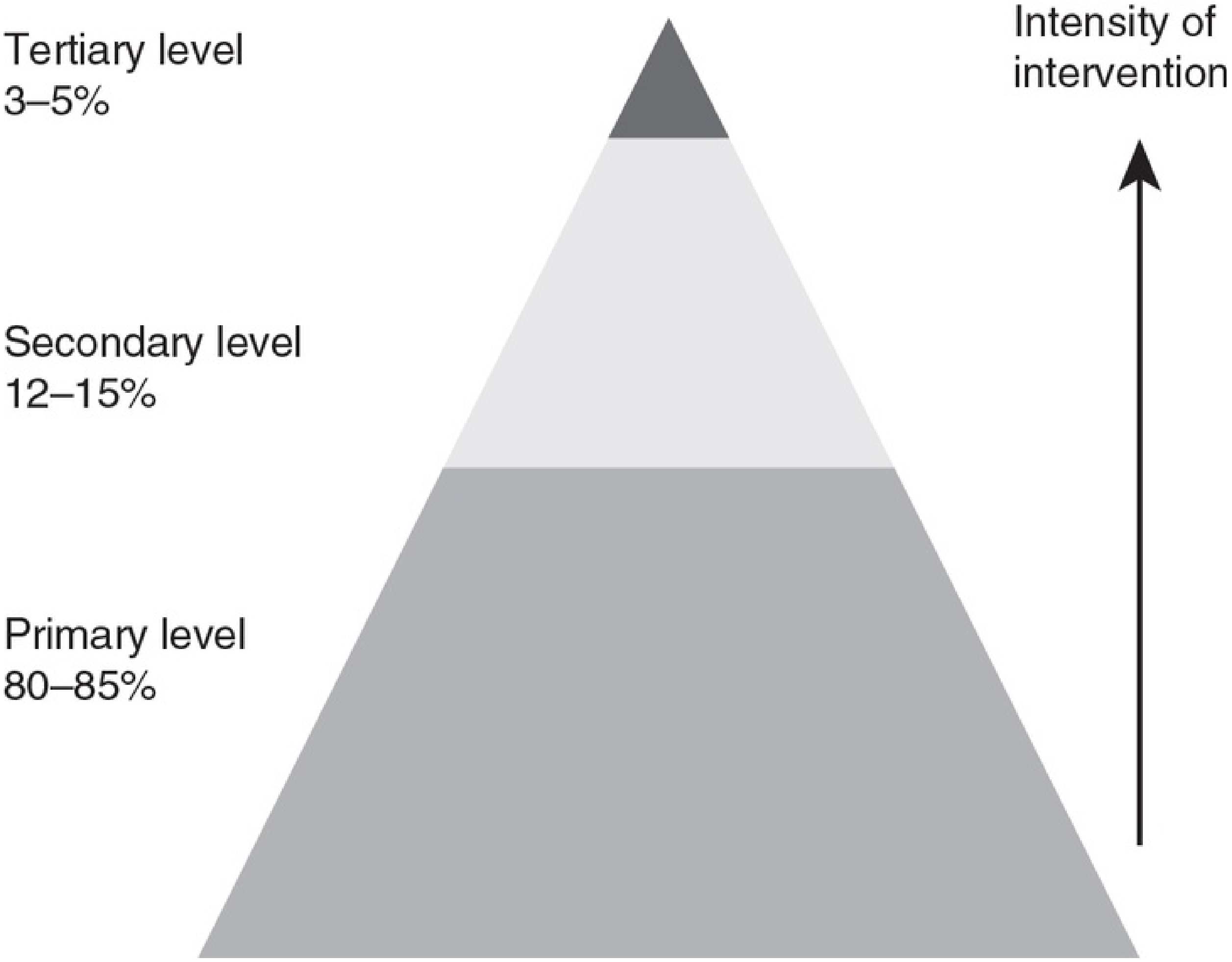
Levels of Prevention

Levels of Prevention

A fundamental principle of RtI, just as in public health prevention models, is to emphasize prevention. In schools, the goal is to prevent academic and behavioral difficulties. This principle is represented as graduated levels of prevention. Prevention activities have three levels: primary, secondary, and tertiary. These levels vary by the size of the targeted population and the intensity of the intervention. The intensity of the intervention is matched to the level or severity of the students' needs. Figure 1 graphically depicts the population size and intervention intensity. As the need becomes more acute, the more intense is the intervention. These prevention levels are often described as tiers of intervention. The primary prevention level corresponds to Tier 1, secondary level to Tier 2, and tertiary level to Tier 3. In practice, schools have generally implemented three to four tiers within the three prevention levels and may or may not include special education instruction as a separate tier. Some schools have implemented as few as two tiers and other schools as many as five tiers.

Figure 1 Preventive levels and student prevalence rates

Prevention levels and population targets



In the public health framework, primary-level preventive activities such as inoculations (e.g., flu shots), screenings (e.g., vision acuity screening), and information dissemination (e.g., value of healthy diets) are emphasized to reduce the occurrence of illness and disabling conditions. These preventive activities are directed to the population as a whole, are least obtrusive, comparatively inexpensive, and have a large-scale impact. In schools, this primary prevention level is also the most important prevention level and emphasizes all students' engagement in a high-quality, evidence-based core curriculum for a substantial amount of time every day. As a percentage, this primary level of prevention targets all of the students, but with an expectation of effectiveness for 80–85% of the population.

Because inoculations, however, are not always effective, treatment or interventions are needed to assist the students. These secondary and tertiary interventions are needed to assist the students. These secondary-and tertiary-level interventions are characterized as

- supplementary to the core curriculum,
- delivered outside of the general education classroom,
- have a limited duration,
- target-specific goals,
- delivered in small group settings, and
- assessed with frequent progress monitoring.

The secondary level is directed at 12-15% of the students. Secondary-level interventions are unique from the core curriculum. The student receives the secondary-level intervention to supplement the core curriculum, not to supplant that general education classroom instruction. The interventions are evidence-based and delivered with high fidelity, which generally requires a skilled instructor. The interventions are designed for 9–15 weeks' duration, delivered 4–5 times a week for 45–60 minutes per session in a small group setting. In practice, these interventions are validated standard treatment protocols selected because they yield efficacious results. High implementation or treatment fidelity is important to achieving the intended outcomes.

For a very small segment of the population (3–5%), an even more intensive intervention is needed. This tertiary prevention level is akin to a hospital's intensive care unit. In contrast to the secondary prevention level, the tertiary level involves a smaller instructional group (e.g., three or fewer students), more specific skills that are monitored with higher frequency, and a clinical or diagnostic approach to instructional planning and delivery. One might think of this intervention as a single subject research design. Because validated interventions were not as effective as desired, a more individualized, problem-solving approach is required. These interventions can also become more intense through the frequency with which they are presented, the length of the instructional sessions, and increased opportunities for responding. The instructor's systematic and diagnostic approach to teaching is critically important to the student's learning and achievement.

Data-Based Decision Making

The fourth RtI component, data-based decision making, is often represented as the center of the RtI framework. In this component, explicit, quantitative decision rules govern students’ level or tier assignment. These rules also provide a quantitative basis for judging when a student’s assignment should continue (e.g., continue in Tier 1) or be changed (e.g., assigned to Tier 2). The decision rules provide a transparent guide for teachers, students, and parents not only of the academic and behavioral goals but also of the student’s progress for achieving those goals. In practice, greater consistency and equity is achieved with district-level decision rules. School-based teams have a fixed schedule for reviewing students’ screening and progress-monitoring assessment results to determine the course of action for all of the students. The screening data generally reflects a student’s current level of performance and risk for academic or behavioral difficulties. The progress monitoring data reflect two indicators: the student’s current level of performance and the rate of improvement. For example, for a student receiving targeted instruction, the performance level may indicate a level that is just below the peer group, but a high, positive trajectory rate of improvement. In that case, the school’s data team might conclude that a less intensive intervention is warranted (e.g., change of assignment from a secondary level to a primary level).

Caveats

Evidence from numerous schools suggests that an RtI approach can improve reading, math, and behavioral outcomes for many students. Those findings, however, are not universal. A related point is when implemented with fidelity, an RtI approach will identify what has not worked for improving a student’s learning and performance, but those data do not predict what will be effective. Such decisions then become guided by highly qualified educators who are well versed in problem solving based on curricular, instructional, and psychological components of learning and achievement.

No studies have compared an RtI approach for specific learning disabilities determination to competing models. Furthermore, while standardized, norm-referenced tests commonly used in the psychoeducational battery have indices of reliability and validity, the complexities of a high fidelity implementation of RtI create even further challenges. This concern is not to say that RtI is without merit but to add a word of caution, as one considers the evidential and consequential validity evidence.

Practice Challenges

Variation in Practice

While the technical challenges of RtI implementation might be significant in some settings, practitioners usually have more difficulty with the social and cultural shifts. Technical challenges in RtI implementation include the selection and usage of screening and progress-monitoring measures and providing more intensive levels of intervention across the preventive levels. The social and cultural challenges include the shift in thinking about students' failures and in staff's roles and responsibilities. For the vast majority of schools, their modus operandi require significant changes for an effective, high implementation fidelity RtI implementation. With so many decisions regarding the implementation of the four components, schools will likely show variation. One can expect, though, that the four components are well specified in a procedures manual and that school staff are working toward consistent implementation.

Competing Values: Helping Students

One of the challenges that school implementing teams confront is recognizing that although most all students can benefit from more focused instruction in a secondary-level intervention, providing most students with that intensity of intervention actually defuses the intensity of instruction for those students most in need. Thus, agreement and adherence to data-based decision rules is critical for the framework to work as intended. In such situations, schools have generally recognized the importance of budgeting additional professional development and resources to the primary prevention level. In that way, more students are successful in the core curriculum.

Role Changes

As one considers the RtI framework and its components, numerous decisions are required for implementation (e.g., how many tiers will be implemented? which screening and progress monitoring assessments will be chosen? how often and when will they be administered? what interventions will be associated with each tier? and what will be the data-based decision rules?). Answers to these questions and similar ones are all necessary for implementation, but in comparison, the decisions about the role of staff are much more difficult. Staff roles change in implementation and thus extensive consideration is needed about how to best support staff as they make these changes. The support includes professional development, coaching with feedback, and time to adjust.

Daryl F. Mellard

See also Curriculum-Based Assessment; Curriculum-Based Measurement; Data-Driven Decision Making; Evidence-Based Interventions; Learning Disabilities; Multicultural Validity; Progress Monitoring; SchoolWide Positive Behavioral Support

Further Readings

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). Evaluation of response to intervention practices for elementary school reading (NCEE 2016-4000). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Berninger, V. (2006). Research-supported ideas for implementing reauthorized IDE with intelligent professional psychological services. Psychology in the Schools, 43, 781–796. doi:10.1002/pits.20188

Clark, B., Doabler, C. T., Smolkowski, K., Baker, S., Fien, H., & Cary, M. S. (2016). Examining the efficacy of Tier 2 kindergarten mathematics intervention. Journal of Learning Disabilities, 49, 152–165. doi:10.1177/0022219414538514

Compton, D. L. (2000). Modeling the growth of decoding skills in first-grade children. Scientific Studies of Reading, 4, 219–259. doi:org/10.1207/S1532799XSSR0403_3

Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for schoolwide positive be9havior support. Focus on Exceptional Children, 42(8), 1–14.

Kovaleski, J., & Glew, M. C. (2006). Bringing instructional support teams to scale: Implications of the Pennsylvania experience. Remedial and Special Education, 27, 16–25. doi:10.1177/07419325060270010101

Mellard, D. F., & Johnson, E. (2008). RTI: A practitioner’s guide to implementing response to intervention. Thousand Oaks, CA: Corwin Press.

Sugai, G., & Horner, R. H. (2009). Defining and describing schoolwide positive behavior support. In W. Sailor, G. Dunlap, G. Sugai, & R. Horner (Eds.), Handbook of positive behavior support (pp. 307–326). New York, NY: Springer.

Vanderheyden, A. (2011). Technical adequacy of response to intervention decisions. Exceptional Children, 77, 335–350. doi:10.1177/001440291107700305

Tineke A. Abma Tineke A. Abma Abma, Tineke A.

Responsive Evaluation

Responsive evaluation

1426

1430

Responsive Evaluation

Responsive evaluation is an approach to the formal evaluation of educational and social service programs. It takes the issues and concerns of various stakeholders as a point of departure to determine the quality and worth of a program or practice. It favors personal experience and draws upon the ordinary ways people perceive quality. More than other evaluation approaches, it focuses on the meaning and context and the cultural plurality of people. Robert Stake coined the term *responsive evaluation* in the mid-1970s, and his ideas have opened new vistas to evaluation. This entry provides an overview of the relevance, original ideas, development, and strands within responsive evaluation.

Relevance

When responsive evaluation was introduced by Stake, it was characterized by the fact that it takes the concerns and issues of stakeholders as criteria for evaluation. Stake claimed that an evaluation is responsive if it orients more to program activities than program intents; if it responds to audience requirements for information; and if the different value perspectives held by stakeholders are referred to in reporting program success and failure.

Stake’s initial ideas helped accelerate a transformation of the evaluation enterprise into its current pluralistic character. Because of their widespread appeal and continuing importance, these ideas have also permeated many varied strands of evaluation theory and practice. Responsive evaluation is considered relevant, because

- postmodern society has become more pluralistic, and so evaluation approaches that deal with this plurality of values and interests are needed;
- people are less inclined to accept the wisdom of experts, and so evaluation approaches that legitimize and incorporate the ordinary wisdom, voices, and understandings of multiple stakeholders are needed to improve practice; and
- the quality and meaning of programs cannot be understood and reduced to a set of simple indicators—an insistence on complexity and representations of multiple realities and contexts is needed.

All of these needs are addressed in responsive evaluation.

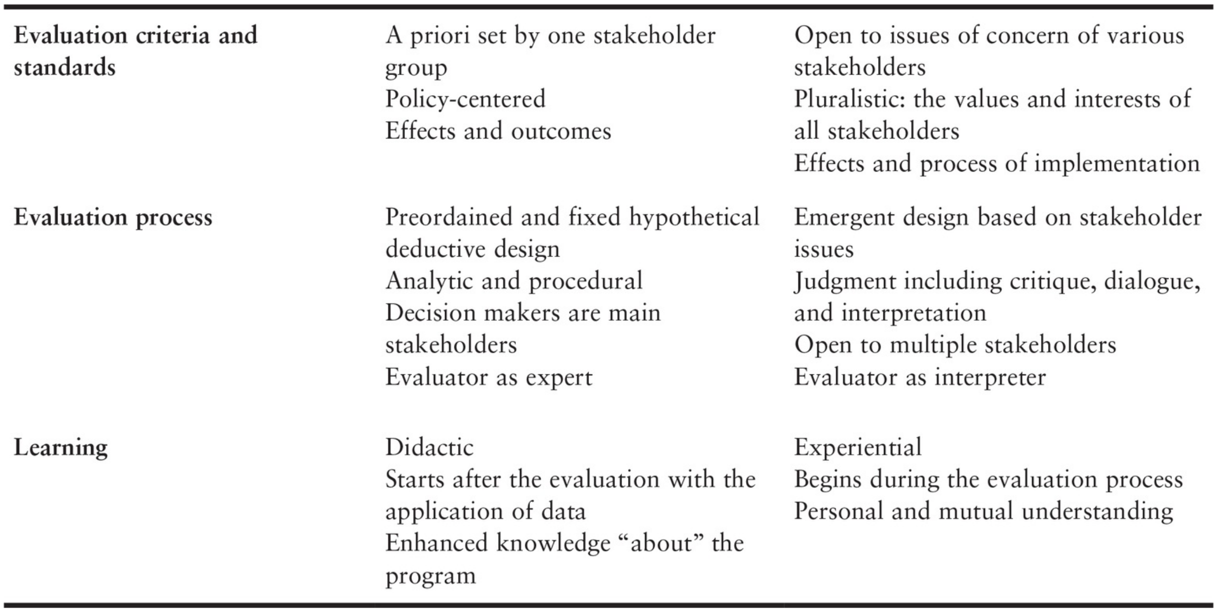
Original Ideas

With the public address titled “Program Evaluation, Particularly Responsive Evaluation” given during a sabbatical in Sweden in 1973, Stake offered a new vision of and rationale for educational and social program evaluation to the then fledgling evaluation communities. In this vision, evaluation was reframed—from the application of sophisticated analytic techniques to address distant policy makers’ questions of program benefits and effectiveness “on the average,” to an engagement with on-site practitioners and stakeholders about the quality and meanings of their practice.

As subsequently elaborated by Stake and others, responsive evaluation proposed a new kind of evaluation knowledge and a new way of conceiving the evaluator role. Evaluation knowledge was no longer to be thought of solely in terms of causal, propositional knowledge but to include the socially constructed meaning and worth of programs in their context. This implied a shift from a postpositivist to a hermeneutic and constructivist approach to research. Over the years, Stake has argued for a holistic understanding of a program from the perspectives of those engaged in it, thus lending depth and richness to understanding the success or failure of programs. Scholars promoted naturalistic inquiry and qualitative research methods as most appropriate to gain insight into participants’ insider perspectives, program uniqueness, and context. Moreover, these scholars argued for ways of reporting evaluation that appeal to and connect with the ordinary ways in which participants and stakeholders make sense of the world.

Responsive evaluation can be compared to a standards-based approach. The aim of standards-based evaluation—or performance measurement—is to measure the quality of a program by comparing the factual effects with intended program goals and standards. A fixed set of criteria and standards are used, usually determined by one stakeholder group (policy/decision makers). The assessment is concentrated on the realized outcomes (and less or not on the input, process, and context). This type of evaluation is prevalent within contemporary society. It is policy centered and stands in contrast to the pluralistic character of responsive evaluation.

In responsive evaluation, the design emerges on the basis of a conversation with and among stakeholders and their issues of concern. To be able to respond to the issues of multiple stakeholders, the design of responsive evaluation is emergent. In the process, the evaluator deals not only with decision makers but also with all those whose interests are at stake and whose input is relevant to the assessment of the worth of the program. Quality assessment is not a procedure of comparing results with criteria and standards but includes moral and political speculation, critique, interpretation, dialogue, and judgment. The responsive evaluator typically acts as an interpreter and must have what Stake called anthropological sensitivity—an ability to pay careful attention to the concrete details of people’s experiences; their activities over time; and their physical, cultural, and social context. Table 1 summarizes the main differences between responsive and standard-based evaluation.



Evaluation criteria and standards	A priori set by one stakeholder group Policy-centered Effects and outcomes	Open to issues of concern of various stakeholders Pluralistic: the values and interests of all stakeholders Effects and process of implementation
Evaluation process	Preordained and fixed hypothetical deductive design Analytic and procedural Decision makers are main stakeholders Evaluator as expert	Emergent design based on stakeholder issues Judgment including critique, dialogue, and interpretation Open to multiple stakeholders Evaluator as interpreter
Learning	Didactic Starts after the evaluation with the application of data Enhanced knowledge “about” the program	Experiential Begins during the evaluation process Personal and mutual understanding

Development and Strands Within Responsive

Evaluation

When Stake completed evaluations in the mid-1970s, he was struck by the finding that most programs did not realize their goals and intentions. He wondered whether those programs were failing or whether the methodologies and instruments used simply were not good enough to capture what is going on in reality. Responsive evaluation was thus in part a critical response to the dismay Stake felt about the narrow selection of data being used for formal evaluation. He saw standards-based evaluation as the dominant approach emphasizing strong (preferably experimental and quantitative) measurement procedures and legitimizing only two kinds of data—goals and outcomes—thus neglecting input, process, and context.

When Stake proposed responsive evaluation, he was partly reflecting the ideas of Tom Hastings, Lee Cronbach, Mike Atkin, Barry MacDonald, and David Hamilton. They spoke of the necessity of organizing the evaluation of educational programs around what was happening in classrooms, drawing attention to what educators and students were *doing*, how they saw themselves and others, the language they spoke, and in what kind of sociopolitical context they were working. Later refinements of responsive evaluation were spurred by Ernest House, Stephen Kemmis, Egon Guba, and Yvonna Lincoln; restated by Linda Mabry, Thomas Schwandt, Helen Simons, Yoland Wadsworth, Ian Stronach, and Stafford Hood; and updated in a 2001 issue of *New Directions for Evaluation*, edited by Jennifer Greene and Tineke Abma.

Being responsive meant for Stake that an evaluator should take into account the multiple perspectives in a manner that was as truthful as possible to the values of each stakeholder and to share evaluation accounts relevant and meaningful to multiple audiences. Although Stake had proposed a pluralistic approach, Guba and Lincoln went a step further, arguing for an approach that engaged stakeholders more actively in a collaborative evaluation process, which they called *Fourth Generation Evaluation*. Guba and Lincoln spoke of evaluation as a democratic and interactive process of negotiation with and among stakeholders. Later, Schwandt and Abma would refer to this process as hermeneutic dialogue. Listening, probing, and a search for meaning characterize this process rather than confronting, attacking, and defending. Central features of dialogue are openness, respect, inclusion, and engagement.

This notion of dialogical understanding was connected to Gadamer’s philosophical hermeneutics. The latter approach uses hermeneutic dialogue to engage stakeholders in a learning process to help them to better understand themselves and each other and, hence, to place their own viewpoints into perspective. Stakeholders thus gain a better understanding of a program or practice through the combination and amalgamation of various perspectives. From a hermeneutic perspective, human life is essentially a process of understanding. Through stories, people make sense of their world and are interconnected with each other. Hermeneutic dialogue takes the complexity of human life (embedded in their stories and experiences) as a starting point for mutual learning processes in which all stakeholders change by their interaction. Social change and learning processes occur when people extend their horizon by the appropriation of new perspectives. Dialogue in this hermeneutic sense is an ongoing and cyclical process among stakeholders, aimed at reciprocal understanding and acceptance. Consensus is not the ultimate goal of this kind of evaluation, as this is never an absolute value—conditions change over time, and a lack of consensus and ambiguities, expressed through the narratives of stakeholders, generate reasons to interact and continue ongoing dialogues. The dialogical process among stakeholders implies new roles for the evaluator; among them, the role of a facilitator creates conditions for genuine dialogues teacher, and Socratic guide.

The responsiveness to a wider set of audiences includes what Jennifer Greene has called a participatory ethics. Stakeholders become partners and coresearchers initiated into the evaluation process and learn during and through active engagement in the process. Within this participatory framework, evaluation should actively steer toward the inclusion of marginalized voices to prevent "epistemic injustice." Moreover, participants are not solely taken serious for the information they are able to provide; their participation is of intrinsic value and given in by democratic principles. This participatory strand within responsive evaluation reflects a value-committed stance working for social justice, equality, empowerment, and emancipation. The work of Abma, Greene, and Wadsworth reflects this strand of responsive evaluation working toward social change and empowerment based on the experiential knowledge of multiple stakeholders. Such work explicitly pays attention to power differentials. It is important to be careful in defining who is considered to be a marginalized group, to avoid stigmatization and exclusion of other groups that might even have less voice. To foster a genuine dialogue, the less powerful should first be given the opportunity to bring their issues to the fore. Guba and Lincoln have provided authenticity and fairness criteria to check whether the process is fair and enhancing the mutual understanding and empowerment of participants.

Case Examples

It is difficult to tell from an evaluation report whether the investigation itself was "responsive." A final report seldom reveals how issues were negotiated and how audiences are served. Examples of studies that were intentionally responsive can be found in the field of education, in social policy, and in health care.

Tineke A. Abma

See also Collaborative Evaluation; Constructivist Approach; Culturally Responsive Evaluation; Democratic Evaluation; Empowerment Evaluation; Naturalistic Inquiry; Qualitative Research Methods; Stakeholders

Further Readings

Abma, T. A. (2006). The practice and politics of responsive evaluation. American Journal of Evaluation, 27(1), 31–43.

Abma, T. A., Nierse, C., & Widdershoven, G. A. M. (2009). Patients as research partners in responsive research. Methodological notions for collaborations in research agenda setting. Qualitative Health Research, 19(3), 401–415. doi:10.1177/1049732309331869

Abma, T. A., & Stake, R. E. (2001). Stake's responsive evaluation: Core ideas and evolution. In J. Greene & T. A. Abma (Eds.), Responsive evaluation: New directions for evaluation, No. 92 (pp. 7–23). San Francisco, CA: Jossey-Bass.

Greene, J. C. (2006). Evaluation, democracy, and social change. In I. Shaw, J. Greene, & M. Mark. (Eds.), The SAGE handbook of evaluation (pp. 118–140). Thousand Oaks, CA: SAGE.

Greene J. C., & Abma, T. A. (Eds.). (2001). Responsive evaluation: New directions for evaluation (92, winter). San Francisco, CA: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1981). Effective evaluation. Beverly Hills, CA: SAGE.

Guba, E. G., & Lincoln, Y. S. (1989). Fourth generation evaluation. Newbury Park, CA: SAGE.

Schwandt, T. A. (2002). Evaluation practice reconsidered. New York, NY: Peter Lang.

Stake, R. E. (1975). To evaluate an arts program. In R. E. Stake (Ed.), Evaluating the arts in education: A responsive approach (pp. 13–31). Columbus, OH: Merrill.

Stake, R. E. (2004). Standards-based and responsive evaluation. Thousand Oaks, CA: SAGE.

Kevin A. Hallgren Kevin A. Hallgren Hallgren, Kevin A.

Restriction of Range

Restriction of range

Restriction of Range

Restriction of range occurs when a variable has less variability in a study sample than in the full population. The restricted range can be present for observed variables (dependent or independent variables) or in other variables that were not measured. Restricted range can affect statistical inferences, typically in the direction of underestimating the effect sizes and underestimating validity coefficients of associations between predictors and outcomes.

For example, consider the scenario in which scores from a college entrance examination, such as the SAT (*x* variable), are used to predict first-year college grade point average or GPA (*y* variable). Many colleges base their admission criteria on SAT scores, and therefore first-year college GPA data would only be available for a restricted sample (i.e., students with SAT scores above a certain level) rather than the full population of students taking the SAT. If a researcher aims to test the validity of SAT scores in predicting first-year college GPA, the sample would have a restricted range because data on first-year college GPA would not be available across the full distribution of students taking the SAT.

More specifically, this example illustrates *direct* restriction of range because the restricted range is directly attributable to a variable that is restricted and also measured and used in the statistical analysis (i.e., SAT scores). However, restriction of range may also be *indirect* if the restricted range in the sample is due to a variable that is unmeasured or not included in the statistical analysis, but nonetheless causes restriction in the ranges of other variables in the analysis. For example, if college admission was decided entirely by high school GPA (not SAT scores), then an analysis of SAT scores predicting first-year college GPA could still be indirectly restricted if the restricted variable (i.e., high school GPA) was correlated with other variables in the analysis (e.g., SAT scores) and caused them to also be restricted.

Restricted range can create statistical problems, particularly related to validity coefficients and effect size estimates. Typically, restriction of range reduces correlation coefficient magnitudes and other measures of effect size (e.g., R^2) when a variable is directly or indirectly restricted. Reduced variability limits the degree to which that variable can predict another variable, and effect sizes in a restricted sample are generally smaller than effect sizes in an unrestricted sample. (For an extreme example: Imagine a college that only accepted students with perfect SAT scores. It would have no ability to predict first-year college GPA from SAT scores, and therefore any true correlation between first-year college GPA and SAT scores would be reduced to zero.) Attenuated effect sizes, in turn, can affect interpretations about the utility of a measure in predicting an outcome, such that a predictor variable will typically appear to have lower validity in the restricted sample than it actually has for the full population. Estimates of reliability (e.g., interrater reliability, internal reliability) are also typically attenuated due to restricted range for the same reason.

To identify and correct for the consequences of restricted range, the data source and study design should be carefully examined for potential restriction of range, particularly when an analysis focuses on assessing validity coefficients or effect size estimation. Many study designs have inclusion and exclusion criteria that can restrict the range of the resulting data, and these factors should be considered. In other cases, studies may unintentionally create restricted range by analyzing subsets of data that create restricted range. For example, the correlation between high school GPA and SAT scores will usually be higher if one computes a single correlation estimate across the full range of high school GPAs and lower if one computes three separate correlation estimates each within restricted subsamples of students with only low, intermediate, or high GPAs. These and other design factors should be carefully considered by the researcher to assess whether they may lead to direct or indirect restriction of range.

Descriptive statistics should also be carefully examined to identify possible restricted range. Univariate statistics, including means, medians, standard deviations, ranges, and graphical methods (e.g., histograms, scatterplots), should be examined for each variable in the analysis and compared to its expected values in the larger population.

Study design considerations can often prevent restricted range. However, preventing restricted range is not always feasible and statistical corrections have been developed for these cases. Most corrections aim to more accurately represent the unrestricted (population-level) correlation between a predictor and an outcome, which will typically be higher than the attenuated correlation obtained with a restricted sample. Although it is generally agreed that restricted range can cause statistical problems, corrections for it are often not utilized in practice.

In brief, most methods correct for restricted range by incorporating estimates of the unrestricted-population variance when correlation indices are computed. Therefore, these assume that the variances (or standard deviations) of the full, unrestricted population are known or can be estimated. Most of these methods also assume that relationships between variables continue to be linear beyond the point of restriction (i.e., variables have the same association in the observed and unobserved segments of the population), which may not always be testable or accurate.

Kevin A. Hallgren

See also Convenience Sampling; Predictive Validity; Psychometrics; Representativeness; Scatterplots; Selection Bias; Validity; Variance

Further Readings

Bland, J. M., & Altman, D. G. (2011). Correlation in restricted ranges of data. British Medical Journal, 342, d556. Retrieved from http://dx.doi.org/10.1136/bmj.d556

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. Personnel Psychology, 55, 807–825.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches of restriction of range in correlation analysis. Practical Assessment, Research, & Evaluation, 14(5), 1–9.

Zimmerman, D. W., & Williams, R. H. (2000). Restriction of range and correlation in outlier-prone distributions. Applied Psychological Measurement, 24(3), 267–280.

Chunmei Zheng Chunmei Zheng Zheng, Chunmei

Results Section

Results section

1431

1432

Results Section

The Results section is a part of a research paper in which the author describes findings as clearly and objectively as possible after a series of analyses. It is a summarized report with narrative text, supporting evidence, and sometimes illustrative examples, tables, or figures. Interpretations for a specific result are included in the Discussion section that follows the Results section. A research paper typically consists of five sections: Introduction, Methods, Results, Discussion, and Conclusions. The Results section is a crucial part of a research paper because it contains answers to research questions. This entry describes how to organize results, how to present them, how to demonstrate findings with figures and tables, and some general guidelines to write the Results section. It concludes with an explanatory example.

A well-organized Results section is easy to follow and understand. With a simple design, it usually starts with a few sentences briefly summarizing the research questions and main analyses. With a complicated design, an introduction paragraph provides more details. Authors then present key findings with supporting data and materials. The order of the presentation of key findings is similar to that in other sections of the research paper, so that there is coherence among the sections. Specifically, the sequence and structure of the results follows the same order of the investigated research questions. For instance, if the research questions contain subheadings, then the results should follow the same structure.

The goal of the Results section is to help readers understand the presented statements; therefore, the final results should be concise, simple, clear, and objective. To achieve this goal, the Results section may contain verbal and numerical explanations, examples, or numbered graphs and figures to help present results effectively. Moreover, the narrative text and supportive materials should complement each other. For statistical analysis, the Results section often includes descriptive analyses. For a quantitative study, this might include means, standard deviations, or correlations for the overall group characteristics and inferential statistics for a significant test. A summary of statistical analyses is usually embedded in a text description and reported in parentheses. In addition to the summary statistics, statistics such as how much groups are different—for example, males (*M* = 25, *SD* = 3.2) averaged to fall asleep 20 minutes faster than females (*M* = 45, *SD* = 4)—can inform readers about the nature of differences and relationship. For quantitative studies, the Results section might describe themes or conceptual findings, sometimes presented graphically.

When presenting a large amount of information, supporting materials such as tables and figures can help readers visualize the relationships of variables. If tables and figures are included, they are typically inserted near the relevant text description but may be placed at the end. They are numbered sequentially and consecutively, as authors will refer to them in a text such as Table 1, Table 2, Figure 1, Figure 2, and so on. The format and vocabulary among all tables and figures should be consistent.

There are some general guidelines regarding grammar when reporting results. For instance, results are generally written in past tense because all hypothesis tests should have been completed by the time of writing. However, tables, figures, and graphs are usually referred to in the present tense. For instance, the results can be written as "Overall, a variable had a significant impact for … ," but authors can refer to tables by writing, "Table 1 shows that …" Although active voice should be used as much as possible, passive voice is acceptable. All relevant findings, regardless of significant, nonsignificant, or negative results, should be reported in the Results section. Unexpected results can be important findings even though not consistent with the predicted results and may suggest a further study.

An example is provided here to demonstrate how to write a Results section. Assume the research question of a paper is to study whether parents' height has a significant impact on their children's height. Both parents' and children's heights were collected and analyzed with the regression method to examine this research question. The results can be written as follows:

> A simple linear regression was conducted to examine whether parents' height can significantly predict their children's height. The results of the regression indicated that parents' height explained 30% of the variance, $R^2 = .30$, $F(1, 14) = 10.8$, $p < .01$. Parents' height significantly predicted children's height ($\beta = .43$, $p < .001$).

Chunmei Zheng

See also Part Correlations; Partial Correlations; Post Hoc Analysis; Posttest-Only Control Group Design; Pre-experimental Designs

Further Readings

Gilgun, J. F. (2005). "Grab" and good science: Writing up the results of qualitative research. Qualitative Health Research, 15(2), 256–262.

Jalalian, M., & Danial, A. H. (2012). Writing for academic journals: A general approach. Electronic Physician, 4(2), 474–476.

Kliewer, M. A. (2006). Writing it up: A step-by-step guide to publication for beginning investigators. Journal of Nuclear Medicine Technology, 34(1), 53–59.

Lertzman, K. (1995). Notes on writing papers and theses. Bulletin of the Ecological Society of America, 76(2), 86–90.

Perneger, T. V., & Hudelson, P. M. (2004). Writing a research article: Advice to beginners. International Journal for Quality in Health Care, 16(3), 191–192.

Websites

Writing the Results:
http://users.clas.ufl.edu/msscha/ThesisCSS/thesis_results.html

Kathleen Sexton-Radek Kathleen Sexton-Radek Sexton-Radek, Kathleen

Lucinda Simmons Lucinda Simmons Simmons, Lucinda

Reverse Scoring

Reverse scoring

1432

1434

Reverse Scoring

A common attitude scaling format, Likert-type scaling, presents a statement and asks respondents to agree or disagree, and scores range from, for example, 1 to 5. Sometimes the same group of statements on a single measure is stated in different “directions.” That is, sometimes a 5 indicates a high level of endorsement of a particular attitude, whereas on other items, a 5 means a low level of endorsement of that attitude. Before responses can be combined into a single meaningful total score, all items must be in the same direction. To accomplish this, the scores for those items that are in an opposite direction are “reversed.” High scores become low scores and low scores become high scores.

Scores are reversed in a straightforward manner that depends on the range of possible scores for the selected items. Using the common 1, 2, 3, 4, 5 format, where 1 = *strongly disagree* and 5 = *strongly agree*, one would reverse score in this way:

> 1s become 5s, 2s become 4s, 3s remain as 3s, 4s become 2s, and 5s become 1s.

Reverse scoring is necessary when research instrument developers have purposefully written a group of items with some items in a different direction than others. A mix of directions in attitude (or any self-report) statements is sometimes designed to break a mental response set in the respondent or force increased concentration when responding. Reversing some questions is also thought to reduce acquiescence and boredom of respondents. However, respondents may misinterpret the test statements when the wording is reversed. It is believed that this may occur due to awkward phrasing on items that are written in reverse, such as “I don’t often read the funny papers.” Some populations with concentration difficulties, such as the elderly and children diagnosed with attention-deficit/hyperactivity disorder, have been studied with evidence of the failure to understand or attend to reverse scored items. It has also been suggested that reversed items may actually measure a construct different than the one intended by a researcher, and factor analyses frequently load items on different factors when they are worded in reverse. Consequently, additional analysis is recommended to indicate the parity of reversed items to other items on the test.

Reverse scored items on assessment scales used in personality theory measurement and clinical symptomology have found some compromise in internal consistency for reverse scored items among older adults. Translating an instrument from one language to another also raises concern, as the linguistic similarity between 2 items written in opposite directions may break down when in another language or cultural context.

The psychometric characteristics of items when reversed have also been studied. Internal consistency tends to be relatively equal when comparing reversed and original items, though it is not always the case and should be examined in each given study. Item response theory analyses have been conducted on selected personality measures to confirm the unidimensional nature of a group of items, whether reversed or not. Researchers have concluded that the different samplings of participants examined by item response theory analysis reflected differing response styles in terms of both the multidimensional nature of the questionnaire and responses to reverse scoring. This item response theory work and other studies have sometimes found that self-reported traits are more complex than reverse scoring ratings may be able to capture.

The general conclusions from the empirical literature do not fully support the need for, or utility of, including reversed items on attitude instruments. Although the reasons for reverse scoring are to interrupt response sets and discourage acquiescing answering, there is little research to suggest that it is necessary or useful. The future direction lies with the test developer to write items that are parsimonious, well researched, and empirically valid. Then, the careful design of a response method that will allow for the full conveyance of the participant’s response should be selected.

Kathleen Sexton-Radek and Lucinda Simmons

See also Attitude Scaling; Likert Scaling; Rating Scales

Further Readings

Brunner, G. C.(2016). Scale-related pet-peeves: Don’t use reverse-coded items in scales. Southern Illinois University Office of Scale Research Department of Marketing Blog. Retrieved May 18, 2016.

Carlson, M., Wilcox, R., Chou, C., Yang, F., Blanchard, J., Marterella, A., & Clark, F. (2011). Psychometric properties of reverse-scored items on the CES-D in a sample of ethnically diverse older adults. Psychological Assessment, 23(2), 558–562. doi:10.1037/a0022484

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 60, 151–174.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37, 504–528.

Llorente, E., Warren, C. S., de Eulate, L. P., & Gleaves, D. H. (2013). A Spanish version of the sociocultural attitudes towards appearance questionnaire-3 (SATAQ-3): Translation and psychometric evaluation. Journal of Clinical Psychology, 69(3), 240–251. doi:10.1002/jclp.21944

Rodebaugh, T. L, Woods, C. M, Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (2006). The factor structure and screening utility of the Social Interaction Anxiety Scale. Psychological Assessment, 18(2), 231–237.

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. Organizational Research Methods, 15(3), 363–384.

VonSonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. PloSONE, 8(7), e68967. doi:10.1371/journal.pone.0068967

Weijters, B., Baumgartner, H., & Schillewaet, N. (2013). Reversed item bias: An integrative model. Psychological Method, 18(3), 320–334. doi:10.1037/a0032121

B. Evan Blaine B. Evan Blaine Blaine, B. Evan

Robust Statistics

Robust statistics

1434

1436

Robust Statistics

Robust statistics are procedures that maintain nominal Type I error rates and statistical power in the presence of violations of the assumptions that underpin parametric inferential statistics. Since George Box coined the term in 1953, research on robust statistics has centered on the assumption of normality, although the violation of other parametric assumptions (e.g., homogeneity of variance) has their own implications for the accuracy of parametric procedures. This entry looks at the importance of robust statistics in educational and social science research and explains the robustness argument. It then describes robust descriptive statistics, their inferential extensions, and two common resampling procedures that are robust alternatives to classic parametric methods.

Robust statistics are important tools for educational and social science researchers because of three well-established findings. First, parametric methods (e.g., ANOVA, least squares regression) are the most commonly used procedures for significance testing in the social sciences; some estimates indicate that over 90% of published articles use a parametric significance test. Second, surveys of the educational and psychological literature show that nonnormally distributed data is the rule rather than the exception. Third, even modest departures from normality can substantially compromise both the Type I error rate and the power of parametric inferential procedures.

The robustness argument refers to the long-standing claim in the social sciences that parametric procedures such as the *t* test are “robust to violations” of the assumption of normality, meaning that the tests maintain accurate Type I error rates in the face of nonnormality. Originating in several research articles from the 1970s, the robustness argument has been repeated in introductory statistics textbooks, asserted by researchers in defense of their use of parametric methods, and over time become accepted as fact in the social science research community.

The near ubiquity of parametric procedures for significance testing in social science research speaks to the acceptance of the robustness argument. However, the research underlying the robustness argument has been criticized both for its methods and interpretation of results. Subsequent research has substantially, if not convincingly, established that beyond some very specific circumstances in which parametric procedures are in fact robust to violations of the assumption of normality, the robustness of *t* tests and other parametric procedures to violations of normality is the exception rather the rule.

The robustness argument invokes the central limit theorem, which provides for normal sampling distributions of the mean (given adequate sample size) even when the parent population is not normally distributed. However, the central limit theorem says nothing about the distribution of *t*, from which probabilities are derived for *t* tests of null hypotheses and *t* quantiles derived for constructing confidence intervals (CIs). Simulation studies show that under conditions of nonnormality, inferences based on the *t* distribution are inaccurate (i.e., nominal Type I error rates are not maintained) and can be very inaccurate even with modest departures from normality in the parent population. Combined with the commonality of nonnormally distributed data mentioned earlier, the influence of the robustness argument on statistical practices has broad implications for research literatures in education, psychology, and beyond.

The most pernicious departures from normality, from the standpoint of undermining parametric significance tests, are those that take the form of heavy tailed distributions. Heavy tailed (also called contaminated normal) distributions are common in educational research where a target population is contaminated with cases from subpopulations that have different means and variances, or from the presence of outliers, or both. Worse, heavy tailed distributions appear to be normal by visual inspection, and their nonnormality often goes undetected by tests of normality. Robust descriptive statistics are, by definition, resistant to the influence of outliers, and inferential procedures that use robust descriptive statistics inherit the same resistant quality.

Common robust descriptive statistics include the trimmed mean and variance, Winsorized mean and variance, and M-estimators. As a group, these statistics moderate the influence of outliers or heavy tails on estimates of location and moderate the influence of outliers or heavy tails on estimates of location and variability and are much preferable to data transformations as methods to deal with outliers or restore normality. Trimming removes a set percentage of cases in the upper and lower tails and calculates the mean or variance of the remaining cases. The median, which is widely appreciated as being resistant to the influence of outliers, is a 50% trimmed mean and therefore a robust estimator of location.

Winsorizing involves the systematic recoding of cases in the tails of a skewed or heavy tailed distribution, with the mean and variance calculated from the recoded data. Like Winsorizing, M-estimators also reassign values to observations in the tails of a distribution but do so based on one of several estimating functions. When these robust descriptive statistics are used in parametric inferential procedures, such as when a *t* test is calculated with trimmed means and variances, those procedures in turn become more robust. Robust descriptive and inferential statistics can be generated in most modern statistical software packages.

The robust procedures just described rely on theoretical probability distributions (e.g., *t*) to approximate the underlying distribution and generate probabilities for inference but do so with robust estimators of mean and variance. This category of robust inferential procedures is therefore still *parametric*. In contrast, other robust methods create empirical probability distributions from sample data and use those distributions for inference and estimation.

Certain robust procedures are freed from parametric assumptions, such as the assumption of normality, because the underlying probability distribution is directly estimated from sample data rather than approximated by a mathematical distribution. Two common examples are the bootstrapped CI and the permutation test for a mean difference. A bootstrapped 95% CI for estimating μ is produced via a resampled distribution of thousands of sample means. From that distribution, the 2.5% and 97.5% quantiles become the lower and upper limits, respectively, of the CI. A permutation test for a mean difference also starts with sample data, creating a probability distribution of mean differences from thousands of independent shufflings of scores into two random samples, each generating a mean difference, from which a *p* value for the observed mean difference can be retrieved. Resampled, robust alternatives exist for most parametric inferential procedures and are also part of most statistical software packages.

B. Evan Blaine

See also Random Assignment; Winsorizing

Further Readings

Erceg-Hurn, D., & Mirosevich, V. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. American Psychologist, 63(7), 591–601.

Huber, P. (1972). The 1972 Wald Lecture robust statistics: A review. Annals of Mathematical Statistics, 43(4), 1041–1067.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156–166.

Wilcox, R. (2009). Robust data analysis. In R. Milsap & A. Maydeu-Olivares (Eds.), The SAGE handbook of quantitative methods in psychology (pp. 387–403). Thousand Oaks, CA: SAGE.

RTI

RTI

1436

1436

RTI

See Response to Intervention

James Dean Brown James Dean Brown Brown, James Dean

Rubrics

Rubrics

1436

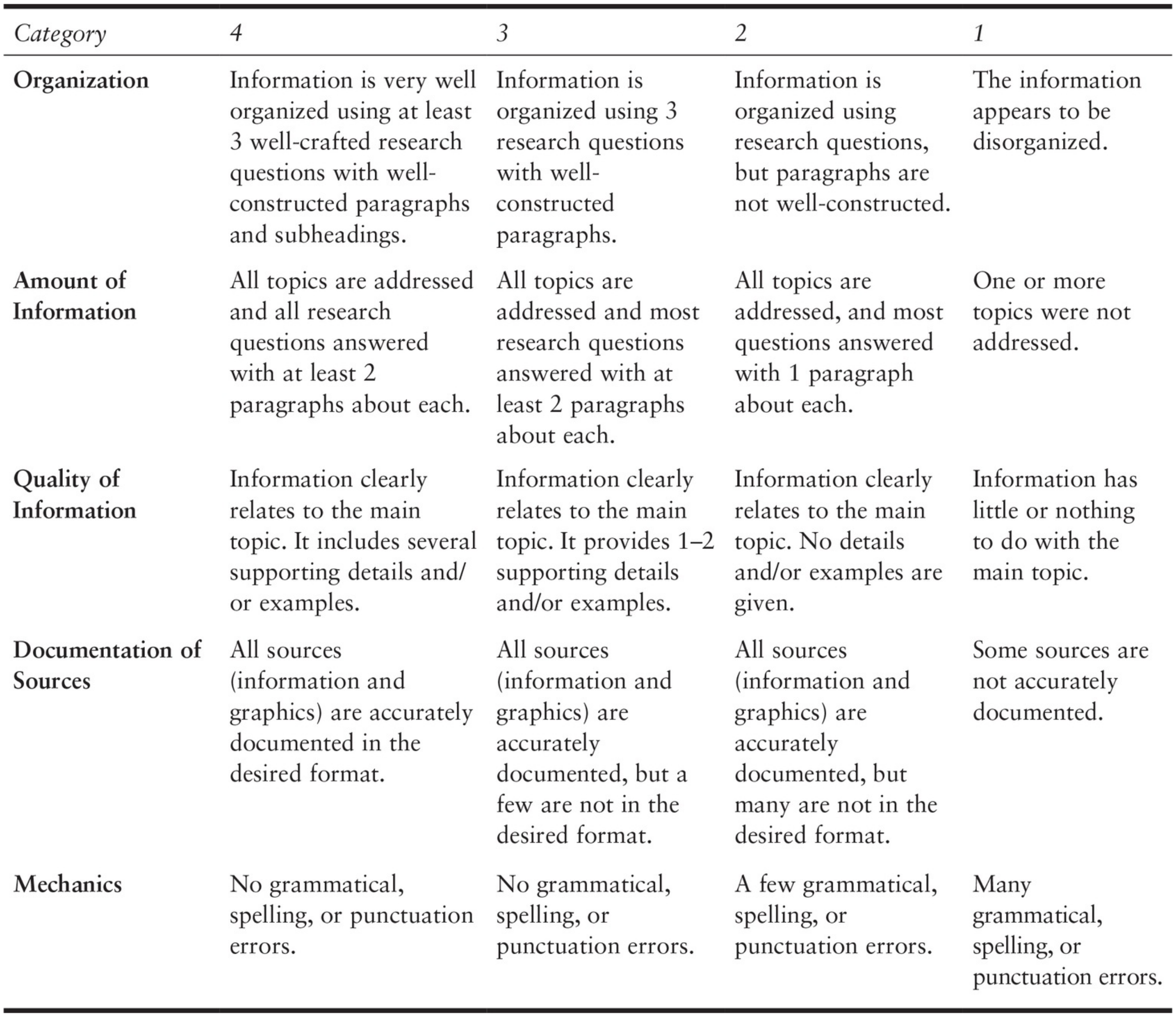
1440

Rubrics

A rubric is a tool that teachers and testers can use for assessing, scoring, and providing feedback to examinees for any performance, simulation, or task that those examinees are required to do as part of a classroom activity, formal classroom assessment, or institutional testing. At minimum, a rubric is a written grid that has possible scores on one dimension and descriptions of the characteristics of performances at each score level in the cells of the grid. Some rubrics are also designed for separately rating multiple categories or criteria that are labeled on a second dimension. This entry explains three types of rubrics (analytic, holistic, and checklist rubrics) and discusses challenges that teachers face in using rubrics in their classroom testing as well as the issues faced by testers when designing and using large-scale, high-stakes examinations (including rater training and various potential problems in statistical analysis).

Consider a teacher who needs to assess and provide feedback to students on their end-of-term written reports. The teacher might decide that she wants to assess, score, and provide feedback on the following categories: organization, amount of information, quality of information, documentation of sources, and mechanics. The same teacher might decide that she wants to use a 4-point scale for each of her five categories for a total of 20 points. Table 1 shows a rubric that she might use that has her five categories labeled down the left side and scores across the top. Notice that this rubric also provides descriptions of the characteristics of performances at each score level for each category in the cells of the grid.

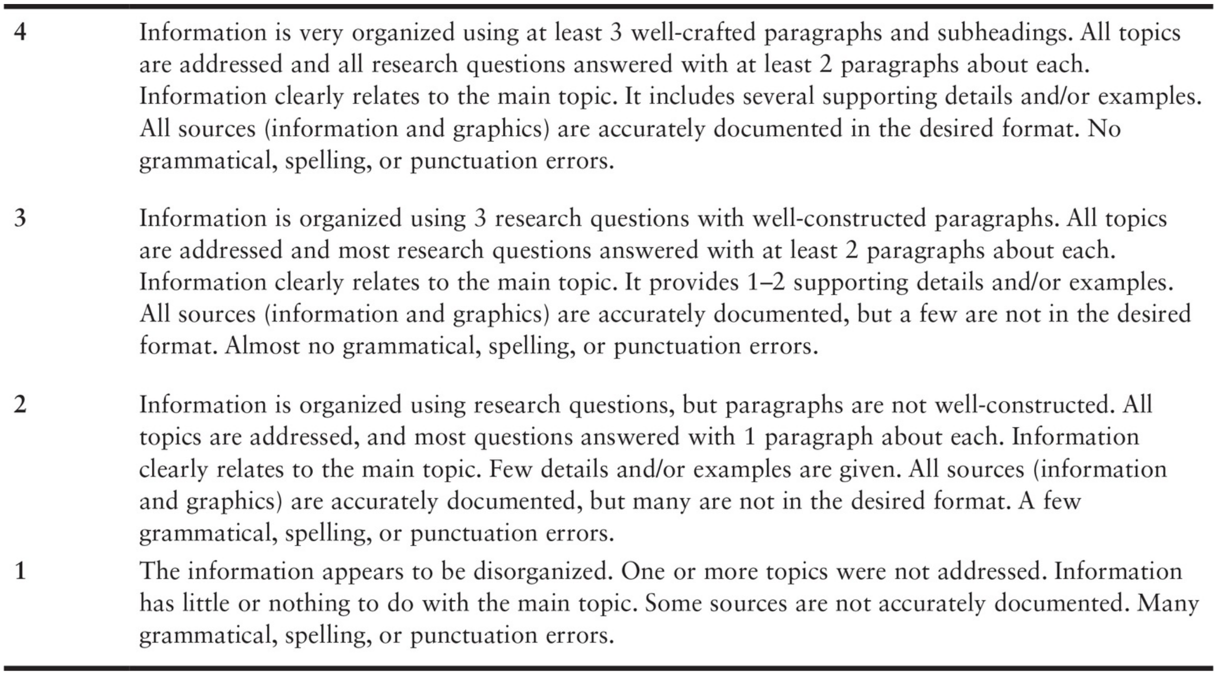
Student's Name: ______________________



Category	*4*	*3*	*2*	*1*
Organization	Information is very well organized using at least 3 well-crafted research questions with well-constructed paragraphs and subheadings.	Information is organized using 3 research questions with well-constructed paragraphs.	Information is organized using research questions, but paragraphs are not well-constructed.	The information appears to be disorganized.
Amount of Information	All topics are addressed and all research questions answered with at least 2 paragraphs about each.	All topics are addressed and most research questions answered with at least 2 paragraphs about each.	All topics are addressed, and most questions answered with 1 paragraph about each.	One or more topics were not addressed.
Quality of Information	Information clearly relates to the main topic. It includes several supporting details and/or examples.	Information clearly relates to the main topic. It provides 1–2 supporting details and/or examples.	Information clearly relates to the main topic. No details and/or examples are given.	Information has little or nothing to do with the main topic.
Documentation of Sources	All sources (information and graphics) are accurately documented in the desired format.	All sources (information and graphics) are accurately documented, but a few are not in the desired format.	All sources (information and graphics) are accurately documented, but many are not in the desired format.	Some sources are not accurately documented.
Mechanics	No grammatical, spelling, or punctuation errors.	No grammatical, spelling, or punctuation errors.	A few grammatical, spelling, or punctuation errors.	Many grammatical, spelling, or punctuation errors.

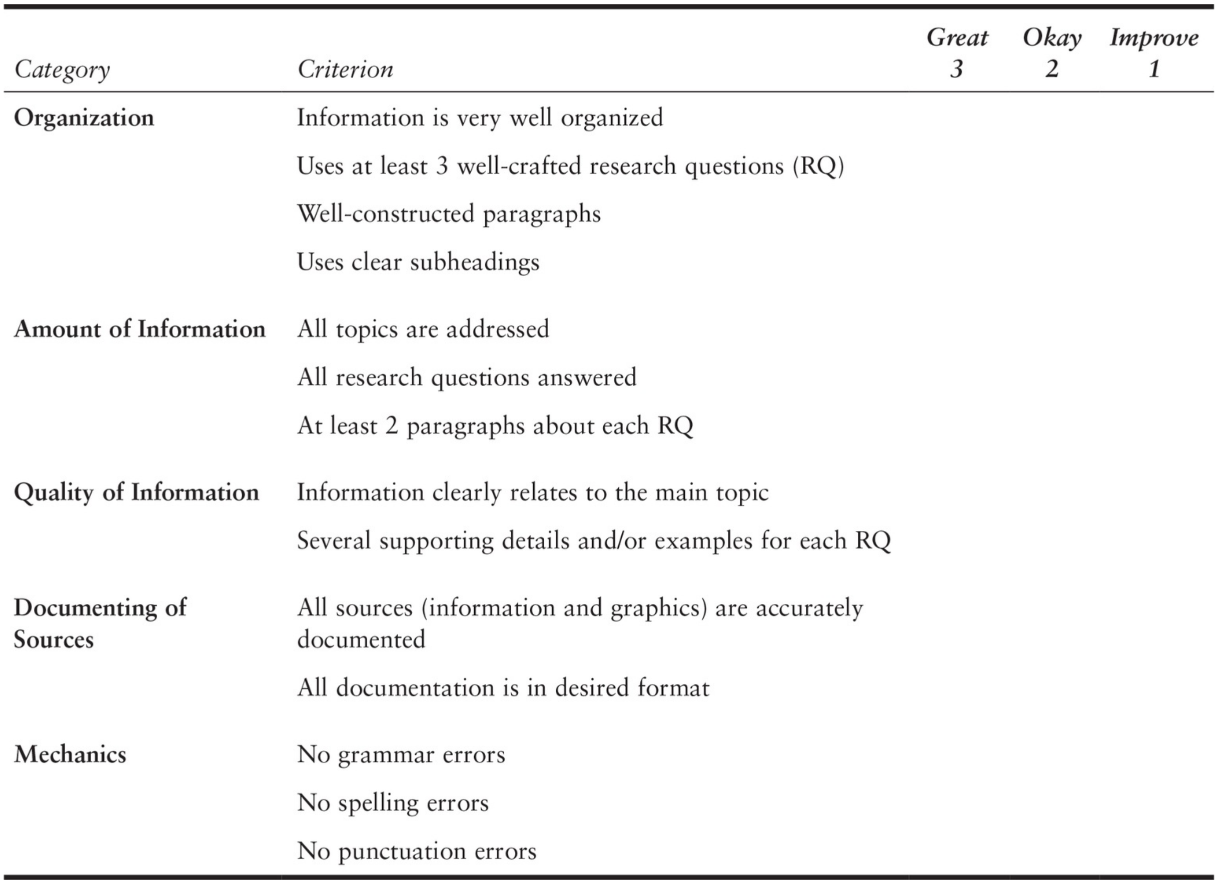
This rubric is an example of an *analytic rubric* because it assesses and analyzes multiple categories at one time, giving separate feedback on each in what is sometimes called analytical scoring. Of the different types of rubrics, analytic rubrics take the most time to apply because multiple judgments have to be made about different categories and levels. As a result, analytic rubrics tend to be used by teachers in small-scale assessment for classroom diagnostic, progress, and achievement assessments, wherein teachers feel it is worth the time and effort to provide students with useful feedback that will foster learning through feedback based on clear descriptions of their performances at different levels.

In contrast, the rubric shown in Table 2 is called a *holistic rubric* because it is designed to assign each student a single holistic score without breaking down the categories involved in what is often called holistic scoring. Holistic rubrics are relatively quick and efficient to apply because only one judgment needs be made for each student, but holistic rubrics do not provide as much feedback as analytic rubrics. As a result, holistic rubrics tend to be used in large-scale proficiency, placement, and achievement testing where the focus is on scoring quickly and efficiently, and feedback to the students (beyond reporting a single score) is not necessary.



4	Information is very organized using at least 3 well-crafted paragraphs and subheadings. All topics are addressed and all research questions answered with at least 2 paragraphs about each. Information clearly relates to the main topic. It includes several supporting details and/or examples. All sources (information and graphics) are accurately documented in the desired format. No grammatical, spelling, or punctuation errors.
3	Information is organized using 3 research questions with well-constructed paragraphs. All topics are addressed and most research questions answered with at least 2 paragraphs about each. Information clearly relates to the main topic. It provides 1–2 supporting details and/or examples. All sources (information and graphics) are accurately documented, but a few are not in the desired format. Almost no grammatical, spelling, or punctuation errors.
2	Information is organized using research questions, but paragraphs are not well-constructed. All topics are addressed, and most questions answered with 1 paragraph about each. Information clearly relates to the main topic. Few details and/or examples are given. All sources (information and graphics) are accurately documented, but many are not in the desired format. A few grammatical, spelling, or punctuation errors.
1	The information appears to be disorganized. One or more topics were not addressed. Information has little or nothing to do with the main topic. Some sources are not accurately documented. Many grammatical, spelling, or punctuation errors.

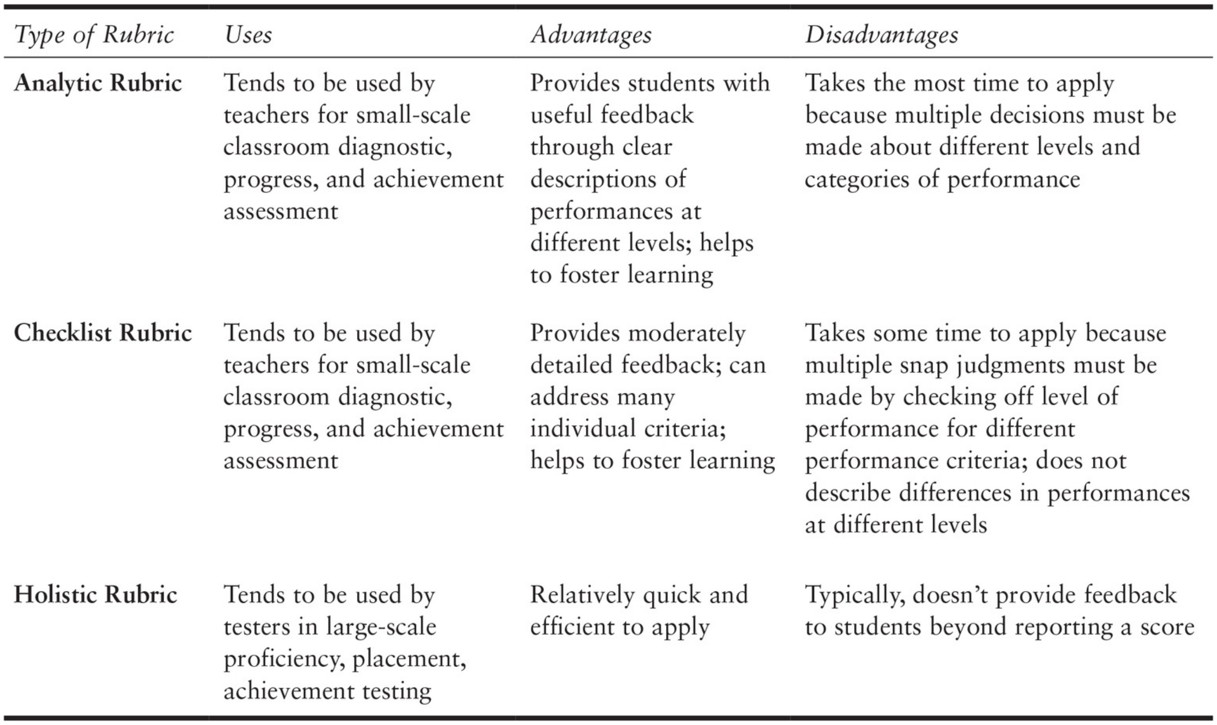
Table 3 shows a *checklist rubric* that represents a compromise somewhere between the detailed feedback provided by an analytic rubric and the efficiency of a holistic rubric. Checklist rubrics provide moderately detailed feedback (though generally not as detailed as holistic rubrics), but they also take some time to apply because the teacher needs to make multiple quick judgments and put a check mark next to each criterion. In addition, checklist rubrics do not describe differences in performances at different levels. For example, the checklist rubric shown in Table 3 provides feedback for 14 criteria in five categories about students' written reports, but only in terms of whether it was *great, okay*, or needs to *improve* (or scores of 3, 2, and 1).



Category	*Criterion*	*Great* *3*	*Okay* *2*	*Improve* *1*
Organization	Information is very well organized			
	Uses at least 3 well-crafted research questions (RQ)			
	Well-constructed paragraphs			
	Uses clear subheadings			
Amount of Information	All topics are addressed			
	All research questions answered			
	At least 2 paragraphs about each RQ			
Quality of Information	Information clearly relates to the main topic			
	Several supporting details and/or examples for each RQ			
Documenting of Sources	All sources (information and graphics) are accurately documented			
	All documentation is in desired format			
Mechanics	No grammar errors			
	No spelling errors			
	No punctuation errors			

Teachers and Rubrics

Notice that the rubrics in Tables 1–3 are all designed for scoring and/or giving feedback on students’ end-of-term written reports and therefore share a similar purpose (indeed, the words used in all three tables are very similar). However, the analytic, holistic, and checklist rubric formats offer different advantages and disadvantages. Table 4 summarizes these advantages and disadvantages for the three formats—primarily from a teacher/pedagogical perspective. Naturally, similar instruments could be developed for many other purposes related to any type of classroom performance, simulation, or task.



Type of Rubric	*Uses*	*Advantages*	*Disadvantages*
Analytic Rubric	Tends to be used by teachers for small-scale classroom diagnostic, progress, and achievement assessment	Provides students with useful feedback through clear descriptions of performances at different levels; helps to foster learning	Takes the most time to apply because multiple decisions must be made about different levels and categories of performance
Checklist Rubric	Tends to be used by teachers for small-scale classroom diagnostic, progress, and achievement assessment	Provides moderately detailed feedback; can address many individual criteria; helps to foster learning	Takes some time to apply because multiple snap judgments must be made by checking off level of performance for different performance criteria; does not describe differences in performances at different levels
Holistic Rubric	Tends to be used by testers in large-scale proficiency, placement, achievement testing	Relatively quick and efficient to apply	Typically, doesn't provide feedback to students beyond reporting a score

Generally, teachers will find rubrics especially useful for assessing and giving systematic written feedback to students on various performances, simulations, or tasks that teachers use in giving students opportunities to demonstrate their abilities or knowledges on material or skills learned in class. While multiple-choice, true-false, and matching items might be useful for assessing students' passive abilities or knowledges by requiring them to recognize and select from options supplied to them, assessing their abilities to actively apply skills and knowledge will require giving them opportunities to perform, to participate in a simulation, or to complete a task. It is in assessing such performances, simulations, or tasks that rubrics take on the most importance as tools that can usefully supplement the other more passive forms of assessment.

Fortunately, teachers need not develop rubrics from scratch. A number of online resources are available to facilitate designing and developing rubrics. For instance, the *Teachnology* website provides general links to example rubrics and rubric makers available elsewhere on the Internet. More specifically, the *Rubistar* website provides free tools for developing rubrics for oral projects, products, multimedia, science, research and writing, work skills, math, art, music, and reading—many of which can be adapted to other fields and purposes. After registering (for free) and exploring the website for a few minutes, creating the analytic rubric shown in Table 1 using the Rubistar templates took fewer than 10 minutes (based on selecting available descriptors). Naturally, copying the rubric to a Microsoft Word document and tailoring the descriptors to create Table 1 took a bit longer. The holistic rubric in Table 2 was then adapted from Table 1 by using Microsoft Excel to move things around. Next, the checklist in Table 3 was further adapted from Table 2 by using the design and layout menus and functions for tables in Microsoft Word.

Testers and Rubrics

Rubrics present additional challenges for large-scale, high-stakes test designers who want to create examinations that assess knowledges and abilities beyond the typical passive select-an-answer tests. Indeed, rubrics may prove indispensable for scoring performance tests of examinees' proficiencies or abilities to perform tasks or interact in simulations. For example, rubrics have been designed and used for scoring performance tests of various kinds appropriate for testing police officers, pilots, medical doctors, nurses, and students of all kinds—especially while they perform various tasks or in simulations. However, the application of rubrics in large-scale, high-stakes testing requires attending to several issues.

First, when rubrics are applied in large-scale, high-stakes testing, multiple raters are typically used to score examinee performances (either by summing or by averaging across raters). In such cases, rater training is often necessary in which rubrics can be used to

- guide the raters to score based on the same set of written performance characteristics,
- focus the raters during training to score on a single scale or set of scales,
- standardize raters' scoring during training as they practice being self-consistent and consistent with each other while rating sample examinee performances,
- allow raters to review the rubric during recalibration activities, and
- lead to demonstrably reliable scores by aligning the scoring practices of multiple raters.

Second, rubrics also present testers with several challenges related to analyzing the results statistically. Classical test theory analyses (e.g., mean, mode, median, standard deviation, and range) may prove adequate for examining the distributions of scores for a particular sample of examinees in a specific setting. However, care must be taken in calculating the classical test theory reliability of rubric-based scores because, by and large, the units of analysis are not simply right-or-wrong answers coded as 1 or 0, as would be appropriate when using Kuder–Richardson Formula 20 or 21. Because the scores tend to be based on units of analysis that are weighted scales (e.g., 1–4, 1–5, 1–20), testers interested in classical test theory reliability will need to consider using interrater reliability approaches or Cronbach's α.

In addition, where the resources and knowledge are available, testers can learn a great deal about the effectiveness of their rubric-based scales—especially if they are based on analytic or checklist rubrics—by using generalizability theory (G theory) or multifaceted Rasch analyses. G theory can prove particularly useful for determining how many raters, rating categories, and rating occasions, for example, might be maximally dependable when using a rubric to score a particular performance test. Multifaceted Rasch can be especially useful for spotting and adjusting for inconsistencies in various facets of measurement like raters who are particularly lenient, severe, or inconsistent; tasks or rating categories that are far too difficult or easy for the examinees involved; or inconsistencies in one facet across other facets (called bias interactions).

James Dean Brown

See also Analytic Scoring; Classical Test Theory; Generalizability theory; Holistic Scoring; Rasch Model

Further Readings

Arter, J., & McTighe, J. (2001). Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. Thousand Oaks, CA: Corwin.

Campbell Hill, B., & Ekey, C. (2010). The next-step guide to enriching classroom environments: Rubrics and resources for self-evaluation and goal setting for literacy coaches, principals, and teacher study groups, K–6. Portsmouth, NH: Heinemann.

Glickman-Bond, J., & Rose, K. (2006). Creating and using rubrics in today's classrooms: A practical guide. Norwood, MA: Christopher-Gordon.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation, 7(25). Retrieved October 26, 2010, from http://PAREonline.net/getvn.asp?v=7&n=25

Moskal, B. M. (2000). Scoring rubrics: What, when and how? Practical Assessment, Research & Evaluation, 7(3). Retrieved October 26, 2010, from http://PAREonline.net/getvn.asp?v=7&n=3

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. Educational Leadership, 55(2), 72–75.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. Practical Assessment, Research & Evaluation, 9(2). Retrieved January 9, 2011, from http://PAREonline.net/getvn.asp?v=9&n=2

Websites

Rubistar: http://rubistar.4teachers.org/index.php

Teachnology: http://www.teachnology.com/web_tools/rubrics/languagearts/

University of Hawai‘i at M¯anoa: https://manoa.hawaii.edu

S

Samantha F. Anderson Samantha F. Anderson Anderson, Samantha F.

Scott E. Maxwell Scott E. Maxwell Maxwell, Scott E.

Sample Size

Sample size

1441

1443

Sample Size

Sample size in the context of educational research refers to the number of participants in an experiment or study. The sample size has implications for how accurate the estimate of the effect under study will be (precision) as well as how detectable the effect will be (statistical power). This entry discusses the role sample size plays in both precision and power as well as how to plan the appropriate sample size for an educational study.

Precision

To understand how sample size is involved in the accuracy of an effect, consider a basic political poll. Suppose that 20 individuals sampled at random were interviewed about their choice between two different candidates. This type of study is modeled with the binomial distribution. If 12 of the 20 individuals reported a preference for Candidate A, is this sample proportion of .6 enough evidence to conclude that Candidate A has a majority lead in the population from which individuals were randomly sampled? To answer this question, the concept of standard error is needed. Right now, there is only a single sample of voters in the poll. Another sample of 20 different individuals could be collected, and the researcher can imagine that the new sample will probably not result in the exact same proportion of .6 preferring Candidate A. If an infinite number of polls were to be conducted, each based on a random sample of 20 individuals, each sample proportion could be plotted on a graph to create a sampling distribution. A large number of sample proportions from these polls would be clustered around the unknown true population proportion representing preference for Candidate A. But in reality, there is usually only one sample available. Where does this first sample proportion of .6 fall on this sampling distribution? This is where standard error comes in.

Standard error is the standard deviation of the sampling distribution or the square root of its variance. If the standard error is very large, and thus the sampling distribution is very wide, the single sample proportion could be a really inaccurate estimate of the preference for Candidate A. This is because this particular sample proportion could fall in the tails (extremes) of the sampling distribution, and in a wide sampling distribution, these tails are quite far away from the center. On the other hand, if the standard error is small, and thus the sampling distribution is narrow, the researcher can be more confident that this single sample proportion is a good guess at the true proportion of individuals in the population who prefer Candidate A, even if it does fall near the tails.

Thankfully, researchers do not actually need to conduct an infinite number of polls or plot a sampling distribution to determine the standard error. Only the single sample of 20 individuals is needed to calculate the estimated standard error for a proportion. The formula for the estimated standard error of a sample proportion is as follows: where is the sample proportion. So, how can researchers make their standard error small so the single sample can do a good job of accurately estimating the true population proportion? Notice the "*n*" in the denominator of the formula. The letter *n* is often used to denote sample size. Thus, by increasing the sample size, the standard error will decrease, and in return the estimate of the true population proportion will be more accurate.

To make this discussion concrete, suppose the researcher calculates the 95% confidence interval (CI) for a proportion: of the poll, where is the standard normal distribution *z* score associated with the desired confidence level (1.96 for a confidence level of 95%). A 95% confidence level is common, as it relates to the typical α level of .05 or the probability of declaring a result to be significant when it is null in reality. With only 20 respondents, this interval turns out to be [.39, .82]. A 95% CI specifies the interval in which, if researchers were to take 100 samples, in the long run 95 of them would contain the true population parameter (in this case, the true proportion of individuals preferring Candidate A). This means that, while the sample proportion was .6, the true population proportion could plausibly range from .39 to .82. Given that some of the interval extends below .5, it does not overwhelmingly rule out the possibility of majority preference for Candidate B. However, if the researcher instead had a much larger sample size of 200, the 95% CI narrows to .53, .67, which gives stronger evidence for Candidate A’s majority.

More generally, larger sample sizes are associated with greater precision in the estimation of the effect, as was shown with the shrinking CI in the political poll example. As a general rule, if one wants to double the precision in the estimate of an effect, it is not enough to double the sample size. Because “*n*” generally appears under a square root sign in the expression for a CI, one must quadruple the sample size in order to double the precision. Thus, large enough samples are important in educational research, where researchers and practitioners often want to get a trustworthy estimate of the size of the effect in question. And, as was illustrated in the example, the precision associated with an effect estimate can have consequences for the conclusions readers and policy makers can draw from a study.

Planning the Appropriate Size

However, in the political poll example, the researcher had already conducted a study when calculating its precision. It would be advantageous to know how large a sample needs to be to achieve a certain desired precision when researchers are in the planning stages of a study. This is where accuracy in parameter estimation approaches prove helpful. When planning a study, the researcher can specify the desired level of precision in advance. For example, perhaps the researcher involved in the introductory political poll would like to be 95% confident that the true proportion is within ±.1 of the sample proportion estimate. Using this *a priori* precision level, the researcher can determine the necessary number of participants to recruit for the study. For an interval around a single proportion, the formula for the appropriate sample size is , where is the researcher’s best guess at what the population proportion is, is the critical value for a standard normal distribution (1.96, if the planned confidence level is 95%), and ME is the desired margin of error, or ½ the width of the desired CI. In the current example, this would suggest 93 participants.

Similar formulas exist for other types of designs. For example, for the independent *t* test (for testing the mean difference between two groups on a criterion), an approximate formula for the appropriate per-group sample size when the desired precision is in standardized units is , where ME is the standardized margin of error, defined previously. Suppose a researcher expects a Cohen's *d* (a medium standardized effect size, the mean difference between the two groups divided by the standard deviation) of around .5, and wonders how large the sample needs to be to achieve a margin of error of .1 standard deviation units. Using the previous formula, the researcher would need about 800 participants per group to achieve this high level of accuracy.

Statistical Power

Complementary to precision and accuracy is the concept of statistical power. Some research questions pertain more to the existence of an effect, rather than its size. For example, perhaps a researcher is planning a study in the hopes of debunking a popular common sense theory in the field. Here, the direction of the effect, and the fact that it reverses the expected finding, is what is of note. In most educational studies, researchers hypothesize some sort of effect they expect to see, but in order for the effect to be detectable, the noise and error cannot blur the signal. How can researchers give themselves the best chance at detecting the effect they hope to see?

Power depends on effect size, α level, and sample size, and when any three of those four quantities are known, the fourth can be determined. The typical desired power value in educational research is 80%, but recent research suggests that even higher values are preferable. One factor that can increase statistical power is increasing the size of the effect in question. Although the effect size can sometimes be manipulated by using a stronger experimental treatment or more homogeneous groups, this factor is often immutable. However, just as with precision, increasing the sample size will increase statistical power, and the number of participants necessary to achieve a given level of power can be planned in advance. An approximate equation for determining the per-group sample size needed to reach a certain degree of statistical power for an independent *t* test is , where *d* is Cohen's *d* as defined previously, is the critical value from a standard normal distribution (1.96 for an α of .05), and is the critical value from the distribution where the alternative hypothesis is true (.84 for a desired power of .8). Using this approximation, 63 participants per group are needed to have 80% power to detect a medium-sized Cohen's *d* of .5.

In addition to the clear advantage of having a greater probability of detecting the effect of interest (and hence a greater chance of publication), high statistical power carries with it a host of additional benefits, such as more agreement among studies in the literature and a lower rate of false positive studies. Although researchers often set the probability of falsely rejecting the null hypothesis, given that it is true, to be .05, the reverse probability (the probability that the null hypothesis is true, given the study has statistically significant results) is not necessarily that low. The higher the statistical power of studies in the literature, the less likely it is that a study reporting a significant result has in fact mistakenly rejected a true null hypothesis. On the other hand, however, high statistical power can be difficult to achieve. The required sample sizes necessary for high degrees of power can be extremely large, especially if the effect in question is small in magnitude. Furthermore, large samples can not only be expensive to collect but also difficult to find when participants come from niche and minority populations. Despite these difficulties, though, the importance of sample sizes that can achieve high power and precision for both the individual researcher and the field cannot be overemphasized.

Samantha F. Anderson and Scott E. Maxwell

See also Confidence Interval; Experimental Designs; Power; Power Analysis; Type II Error

Further Readings

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2, e124. doi:10.1371/journal.pmed.0020124

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. Psychological Methods, 11, 363–385. doi:10.1037/1082–989X.11.4.363

Kraemer, H. C., & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychological Methods, 9, 147–163. doi:10.1037/1082–989X.9.2.147

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. Annual Review of Psychology, 59, 537–563. doi:10.1146/annurev.psych.59.103006.093735

Wainer, H. (2007). The most dangerous equation: Ignorance of how sample size affects statistical variation has created havoc for nearly a millennium. American Scientist, 95, 249–256. doi:10.1511/2007.65.249

Fei Gu Fei Gu Gu, Fei

SAS

SAS

1444

1447

SAS

SAS (pronounced “sass”) is a computer software package that began as a project at North Carolina State University to analyze large amount of agricultural data collected through U.S. Department of Agriculture grants. Due to the need to develop a general-purpose statistical software package for agricultural data, the resulting program, the Statistical Analysis System, was popularly known as SAS, the basis for the software name and the corporate (i.e., SAS Institute Inc., Cary, NC). Nowadays, SAS is a leading statistical package installed and used at different customer sites, including colleges and universities, governmental entities, pharmaceutical companies, and banks. The actual installation of the software can be customized at each site to satisfy different needs of the specific products. This entry describes the application of SAS in educational measurement and statistics. Specifically, this entry focuses on four SAS products—Base SAS, SAS/STAT, SAS/econometrics and time series (ETS), and SAS/IML—because the functions and procedures included in these products are frequently used by researchers in education.

It is not the intention of this entry to cover all the available tools provided by SAS, but an overview of what are commonly used is given. The content of this entry is organized as follows. First, some data management tools in SAS are introduced. Second, basic applications of statistical methods are briefly covered, which is immediately followed by more detailed discussions on advanced applications in educational measurement and statistics. Finally, this entry concludes with some useful resources. For simplicity and to avoid confusion, a particular SAS procedure will be referred to as PROC XXX instead of the XXX procedure throughout this entry.

Data Management

The prerequisite of any successful application of statistical method is to manage the collected data set, and SAS is developed for data analysis as well as data management. In order to use SAS for data analysis, the user has to first create the SAS data set within the SAS environment. Sometimes, existing SAS data sets are available, but some manipulations are required to produce the resulting SAS data set. To this end, this section introduces some commonly used data management tools provided by SAS.

First, one can use PROC IMPORT (in Base SAS) to import into the SAS environment the external data file of various formats, such as the space-, comma-, tab-, or any type of delimited file (e.g., *.txt,* .dat), Excel file (i.e., *.xls,* .xlsx), or SPSS file (i.e., *.sav). Reversely, the user can use PROC EXPORT (in Base SAS) to export the SAS data set as an external file of different formats. Technically, the features of PROC IMPORT and PROC EXPORT are implemented by the DATA step (in Base SAS). In other words, the user can use the DATA step with appropriate statements and options to import and export data. In addition to the import and export feature, the DATA step has many other features for data management within the SAS environment.

Second, the DATA step can copy or create SAS data sets, add or delete the variables, select the observations according to the customized rules, and assign built-in or user-defined formats to variables. Third, PROC structured query language (PROC SQL; in Base SAS) allows the user to implement SQL to manipulate the SAS data sets. Note that SQL is a standardized language for database management. If one is familiar with SQL, it is very convenient to use PROC SQL.

Fourth, for users comfortable with matrix algebra, SAS/IML may be the preferred choice to manipulate SAS data sets as matrices. Specifically, SAS/IML has only one procedure (i.e., PROC IML), which is an object-oriented programming language and provides a wide variety of statistical and text functions. In addition, new functions can be defined in PROC IML for any specialized purposes. In practice, one can use the user-defined function to accomplish challenging data management tasks and even in a repetitive manner. This feature is especially useful when a sequence of SAS data sets needs to be processed in the same but sophisticated way. However, one limitation of PROC IML is that it can store the SAS data sets only as matrices in computer memory, which renders certain tasks infeasible in PROC IML. In contrast, other SAS products can use both computer memory and hard drives to store the SAS data sets.

Finally, another useful tool for challenging and/or repetitive data processing is the macro language (in Base SAS). Typically, the macro language is used to implement a set of operations, and these operations can be defined in a customized loop for a sequence of SAS data sets and/or external data files. In practice, testing companies and governmental entities often need to process a lot of data when individual data sets are collected from each participant and the resulting data sets must be created from processing each of the data sets by a set of customized operations. SAS, unlike many other software packages, provides a good solution to this end.

Basic Applications

SAS/STAT provides a comprehensive coverage of basic statistical methods. All the basic methods taught in the standard curriculum for graduate students in education, which includes the courses in introductory statistics, analysis of variance, regression, and multivariate analysis, can be implemented by various procedures in SAS/STAT. For some procedures, their names are self-evident. That is, PROC ANOVA, PROC CLUSTER, PROC FACTOR, and PROC TTEST correspond with analysis of variance, cluster analysis, factor analysis, and *t* test, respectively. For some other procedures, their names can be easily associated with the methods, for example, PROC CANCORR, PROC CORRESP, PROC DISCRIM, PROC GLM, PROC LOGISTIC, PROC MDS, PROC PLS, PROC PRINCOMP, PROC QUANTREG, and PROC REG; implements canonical correlation analysis; correspondence analysis; discriminant analysis; general linear model; logistic regression; multidimensional scaling; partial least squares regression; principal component analysis; quantile regression; and linear regression, respectively. However, there are few exceptions such that the name of the procedure is not immediately obvious, but still reasonable, for the statistical method. For example, the chi-square independence test for a two-way table needs to be requested by the CHISQ option in the TABLES statement of PROC FREQ. Note that PROC FREQ is the procedure that is included in both Base SAS and SAS/STAT. Also, there are at least three other procedures in Base SAS that implement the basic methods. That is, PROC CORR, PROC MEANS, and PROC UNIVARIATE can calculate descriptive statistics such as Pearson correlation, mean, variance, standard deviation, skewness, and kurtosis. Note that PROC CORR also computes Cronbach’s coefficient α if the ALPHA option is invoked in the PROC CORR statement.

Advanced Applications

This section describes some procedures and functions in Base SAS, SAS/STAT, SAS/ETS, and SAS/IML for advanced methods in educational measurement and statistics, including item response theory, multilevel models, structural equation models, and time series models. The procedures included in this section do not exhaust everything that SAS can do, but it is representative of the methods widely used in educational measurement and statistics.

Item Response Theory

Starting SAS 9.4, PROC IRT (in SAS/STAT 14.1) is introduced as a new procedure to estimate item response theory models. The basic features of PROC IRT enable the user to estimate the Rasch model, one-, two-, three-, and four-parameter models, graded response model, and generalized partial credit model. Moreover, PROC IRT allows both unidimensional and multidimensional models to be estimated as well as performing multiple-group analysis with fixed values and equality constraints imposed within and between groups. In addition, the available factor score estimation includes maximum likelihood, maximum a posteriori, and expected a posteriori methods. Because PROC IRT is a very new procedure, it is expected that more features will be added to PROC IRT to catch up with other existing software packages.

Multilevel Models

The SAS/STAT documentation refers to the random-effects model and random-coefficients model as mixed models. Mixed models are appropriate to fit data with nested structures, for example, students nested within schools or repeated measurements nested within participants. Multiple procedures are available in SAS/STAT to fit various mixed models. Specifically, PROC MIXED is designed to fit the linear mixed models. If nonlinear relationships are the focus, one can use PROC NLMIXED. For the categorical outcome variable in generalized linear mixed models, PROC GENMOD and PROC GLIMMIX can be used. Moreover, there are at least two high-performance procedures for mixed models in SAS/STAT 14.1: PROC HPLMIXED and PROC HPMIXED. The high-performance procedures are designed specifically for analyzing big data sets, which might be the only feasible solution in certain situations. However, the advantage of PROC HPLMIXED and PROC HPMIXED in handling big data sets, compared to PROC MIXED and PROC GLIMMIX, comes with certain limitations. Detailed comparisons among these procedures can be found in the SAS/STAT documentation.

Structural Equation Models

Models for the analysis of covariance structures (sometimes with mean structures involved as well) are popularly known as the structural equation models. PROC CALIS (in SAS/STAT 14.1) offers seven different modeling languages (i.e., COSAN, FACTOR, LINEQS, LISMOD, MSTRUCT, PATH, and RAM) for specifying a very wide class of structural equation models. In practice, it is not necessary to learn all the modeling languages. The purpose of providing different modeling languages is that different researchers may have learned different modeling languages adopted by other software packages such as EQS, LISREL, or Mplus. Thus, users of other software packages may quickly adapt themselves with PROC CALIS. A relatively new feature of PROC CALIS is that, starting with SAS/STAT 13.1, high-quality graphical output of path diagrams can be generated from PROC CALIS. Moreover, the generated path diagrams can be edited in the ODS Graphics Editor (in SAS/GRAPH), which is an interactive GUI-based tool for editing and customizing plots, to cater to publication needs. Like PROC IRT, new features were, are, and will continue to be, added to PROC CALIS in future versions.

Time Series Models

Time series or process analysis is not typically taught to graduate students in education, but this type of data analysis does have appealing and sound theoretical foundations in educational research (e.g., cognitive processes, functional magnetic resonance imaging). Traditionally, this branch of statistics is very active in econometrics. Therefore, the procedures developed for time series analysis are included in SAS/ETS. In the past decades, there are quite a few procedures in SAS/ETS to analyze time series data, including PROC ARIMA, PROC AUTOREG, PROC FORECAST, PROC MODEL, PROC PANEL, PROC UCM, and PROC VARMAX. Particularly, a new procedure (i.e., PROC SSM) was introduced in SAS 9.3 (SAS/ETS 12.1), which enables linear state space modeling of time series and longitudinal data. From both applied and theoretical researchers, PROC SSM is an extremely useful procedure because many time series models and structural equation models can be reformulated as special cases of state space model. Applications of the procedures in SAS/ETS are as easy as those in SAS/STAT, as long as the user is familiar with the corresponding statistical methods.

Besides the procedures in SAS/ETS, there are some useful functions and subroutines for time series analysis provided by SAS/IML such as ARMASIM, KALCVF, KALCVS, KALDFF, KALDFS, and VARMASIM. For statisticians developing new methods, the functions and subroutines in PROC IML may be incorporated in customized programs to test new models, including time series models. This sort of research often requires Monte Carlo simulation, in which the user may need several SAS products: macro language and the DATA step in Base SAS, SAS/IML, and/or some procedures in SAS/ETS.

Useful SAS Resources

Overall, SAS is an enormous package. For an encyclopedia entry, it is impossible to cover all aspects of SAS applications. However, users can explore and teach themselves how to use SAS from various SAS resources. First, SAS has a technical support page where the user can find SAS documentation, SAS papers, and other information. Alternatively, one can always send e-mails to support@sas.com to seek solutions. Second, SAS holds an annual conference in April every year, called SAS Global Forum, for users around the world. The conference accepts various manuscripts ranging from statistics and data analysis, clinical trials, text analytics to any traditional or innovative SAS development and applications. The conference is an opportunity for users to network with SAS developers and other SAS users. For a student whose manuscript is accepted, the student may be selected as one of the 10 SAS Student Ambassadors every year, such that SAS will cover all costs associated with the conference (e.g., transportation, registration, and meals). For university professors, SAS provides two SAS books upon request through the academic evaluation copy program.

Fei Gu

See also EQS; LISREL; R

Websites

SAS Support: https://support.sas.com

Brenda Hannon Brenda Hannon Hannon, Brenda

SAT

SAT

1447

1451

SAT

The SAT is a standardized test that is widely used for college and university admissions in the United States. Created by the College Board, the SAT is intended to assess a student's readiness for college, and in theory, it furnishes colleges and universities with a common criterion for comparing applicants. Since its debut in 1926, the name and the scoring of the SAT have changed multiple times; it was called the Scholastic Aptitude Test and then the Scholastic Assessment Test before becoming simply the SAT. This entry first discusses the structure and scoring of the SAT. It then looks at research on how well the SAT predicts college success, which constructs predict how well students will perform on the SAT, and gaps in performance on the SAT among different groups of students.

Structure of the SAT

The College Board announced in 2014 that it would be overhauling the SAT, and the new version was administered for the first time in 2016. The redesigned SAT includes two mandatory sections: (1) Math and (2) Reading, Writing, and Language. It also includes one optional written essay. Students are allotted 3 hours to complete the two mandatory sections and an additional 50 minutes to complete the optional essay.

Generally speaking, the questions in both the Math and Reading, Writing, and Language sections range from easy to hard. Although easier questions frequently appear near the beginning of a section and more difficult questions frequently appear near the end of a section, this format is not necessarily true for all sections.

In addition, according to the College Board, the redesigned SAT assesses skills that are more predictive in college and beyond. For instance, it places more emphasis on reasoning skills in context (e.g., inferring meanings of words from context, editing a passage) rather than skills in isolation (e.g., what is the definition for *fecund*?).

The Reading, Writing, and Language section includes two subtests: one subtest that assesses reading and a second subtest that assesses writing and language. Like its predecessors, the Reading subtest uses multiple-choice questions; however, unlike its predecessors, more emphasis is placed on extracting, thinking about, and interpreting information from passages.

According to the College Board, the questions in the Reading subtest are analogous to those asked in a lively, thoughtful, evidence-based debate. That is, these new questions assess when a student has command of the evidence. Some questions might directly ask a student to locate a specific piece of information, such as finding evidence in a passage that supports an answer, identifying how authors use evidence to support their claims, or finding a relationship between the passage and its accompanying graphics. Other questions will ask a student to understand what is implied (e.g., use contextual clues to infer the meaning of a word and decide how word choices shape meaning, style, and/or tone of a passage) or analyze content stated or implied by a passage (e.g., assess hypotheses, interpret data, and consider implications).

The Writing and Language subtest also uses multiple-choice questions and passages with accompanying graphics, but unlike its predecessors, the Writing and Language subtest assesses a student's ability to edit and improve passages, including passages that include deliberate errors. More specifically, the Writing and Language subtest assesses three skills that are used while generating a paper, namely reading, finding mistakes or weaknesses, and fixing mistakes or weaknesses. Some questions assess a student's command of evidence (e.g., improving how a passage develops information and ideas); other questions assess improvement of word choices (based on words surrounding the to-be-replaced word), expression of ideas (e.g., identify which words or phrases improve how well a passage makes its point), and standard English conventions (e.g., verb tenses, parallel construction, and subject-verb agreement).

The Math section includes two subtests, one that allows students to use calculators and one that does not. Like other sections of the SAT, the Math section uses multiple-choice questions to assess a student’s knowledge about mathematics. However, the Math section also includes questions, called “grid-ins,” that require students to generate answers and report their exact answers rather than select answers from among choices in multiple-choice questions.

According to the College Board, the Math section assesses the use of Math knowledge that students frequently employ in a variety of situations, rather than knowledge of every Math topic. More specifically, it completes an in-depth assessment of the three areas of Math that contribute the most to a wide range of college majors and future professional careers: (1) central concepts of algebra (i.e., mastery of linear equations and systems), (2) problem solving and data analysis (i.e., being quantitatively literate), and (3) advanced mathematics (i.e., manipulation of complex equations).

The Math section also assesses other topics in Math (e.g., geometry and trigonometry) that are most relevant to college and career readiness. Some questions assess students’ fluency in these topics (e.g., how to execute procedures accurately, efficiently, and flexibly), whereas other questions assess students’ conceptual understandings of Math concepts, operations, and/or relations and students’ abilities to analyze situations and identify the critical elements necessary to solve the problem.

Finally, the Optional Essay section requires students to use their reading, analysis, and writing skills. According to the College Board, the new Optional Essay has been totally remodeled. For instance, it is no longer mandatory, students have 50 minutes to complete it (instead of 25 minutes), and it is no longer a position essay, where students must agree or disagree with a position. Rather, the new Optional Essay is much like a typical college writing assignment in which students must analyze a text. So, students will need to read passages, explain how author(s) build their argument, and then support their explanation with evidence from the passage.

Scoring

Scoring on the redesigned SAT is based on the number of questions that are answered correctly. All questions are weighted equally and, most importantly, unlike with its predecessors, there is no penalty for guessing.

Students receive an overall SAT score and two scores based on the two mandatory sections: Math and Reading, Writing, and Language. Raw scores for the two mandatory sections are converted to scaled scores that range from 200 to 800. Because overall scores on the SAT are a composite of the Math and Reading, Writing, and Language sections, the range of overall scores is from 400 to 1600 (i.e., 200 + 200 = 400; 800 + 800 = 1,600). Percentage ranks are reported for all three of these scores, where the 50th percentile is an average ranking.

The redesigned SAT also includes three scores for the optional essay section: (1) Reading, (2) Analysis, and (3) Writing, which each range from 2 to 8.

There are also two cross-test scores: (1) Analysis in history/social studies and (2) Analysis in science that report how well students performed on items focusing on these subject areas in the Reading, Writing, and Language and Math tests. The scores for these two cross-test sections range from 10 to 40.

Finally, there are seven subscores based on questions taken from different sections. For instance, the four subscores for the command of evidence, words in context, expression of ideas, and command of standard English conventions are generated from the Reading, Writing, and Language section. The three subscores for Heart of Algebra, Passport to Advanced Math, and Problem Solving and Data Analysis are generated from the Math section. The scores for each of these subtests range from 1 to 7.

Predicting College Success

Because of its newness, there is limited research on how well the redesigned SAT predicts college success.

Brent Bridgeman, Laura McCamley-Jenkins, and Nancy Ervin looked at an earlier version of the SAT in a 2000 analysis and reported that it reliably explains the academic performance of college/university students. More recent studies substantiate this claim by suggesting that SAT scores can explain as much as 13.5–24.3% of the variance in freshman GPA.

Studies also suggest that the combination of SAT scores with high school GPA is a better predictor of freshman GPA than either the SAT or high school GPA alone. A 2005 study by Rebecca Zwick and Jeffrey Sklar, for instance, showed that when combined, SAT scores and high school GPA accounted for 22% of the variance in freshman GPA for 4,617 University of California students. However, because SAT scores and high school GPA also correlate, Zwick and Sklar also observed individually, high school GPA accounted for 20.5% of the variance in freshman GPA, a finding that suggests that SAT scores only account for an additional 1.5% of the variance in freshman GPA (i.e., 22.0 = 20.5 + 1.5).

Constructs That Predict SAT Performance

Over the years, researchers have persistently criticized the SAT for lacking construct validity, which is the degree to which the construct being investigated is accurately measured and interpreted. A considerable body of research has addressed this concern and proposed a plethora of factors to account for the variance in SAT scores. Some of these factors include social psychology measures (e.g., self-efficacy, test anxiety, locus of control, and performance-avoidance goals), personality measures (e.g., conscientiousness and level of anxiety), measures assessing specific cognitive abilities (e.g., knowledge integration), measures assessing general cognitive abilities/capacity (e.g., working memory and general intelligence), measures assessing knowledge about learning (i.e., metacognitive/metalearning knowledge), and measures assessing socioeconomic factors (e.g., family income and level of parental education).

Brenda Hannon and Mary McNaughton-Cassill examined the relative contributions of social/personality and specific cognitive/learning factors to SAT performance in a study published in 2011. Their research suggests that measures of social/personality factors (e.g., performance-avoidance goals, test anxiety, and locus of control) account for 21.4% of the variance in SAT performance and that measures of cognitive/learning factors (e.g., knowledge integration, epistemic belief of learning, and working memory) account for 37.8% of the variance in SAT performance. Moreover, when the predictive powers of these social–personality and cognitive/learning are combined, they can account for as much as 43.4% of the variance in SAT performance.

Additionally, research examining the predictive power of general intelligence has indicated that general intelligence consistently accounts for a large amount of the variance in SAT performance. In 2004, Meredith Frey and Douglas Detterman observed that depending on the measure of general intelligence used, differences in general intelligence accounted for 28.1–67.2% of the variance in SAT performance. More recent research suggests three factors predict SAT performance: a general cognitive factor (measured by knowledge integration, working memory, and general intelligence), a learning factor (i.e., metacognitive learning), and a social–personality factor (i.e., measured by performance-avoidance goals and test anxiety).

Still other research promotes socioeconomic factors, such as family income and parental education, as predictors of SAT scores; although the amount of variance in SAT performance that is accounted for by these factors is considerably less than that accounted for by the aforementioned cognitive and social–personality factors. For instance, research has shown that parental education and family income accounted for 6.3% and 4%, respectively, of the variance in SAT performance.

Performance Gaps

Since its first administration in 1926, the SAT has been shrouded in controversies, including those about score differences between male and female students and between students of different racial and ethnic groups. Females have routinely scored lower than males on both the verbal and Math sections of the SAT. During the period from 1987 to 2006, the average verbal SAT scores for females and males were 501 and 508, respectively, and the average Math SAT scores were 492 and 528, respectively. Although the causes of these differences are still under debate, recent research suggests that the social–personality factors of test anxiety and performance-avoidance goals account for all variance in SAT performance that is attributed to gender differences.

There are also gaps in performance among students of different racial/ethnic groups on both the verbal and Math sections of the SAT. For example, during the period from 1987 to 2006, the average Math SAT scores for Hispanic versus European American students were 460 and 523, respectively, and the average verbal SAT scores were 456 and 526, respectively.

Like the gender gap, the cause(s) of the racial/ethnic minority gap in SAT performance is poorly understood. Explanations such as stereotype threat and socioeconomic background factors (e.g., family income and parental education) have been proposed. However, more recent research suggests that 55–75% of the ethnic gap between Hispanics and European Americans in SAT scores can be attributed to metacognitive awareness, performance-avoidance goals, and level of parental education. This finding remained true even when gender differences were controlled.

Brenda Hannon

See also Achievement Tests; ACT; College Success; *g* Theory of Intelligence

Further Readings

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). Predictions of freshman grade-point average from the revised and recentered SAT I reasoning test (College Board Report No. 2000–1). New York, NY: College Entrance Examination Board. Retrieved from https://research.collegeboard.org/publications/content/2012/05/predictions-freshman-grade-point-average-revised-and-recentered-sat-i

Cloud, J. (2001). Should SATs matter? Time, 157, 62–76.

Frey, M. C., & Detterman, D. K. (2004). Scholastic achievement or g? Psychological Science, 15, 373–378.

Hannon, B. (2012). Test anxiety and performance-avoidance goals explain gender differences in SAT-V, SAT-M, and overall SAT scores. Personality and Individual Differences, 53, 816–820.

Hannon, B. (2015). Hispanics’ SAT scores: The influences of level of parental education, performance-avoidance goals, and knowledge about learning. Hispanic Journal of Behavioral Sciences, 37, 204–222. doi:10.1177/0739986315573249

Hannon, B., & McNaughton-Cassill, M. (2011). SAT performance: Understanding the contributions of cognitive/learning and social/personality factors. Applied Cognitive Psychology, 25, 528–535.

Kobrin, J. L., Sathy, V., & Shaw, E. J. (2007). A historical view of subgroup performance difference on the SAT reasoning test. (Research Report No.

2006–5). New York, NY: College Board Publications.

Mattern, K. D., Shaw, E. J., & Williams, F. E. (2008). Examining the relationship between SAT, high school measures of academic performance and socioeconomic status: Turning our analysis to unit of measure. New York, NY: College Board Publications.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. Psychological Bulletin, 130, 261–288. Retrieved from http://dx.doi.org/10.1037/0033–2909.130.2.261

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69, 797–811.

Zwick, R. (2012). Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana, 49(2), 23–30.

Zwick, R., Brown, T., & Sklar, J. C. (2004). California and the SAT: A reanalysis of University of California admissions data (Research and Occasional Papers Series). Berkeley, CA: Center for Studies in Higher Education, UC Berkeley. Retrieved January 11, 2013, from http://cshe.berkeley.edu/publications/rops.htm

Zwick, R. & Green, J.G. (2007). New perspectives on the correlations of SAT scores, high school scores, and socioeconomic factors. Journal of Educational Measurement, 44, 23–45. Retrieved from http://dx.doi.org/10.1111/j.1745–3984.2007.00025.x

Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. American Educational Research Journal, 42,

439–464.

Mary M. Chittooran Mary M. Chittooran Chittooran, Mary M.

Scaffolding

Scaffolding

1451

1454

Scaffolding

The first formal definition of the term *scaffolding* was offered in 1976 by Jerome Bruner and his associates who described it as a “process that enables the child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts” (p. 90). Scaffolding involves an adult or teacher who provides supports for students in order to facilitate learning and to aid in task mastery. The teacher systematically builds on students’ experiences and knowledge as they are learning new skills and then gradually withdraws supports as they achieve mastery. This topic has taken on increasing relevance in today’s classrooms, with a growing focus on teacher–student interactions and their role in effective instruction. Scaffolding also relates to current schools of thought such as social constructivism, differentiation of instruction, and student-centered learning, all of which are characterized by flexibility, sensitivity, and accommodation to students’ needs during the learning process. This entry describes the nature and characteristics of scaffolding, provides examples of how it is used in the classroom setting, discusses some of the benefits and challenges related to its use, and finally, addresses future directions in the use of scaffolding.

Nature and Characteristics

A quick survey of the educational literature suggests that scaffolding has become an increasingly popular topic since the 1970s. To understand how scaffolding works, it might be instructive to think of a building that is constructed with temporary structures in place that support it as long as it lacks the integrity to support itself. The scaffolding is gradually removed as the building becomes more stable and can stand alone. In the same way, instructional scaffolding provides learners with support as new knowledge or skills are constructed and is withdrawn as learners become independent enough to stand on their own.

The analogy of a parent teaching a child to ride a bike may also be helpful in this context. Initially, the parent might model for the child how to get on the bike (perhaps one with training wheels) and demonstrate how to pedal to get the bike moving. Gradually, the child is encouraged to get on the bike and start pedaling, with the parent holding on to the bike. As the child continues to practice and becomes more confident, the training wheels are removed and parental support is gradually withdrawn, until balance has been achieved and the child is able to ride independently. If the parent lets go too soon, the child may crash into a neighboring tree; if the parent keeps the training wheels on or holds on longer than is necessary, there is the danger of limiting the child's sense of autonomy and confidence. Throughout the process, the parent allows for both modeling of desired behaviors and the practice of the newly learned skills. The parent–child interaction changes through the experience, with the parent providing direct, firm support at the beginning, to finally withdrawing support altogether, but being available to the child should parental assistance be sought.

Scaffolding is related to the social constructivistic approach of Lev Vygotsky's zone of proximal development, the gap between the point at which students need adult assistance with their learning tasks and the point at which they can work unassisted on those tasks. Teaching in this zone requires social interaction and communication between a "more knowledgeable other" and the learner. Scaffolding is also related to the gradual release of responsibility model described by David Pearson and Margaret Gallagher in 1983, in which teachers initially shoulder the main responsibility for learning tasks and then finally allow students to take on responsibility for their own learning as they become more competent (I do, We do, You do). At the most basic level, scaffolding reflects what is now thought of as good teaching: the ability to be flexible and make accommodations to serve the unique needs of each learner.

Scaffolding is most useful for new learning or for a complex task that can be broken down into various steps. It is based on the continuous assessment of a learner's progress and the provision of scaffolds can be decreased, changed, or increased, depending on the learner's needs; however, the ultimate goal of scaffolding is always to develop independent learners. Keith Sawyer has used the term *instructional scaffolding* to describe the provision of various supports that promote deep learning in students. These supports may include direct teaching, materials and resources, templates, specific guidance on skill development, modeling, coaching, and feedback. Although scaffolding may be used in a number of ways, the method of delivery will vary depending on the unique nature of the tasks and the needs of the students.

There are three essential characteristics of scaffolding. The first relates to the teacher–student interaction, which should be communicative, collaborative, and reflect a shared responsibility for learning. The second has to do with the fact that learning must take place in the student’s zone of proximal development so as to maximize learning. This involves a knowledgeable teacher who is able to evaluate the student’s functioning, including strengths and weaknesses, and identify the level at which instruction ought to occur. If it is too advanced, the student may become frustrated; if it is too simple, the student runs the risk of getting bored. The third characteristic of scaffolding is that the scaffold, or the support and guidance provided by the teacher, is gradually removed as the learner becomes more competent and independent of the teacher.

Eunbae Lee and Michael Hannafin, who offer a comprehensive look at scaffolding under the umbrella of student-centered learning, describe various kinds of scaffolding such as conceptual (the knowledge basis), procedural (the how-tos), strategic (choosing between alternative strategies for problem solution), and metacognitive (evaluating one’s own learning). Other authors have offered descriptions of task, content, and material scaffolding. The following is an example of scaffolding in the case of an instructor teaching a graduate-level research course in survey development: The instructor provides conceptual and content scaffolding when she discusses the purpose, rationale, and knowledge base for using surveys in gathering research data; task scaffolding, when she breaks down the task of survey development into its component parts; procedural scaffolding, when she models a step-by-step procedure to develop and evaluate student surveys; material scaffolding, when she offers the class sample student-developed surveys to use as a guide; strategic scaffolding, when she helps them modify survey items to minimize bias; and metacognitive scaffolding, when she asks them to evaluate and reflect on their completed surveys.

Use in Classroom Settings

There is no one correct way to scaffold instruction, and there appears to be no standard protocol for effective scaffolding. Part of the reason for this, of course, is that effective scaffolding has to do with individualizing instruction, and therefore, there are as many ways to scaffold as there might be students in a classroom. Teachers might use common sense, trial and error, the benefits of their teaching experience, and knowledge of their students' needs to determine the type of scaffolding that will be required.

Scaffolding typically involves several steps, the first of which involves identifying the learning objective or the task. It is important that the task not only be engaging, but that it identifies each skill to be learned. A second step involves the anticipation of errors that the learner might make during the learning process so that teachers can guide students away from making those errors. A third step involves the implementation of the scaffolding, with various types of scaffolding being provided, depending on the needs of the student. Finally, the teacher must consider affective factors such as the learner's self-confidence and anxiety level and offer emotional scaffolding strategies such as encouragement, coaching, and guidance.

Scaffolding involves two important aspects: The first is modeling, which allows students to observe the teacher demonstrating each step of the learning task or strategy to see how it is done, and practice, both guided and independent, where students, working individually or in a group, are able to work collaboratively with their teachers to practice the task or strategy they are learning. Effective scaffolding may also involve error detection and correction that, along with self-talk, is first modeled by teachers to show students how to handle errors and to help them see how self-talk can be used to get them "unstuck" after they have made an error.

The Japanese concept of *kikan-shido*, or "teaching between desks," is essentially an application of the scaffolding approach used with students during independent seatwork that involves a teacher who informally questions, coaches, and guides students as they practice newly learned skills independently at their desks. The intentional use of kikan-shido is thought to elicit deeper learning in students that is far superior to the kind of learning that occurs with traditional seatwork.

Researchers have also determined that the use of scaffolding during teaching can result in not only increased learning, not just of the academic variety but also affective learning; in fact, in one study, affective outcomes were documented through the use of scaffolding with Latino/a youth in high-risk, urban school settings. Thilo Kleikmann and his associates, in 2016, used expert scaffolding settings. Thilo Kleickmann and his associates, in 2016, used expert scaffolding during professional development for elementary school teachers preparing to teach science and found that it was markedly superior to professional development through self-study alone not only in terms of content learning but also in terms of teachers’ motivation and beliefs, as well as in the quality of instruction and student outcomes.

A review of the literature suggests a vast array of applications where scaffolding has been shown to be useful. It can be used in most instructional areas (e.g., reading, history, computer science, foreign language, and STEM subjects), to teach many skills (e.g., reading comprehension, writing a research paper, assembling a science fair project, and estimating probability), to learners of all age levels (preschool through adulthood), and ability levels (e.g., average intelligence, gifted, and learning disabilities).

Benefits and Challenges

A considerable benefit of using scaffolding is that learning can take place gradually, at the pace that is most comfortable for the student. Effective scaffolding builds self-confidence, a sense of self-efficacy, trust in teachers and other adults, and independence. It minimizes frustration and the fear of failure which, in turn, may lead to greater success in learning.

Although scaffolding confers significant benefits in the classroom, it is by no means fully accepted by everyone. For one thing, effective scaffolding that depends on reciprocal interactions requires more from teachers than traditional teaching. It requires constant observation, monitoring and evaluation of student progress, planning and implementation of scaffolding strategies, individualizing instruction for students, and a greater investment of energy. It increases the responsibility for shared accountability between teacher and student, and it requires teacher and student to communicate with each other. It requires understanding that not only does the difficulty level of tasks or strategies vary, but so too, do students’ needs and skill levels. It may simply take more time to teach the same concepts than it would using a traditional mode of delivery and for many traditionally trained teachers, scaffolding may represent the kind and amount of work that they are not prepared to do.

Future Directions

Scaffolding is an approach to learning that allows for experienced, knowledgeable teachers to support learners through the learning process with the ultimate goal of making them independent learners. While there are certainly challenges involved in incorporating scaffolding into the classroom, to many educators the benefits that it confers far outweigh its attendant difficulties. Teachers in training as well as current teachers can be taught about the benefits of scaffolding and given practice in incorporating scaffolding into their teaching. Changes in how scaffolding is delivered may occur; for example, with technological advances, it may be that virtual scaffolding will supplant face-to-face scaffolding someday. Finally, the practice of scaffolding can be further studied in different disciplines and with different populations so that it can be utilized more fully in learning settings and thereby benefit both teachers and learners.

Mary M. Chittooran

See also Zone of Proximal Development

Further Readings

Clark, K. F., & Graves, M. F. (2005). Scaffolding students' comprehension of text. The Reading Teacher, 58(6), 570–580. doi:10.1598/RT.58.6.6

Hogan, K., & Pressley, M. (Eds.). (1997). Scaffolding student learning: Instructional approaches and issues. Cambridge, MA: Brookline Books.

The IRIS Center. (2005). Providing instructional supports: Facilitating mastery of new skills. Retrieved on August 24, 2016, from http://iris.peabody.vanderbilt.edu/module/sca/

Kleickmann, T. Trobst, S. Jonen, A. Vehmeyer, J., & Moller, K. (2016). The effects of expert scaffolding in elementary science professional development on teachers' beliefs and motivations, instructional practices, and student achievement. Journal of Educational Psychology, 108(1), 21–24.

Lee, E., & Hannafin, M. J. (2016). A design framework for enhancing engagement in student-centered learning: Own it, share it, and learn it. Education Tech Research Development, 64, 707–734. doi:10.1007/s11423–015-9422-5

Pearson, P. D., & Gallagher, M. (1983). The instruction of reading comprehension. Contemporary Educational Psychology, 8, 317–344.

Sawyer, R. K. (2006). The Cambridge handbook of the learning sciences. New York, NY: Cambridge University Press.

Stone, C. A. (1998). The metaphor of scaffolding: Its utility for the field of learning disabilities. Journal of Learning Disabilities, 31, 344–364.

Valkenburg, J. (2015). Joining the conversation: Scaffolding and tutoring mathematics. Learning Assistance Review (TLAR), 20(2), 33–45.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 17(2), 89–100. doi:10.1111/j.1469–7610.1976.tb00381.x

Anna C. Faul Anna C. Faul Faul, Anna C.

Scales

Scales

1454

1456

Scales

Scales are a group of items, all of which are intended to measure the same construct. They are often developed using classical measurement theory and are typically short, easy to administer, and score. Scales are integral to the process of assessment and evaluation and need to accurately assess constructs of interest in practice and research. Scales provide the framework for evaluating practice and testing research hypotheses.

The accuracy of a scale to measure what it is intended to measure is determined during the scale development phase. During this phase, it is important to follow certain basic rules to ensure that the scale is as reliable and valid as possible. In this process, the goal is to validate the scale in such a way that it is consistently measuring the construct (reliability) and is actually measuring the construct in question and not something else (validity). Reliability is possible without validity, but validity is not possible without reliability. In the following section, the guiding theory for scale development, namely, classical measurement theory, is described, after which the process of scale development is explained in more detail, together with general guidelines on how to assess reliability and validity.

Classical Measurement Theory

Developed during the 1920s, classical measurement theory is currently the most frequently used theory for instrument development and validation. It is based on the true-score model developed by Charles Spearman in 1904 and consists of two theoretical concepts, namely, true scores and error scores. These concepts are theoretical because it is impossible to obtain the absolute true score or the absolute error score. However, it is possible to say that a true score is that which reflects what the person is actually experiencing and that an error score is the gap between actual experience and what is perceived as that experience. Any observed score (*O*) is therefore equal to the true score (*T*) plus the error score (*E*) and can be presented in the form of the following equation:

$$O = T + E.$$

According to classical measurement theory, reliability is based on the amount of error in an observed score for an individual. If the amount of error is quite small, reliability can be claimed. If however, the error is quite large, the scale is unreliable. Part of classical measurement theory is the domain-sampling model. According to this model, any particular scale can be composed of responses to a random sample of items from a hypothetical domain of items. The purpose of any particular scale will be to estimate the scale that would be obtained if one could employ all the items in the domain. The score a subject would obtain if it were possible to test the whole domain is referred to as the true score. A sample of items is reliable to the extent that the score it produces correlates highly with the true score.

Scale Development

A scale must always be developed within a very specific theoretical framework, as the framework guides item development for the scale. Using the theoretical framework, an operational definition of the construct must be developed to guide the scale developer in the design of the specific items that will measure the construct. The domain sampling model of measurement—that there is an infinite pool of possible items that can measure a construct—is then used to develop the items. The skill lies in choosing the specific items that will lead to high content validity, that is, doing a good job of representing the domain that the researcher is trying to measure. The list is typically developed by writing down one attribute of the defined construct and then writing an item based on that attribute; these two steps are repeated until the required number of items has been generated.

Although reliability of a scale increases with length, the law of diminishing return is important to consider when deciding how many items to include: The gain in reliability is smaller when one moves from 11 to 20 items than when one moves from 1 to 10 items and even smaller when one moves from 21 to 30 items.

items.

After the items have been developed, the next step is to assign values to the items to obtain an indication of the level of magnitude of the variable for a specific person. When assigning values, a good rule of thumb is to allow small values to indicate a lower level or magnitude of the variable that is measured and a large value to indicate a higher level or magnitude of the variable. Category partition scaling, or Likert-type scaling, named after psychologist Rensis Likert, is the method often used to assign values to items. This kind of scaling consists of breaking up a continuum into a collection of equal intervals. The number of response categories is an important decision, and multiple studies have identified 7 ± 2 as the optimal choice. Different strategies can be used to name the categories on a category partition scale. One approach is to define only the end points, another is to name all the categories, and a third is to ask respondents to choose between two opposite positions.

Scale Validation

After a scale is developed, it has to be tested for reliability and validity. It is important to obtain enough diversity and variability to permit examination of the reliability and validity of the newly developed measurement tool, so such tests often include additional scales. A representative probability sample is not necessary and can be replaced with a nonprobability convenience sampling technique, which is much less expensive, as long as heterogeneity can be guaranteed. A sample size of 450–550 cases may be enough to satisfy the requirement of the hypothesis tester, power analyst, and parameter fitter.

Investigating Reliability

Reliability concerns dependability or consistency. It addresses the question: To what degree does the measurement of a variable produce consistent results under similar circumstances. Reliability is based on the amount of error in an observed score. If the amount of error is quite small, reliability can be claimed. Reliability estimates range from 0.0 to 1.0. A satisfactory level of reliability depends on how a measure is used. On one hand, for large sample scientific work, a reliability coefficient of 0.60 may be acceptable work. On the other hand, a reliability of .80 may not be nearly high enough in making decisions about individuals and may require reliability of .90 or above.

A simple equation can be used to calculate α coefficient of scale reliability based on the work done by Lee Cronbach:

$$\alpha = \left(\frac{k}{k} - 1 \right) \left(1 - \sum \frac{s^2}{s_0^2} \right),$$

where k = number of items, s^2 = variance of items, and = variance of total scores.

Investigating Validity

Validity suggests truthfulness and means that the construct is measured accurately. It is quite possible for a measurement instrument to be relatively valid for measuring one kind of phenomenon but entirely invalid to measure other phenomena. Thus, one validates not the measuring instrument itself but the measuring instrument in relation to the purpose for which it is being used. Validity is seen as a matter of degree. Two scales can both be valid in terms of the construct that is being measured, but one can be seen as a more valid tool than the other because it does a better job than the other in measuring the construct in question.

In evaluating validity, a scale is judged in relation to one or more well-defined criteria. The validity of a scale can be described by computing a validity coefficient. Such coefficients are obtained as a proportion estimate or as a correlation coefficient and therefore have a theoretical range of values from 0.0 to 1.0. Validity coefficients tend to be smaller than reliability coefficients and normally range between 0.40 and 0.60.

It is important to ensure content validity (adequacy of sampling the items on which people are measured) in terms of a well-formulated plan and procedure of scale construction before the actual scale is developed rather than evaluate this after construction. Construct validity refers to the ability of a measurement tool to measure the specific theoretical construct it was designed to measure. With construct validity, the relation between the scale and its underlying theory is evaluated. Construct validity is related to content validity; however, content validity refers largely to the sampling of the construct domain and the construction of the measurement tool, whereas construct validity refers to the performance of the device with respect to theoretical expectations. Both content and construct validity can be investigated with the use of confirmatory factor and construct validity can be investigated with the use of confirmatory factor analysis, which essentially consists of methods for finding clusters of related variables. Each such cluster or factor consists of a group of variables whose members correlate more highly among themselves than they do with variables outside the cluster. Such correlations can be seen as the factorial composition of measures and play a part in content and construct validity. Factor analysis is important to content validity in suggesting how to revise instruments. It also provides some of the tools necessary to define internal structures and cross structures for sets of variables in construct validity.

Other Scale Development Theories

Other theories besides classical measurement theory have gained popularity in recent years. Specifically item response theory, which provides information on the interplay between samples and measurement error, has been used as an extension of classical measurement theory to develop scales. With classical measurement theory, the level of an attribute is estimated as the sum of responses to individual items, whereas item response theory generally uses the response pattern to evaluate the level of an attribute. When researchers use classical measurement theory only to develop scales, they do not know how the scale performs at different levels of the construct measured. Item response theory provides the methodology to evaluate important additional characteristics of a scale that classical measurement theory does not provide. It provides more detail in describing measurement error, and these descriptions are sample invariant, making wider application of measurement procedures possible and enhancing their use in practice.

Anna C. Faul

See also Likert Scaling; Reliability; Validity

Further Readings

Baker, F. B. (2001). The basics of item response theory. Madison, WI: ERIC Clearinghouse on Assessment and Evaluation.

Faul, A. C., & Van Zyl, M. A. (2004). Constructing and validating a specific multi-item assessment or evaluation tool. In A. R. Roberts & R. Y. Kenneth

(Ed.), Desk reference of evidence-based practice in health care and human services. New York, NY: Oxford University Press.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York, NY: McGraw-Hill.

Brandon LeBeau Brandon LeBeau LeBeau, Brandon

Scatterplots

Scatterplots

1457

1461

Scatterplots

Scatterplots are graphical displays that explore relationships between variables by plotting points at the coordinates of the variables being plotted. The simplest scatterplots are used to explore bivariate relationships between two variables. These variables are traditionally both quantitative; however, this does not need to be the case. Scatterplots plot the (x, y) coordinates of the two variables of interest and every point in the plot represents an individual data point. This entry explores in more detail the creation, uses, and limitations of scatterplots in educational research.

Basic Scatterplot Creation

In the simplest most common case, scatterplots are made by plotting the (x, y) coordinates of two variables in a data set. The example in Figure 1 uses school district data to show the relationship between the percent proficient in Grade 3 on a standardized achievement test and the percentage of students eligible for free or reduced price lunch (FRL). Each point in the figure is plotted at its (x, y) coordinates and represents a unique school district. For example, the point farthest to the right has (x, y) coordinates of approximately (100, 60) indicating that this school district has 100% of their students eligible for FRL and that approximately 60% of their students were proficient in grade three. Similar statements could be made from every point shown in the scatterplot in Figure 1.

Traditional scatterplots are two dimensional; however, three-dimensional scatterplots can be made that plot points using the (x, y, z) coordinates of three variables. Three-dimensional scatterplots can become difficult to view, particularly in print form; therefore, it is much more common to create two-dimensional scatterplots. An alternative to include additional variables, especially qualitative variables, is to change the shape of the points or facet the plot into separate panels. An example of faceting is shown in Figure 2 where two scatterplots are created, one representing small school districts and a second representing large school districts. These types of figures as shown in Figure 2 are helpful to explore if the relationship changes, or is moderated, as a function of a third variable.

Figure 1 Grade 3 percent proficient by the percentage of students eligible for a free or reduced price lunch for school districts

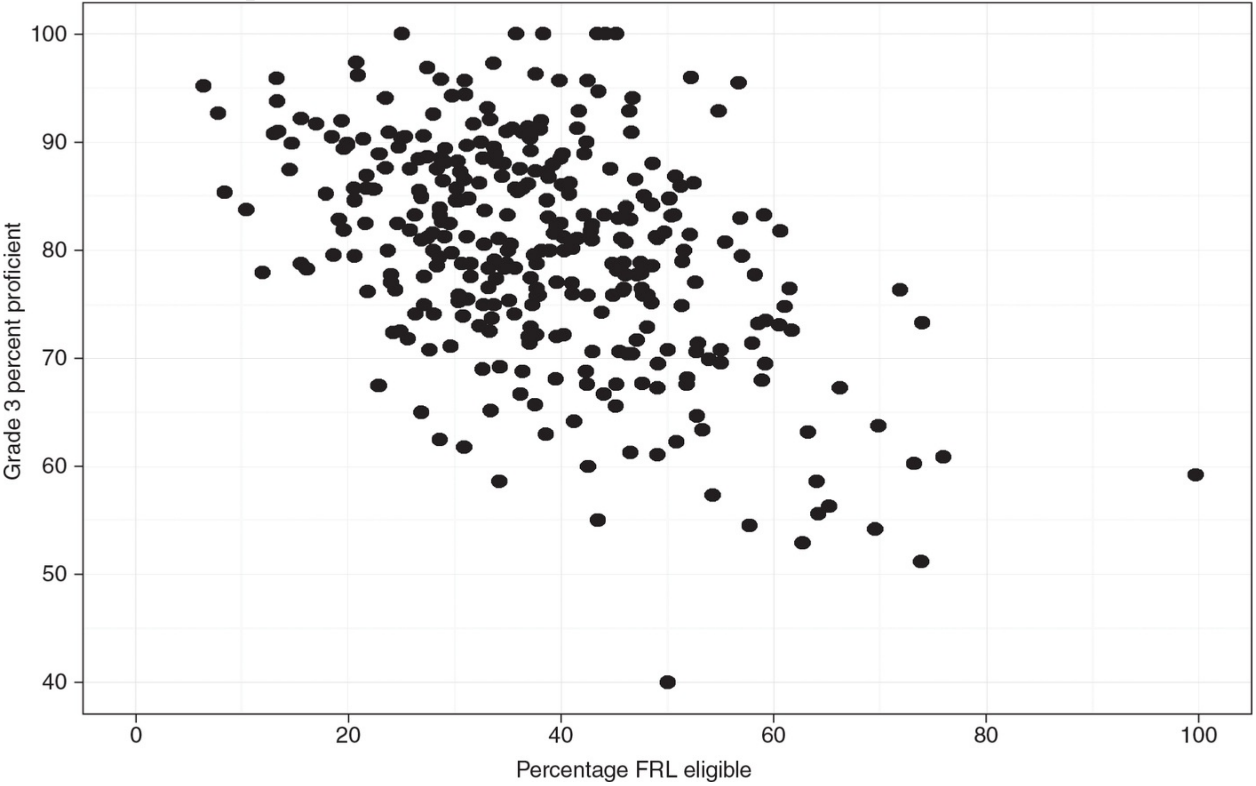
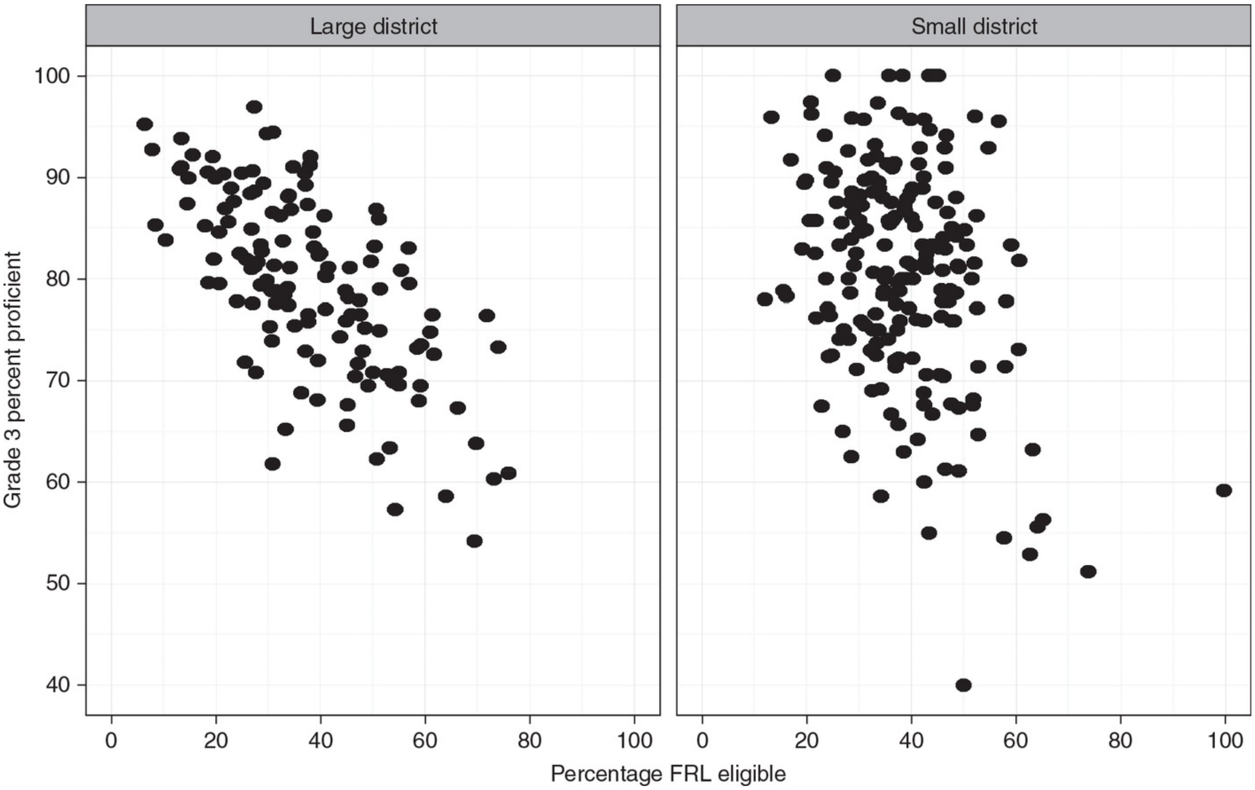


Figure 2 Grade 3 percent proficient by the percentage of students eligible for a free or reduced price lunch and school district size



Data used in the first two figures explore the relationship between two quantitative variables. Data do not need to be quantitative to be plotted in a scatterplot. Instead, data on the *x*-axis could be qualitative, categorical, or ordinal. Figure 3 provides an example of such a plot where the population density of the counties are plotted for various states. Each point in the plot represents a unique county in each state. The primary difficulty in this approach is issues of overplotting. This is discussed in more detail in the limitations section later in this entry.

Uses for Scatterplots

There are many uses for scatterplots in educational research including estimating correlations, exploring the form of bivariate relationships (i.e., linear or nonlinear), creating interaction plots, detecting outliers, and assessing model assumptions. As an example, from Figure 1, one can easily see the negative bivariate relationship between the percentage proficient at Grade 3 and the percentage of students eligible for FRL. Also from Figure 1, the correlation between the two variables could be estimated to be close to −0.5, a moderate to large correlation.

Another use of scatterplots is to assess statistical assumptions for linear models. Figure 4 provides an example of a scatterplot used to assess statistical assumptions for a linear model using the percent proficient at Grade 3 as the dependent variable and the percentage of students eligible for FRL as an independent variable. Figure 4 creates a scatterplot by plotting the predicted values on the *x*-axis and the model residuals on the *y*-axis. As can be seen from the figure, it appears the relationship is roughly linear as there is no strong trend in the residuals across the predicted values. In addition, there do not appear to be large problems with homogeneity of variance as the residuals have a similar range across the predicted values. However, there do appear to be some potential outliers, namely the value with a very small predicted value and the point with a large negative residual. These points could be identified and explored in more detail to determine whether the values are legitimate or some sort of error occurred (e.g., data entry error).

Figure 3 County population density by the state in which the county resides

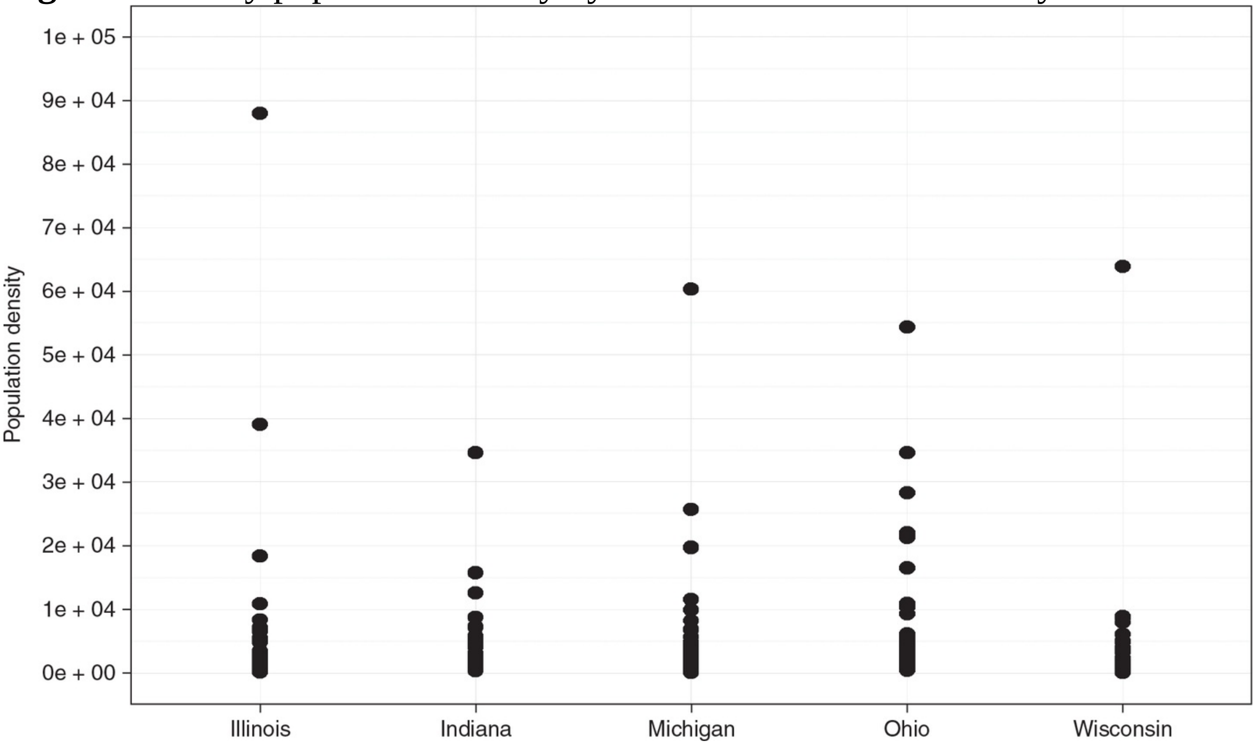
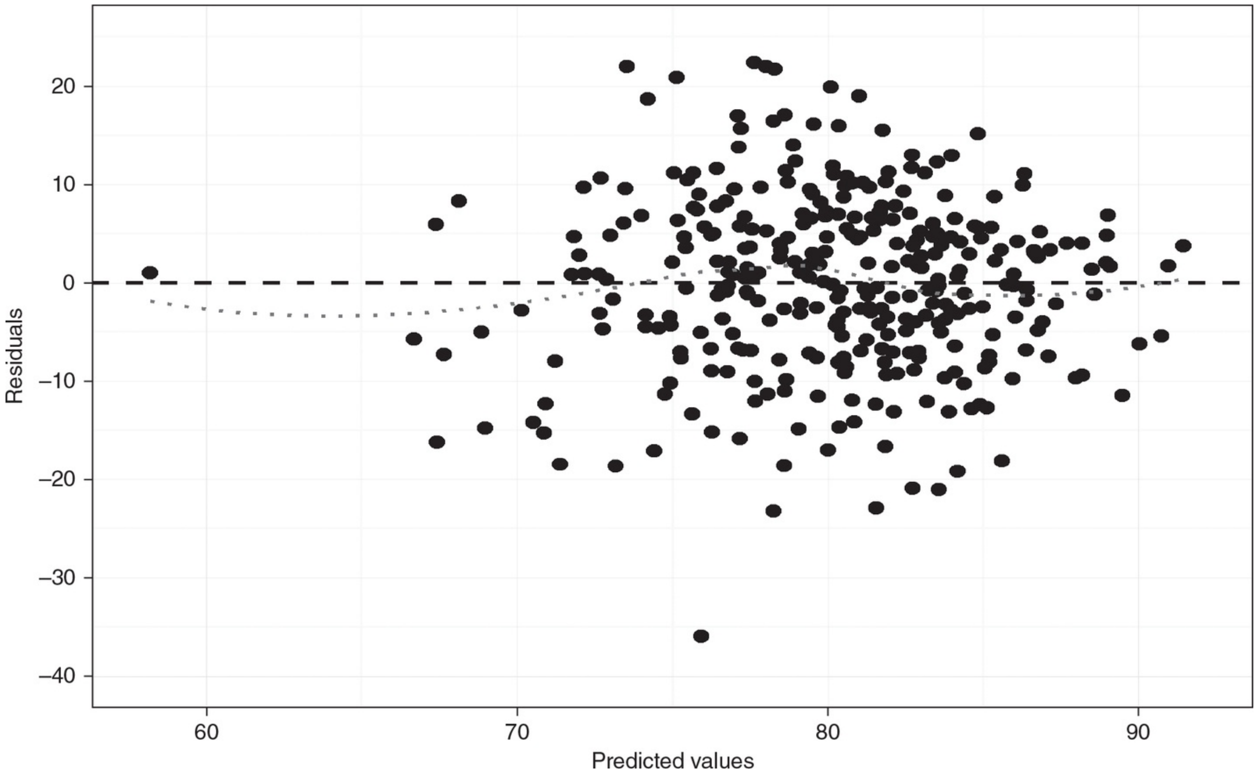


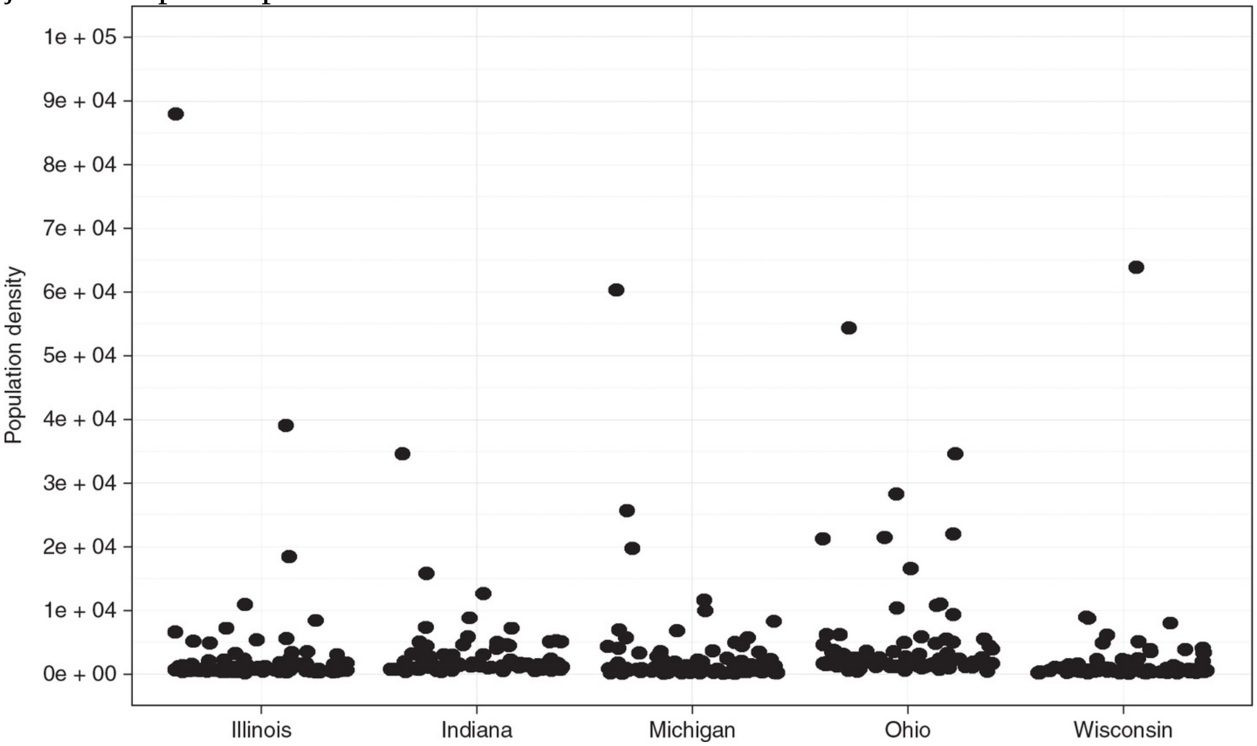
Figure 4 Residuals and predicted values from a linear model used to assess if statistical assumptions have been met



Limitations

Scatterplots can be an extremely useful technique to explore the bivariate relationship between two variables; however, care needs to be taken when creating or interpreting them. The most common issue when creating scatterplots is overplotting. This can occur when there is too much data shown in the graphic or when one variable is discrete (i.e., can only take whole number values). In these situations, it can be difficult or impossible to see all of the data as many points may be represented by a single point. There are many strategies to overcome overplotting, these include adding jitter, randomly sample points to plot, using transparency, or creating interactive scatterplots with the help of HTML and JavaScript libraries. Jittering adds small amounts of random error to spread out points, which is particularly useful when one variable is discrete. Figure 5 recreates Figure 3 while adding jitter to the points. With this plot, it is much easier to see how many data points are present in the data. When plotting many data points, transparency can also improve interpretation where darker portions of the plot indicate areas where many points are overlapping and lighter areas are portions of the data that have relatively fewer points. Finally, randomly selecting a subset of points can be useful to ease overplotting issues as well.

Figure 5 County population density by the state in which the county resides with jitter to separate points on the x-axis



Brandon LeBeau

See also Correlation; Descriptive Statistics; Levels of Measurement; Multiple Linear Regression; Simple Linear Regression

Further Readings

Agresti, A., & Finlay, B. (1997). Statistical methods for the social sciences. Prentice Hall.

Fox, J. (2015). Applied regression analysis and generalized linear models. Thousand Oaks, CA: Sage.

Gerard Michael Jellig Gerard Michael Jellig Jellig, Gerard Michael

School Leadership

School leadership

1461

1462

School Leadership

Leadership can be defined as the set of actions designed to ensure the orderly and effective management and advancement of a system or systems. Leadership is of paramount importance to the success of schools and affects all stakeholders, but most especially students. The role of leadership is to make schools safe and effective ecosystems for learning commensurate with regulatory requirements and community expectations. This entry looks at four specific areas of leadership that are particularly important for school leaders: instructional leadership, organizational leadership, public leadership, and evidence-based leadership.

Instructional Leadership

The primary function of schools through time has been to physically convene students and teachers and transfer a corpus of knowledge, often referred to as the curriculum, from the latter to the former. Today’s schools are more complex, and the transfer is more dialogue based. The outcomes sought include knowledge, skills, behaviors, attitudes, and dispositions. Instructional leadership is the ability to know and manage teaching, learning, and performance.

In today’s schools, teaching and learning are understood as both art and science. The art is in building relationships and establishing a shared culture. Leadership both models how relationships are forged, maintained, and repaired and evaluates the relationships in a building that are critical to student learning outcomes. Classrooms are most conducive to teaching and learning when they are warm, safe, intentionally nonbiased, designed for the learner, and deeply inclusive. Each of these conditions requires great thought and ultimately the endorsement of leadership.

On the scientific side of instructional practice, an effective school leader assumes the dual role of lead learner and expert and is consistently attentive to fields such as neuroscience and neuropsychology and current in pedagogy, curriculum, and assessment. Additionally, the leaders adopt the posture of colearners, indeed the lead learners, recognizing in word and deed the truth that knowledge is not static but rapidly evolving, and they are very much a part of that growth model. A bygone era saw the leader as a dispenser of learning and manager of systems. Today they are consumers of learning and partners in systems leadership.

Organizational Leadership

As organizational leaders, school leaders survey the people, relationships, conditions relative to optimal performance, and change management cycle involved in the school or school system. The organizational leaders recognize that organizations are not static and that they are central to significant organizational change. If the goal is a flatter organization (moving away from a more traditional hierarchy), the school leaders would share research on flatness, model the practices of flatness, conduct regular check-ins with those who might feel loss in flatter structures, articulate progress and challenges to other leaders and the governing board, and allow for concerns and/or fears to surface in a setting where they can respond appropriately.

Organizational leaders need to be able to unite people of divergent backgrounds and belief systems in the service of a shared vision. Organizational leaders manage many relationships within an organization that almost singularly determine outcomes. In the case of the school leader, the stakes are both extraordinarily high and extremely emotional, as the measured outcome is a child’s ability to demonstrate competency in a litany of areas that determines, among other things, placement, promotion, and college admittance. These conditions make it necessary that organizational leaders have developed emotional intelligence or the ability to read their own and others’ emotions and manage processes accordingly.

What school leaders are primarily leading is change, and change can evoke fear in stakeholders as it represents the unknown. Schools are responding to a growing knowledge base about learning and to the digital revolution, and school leaders need to lead this response in an orderly way and with minimal dissonance. To do so, before beginning the change cycle, leaders need the ability to look at the organization with both an external lens (observing what is happening) and an internal lens (anticipating how the organization will navigate change). The organizational leaders must evaluate where each change fits into the greater context of a school's work, where the change is likely to find enemies and allies, how the leaders will support the change and communicate it to others, how they will measure the change, and what will be the unintended consequences. What makes this work so complex is the school leader is the only one assuming this position of inquiry because other stakeholders typically have limited interest or investment in the change.

Public Leadership

Public leadership involves the recognition that no school leader operates in an insular environment. Indeed, school leaders' decisions and the outcomes of these decisions reverberate throughout many publics. Public leadership involves policy and regulatory environments and dealing with the private sector; the municipal, state, and federal governments; and other constituencies. School leaders, when acting as public leaders, must be able to frame issues for all these constituencies in historical and philosophical ways so as to engender support for their work, act as an advocate for their organization, and attain the resources and support necessary for renewal.

The manifestations of public leadership are many. School leaders, for instance, must be connected to the policy-making entities affecting their work. As public leaders, school leaders provide information to local governing boards to inform their decisions and to the state legislators and administrators charged with developing state policies and implementing federal laws such as the Every Student Succeeds Act.

The role of the public leader extends far beyond merely being heard on policy and regulatory matters. Police, fire, and other social services all have a necessary role in the management and leadership of a school. The school leader will determine whether those relationships are transactional or transformational. Will a school leader simply allow a local higher education institution to lease space for its own purposes or begin a wider dialogue about mutual interests and how the university might assist with initiatives to prepare students for higher education and hone the expertise of K–12 teachers? These are the decisions school leaders face when weighing how and where to leverage their leadership.

Evidence-Based Leadership

Evidence-based leadership involves the use of multiple sources of evidence, including published education research and data generated within the school or school system, when making decisions on policies and programs. Evidence-based leadership is a way to determine how best to spend limited resources and to align the organization with one of the primary purposes of education, the vigorous pursuit of truth.

School leaders also model the use of evidence for teachers. Teachers are awash in data that ideally can point the way toward an instructional path, inform them of the efficacy of a unit of study, or indicate where there are holes within a curriculum. Teachers must be competent in working with data in order to conduct sound formative assessment to determine students' readiness for learning, conduct summative assessment to determine the depth and breadth of student understanding, or analyze the results of standards-based assessment given across grades and even schools. An evidence-based leader directs the professional development necessary to build these competencies and ensure that, from the classroom to the main office to the district office, decisions are being made based on the best information available.

Gerard Michael Jellig

See also Emotional Intelligence; Every Student Succeeds Act; Formative Assessment; Stakeholders; Standards-Based Assessment; Summative Assessment

Further Readings

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). School leadership that works: From research to results. ASCD.

Sun, J., & Leithwood, K. (2012). Transformational school leadership effects on student achievement. Leadership and Policy in Schools, 11(4), 418–451.

Stephanie Schmitz Stephanie Schmitz Schmitz, Stephanie

School Psychology

School psychology

1463

1465

School Psychology

The field of school psychology has a primary focus on providing psychological services to children, youth, and families, typically in school settings or in the context of learning. Although school psychology shares interests and training similarities to both clinical and counseling psychology, it is a broader field, encompassing both the psychology and education fields and related theoretical and knowledge bases. School psychologists can work at either the individual or systems levels; and with their background and training, they are able to provide such services as assessment, intervention, prevention, diagnosis, and program development and evaluation.

School psychologists use their knowledge and expertise in learning, behavior, and mental health to promote a child’s success in all areas and across all settings. This entry discusses the history of school psychology and the training and roles of school psychologists.

History of School Psychology

The history of school psychology can be traced back as far as the late 19th-century psychological clinics. Then, in 1905, Alfred Binet and Theodore Simon developed what is considered to be the first test of intelligence. After it was brought to the United States and modified, it was frequently used and for many purposes, one of them being to determine the needs of exceptional, school-age children.

As compulsory schooling was being enforced for all children in the early 1900s, physical exams and psychological inspections, often including intellectual testing, became mandated. Children who failed these psychological and/or medical inspections were often segregated and sometimes placed in separate facilities, receiving services that are often associated with the beginning of special education. The growth of special education services necessitated assistance from various professions, school psychology being one. Therefore, one of the earliest roles of the school psychologist was to use assessment, often in the form of an intellectual test, and to assist in sorting children into categories. It was this role that earned them the title of “gatekeeper,” one that persisted for several decades.

Another important milestone in the history of school psychology was the 1975 passage of the Education for All Handicapped Children Act (PL 94–142), through which an appropriate education for all children, regardless of the presence of a disability, was mandated. School psychologists were once again in demand and their presence in schools increased. The law, which was revised and renamed the Individuals with Disabilities Education Act in 1990, and its subsequent revisions have further supported and shaped the profession, outlining both the services that must be provided to those who have been identified as, or suspected to be, eligible for special education services and the activities that school and related professionals must provide them before, during, and after the evaluation to determine eligibility.

Training of School Psychologists

Due to the number of roles they are expected to play, school psychologists receive specialized training at the graduate level. As of 2013, 240 institutions offered a school psychology program that at least resulted in state certification and/or licensure in the United States. Although some states require only a master’s degree, others require at least a specialist degree (EdS), corresponding to at least 60 credit hours. To receive national certification through the National Association of School Psychologists, school psychologists must have an EdS or a doctoral degree, which corresponds to at least 90 credit hours. Along with coursework, school psychologists must complete a specified number of practicum hours, culminating with a yearlong internship of at least 1,200 hours.

As part of graduate training, the school psychologist’s knowledge base and skills are developed in multiple areas. In 2010, National Association of School Psychologists updated its *Standards for Graduate Preparation of School*

Psychologists, corresponding to the National Association of School Psychologists practice model, each containing 10 domains of school psychology practice. According to these standards, school psychologists should receive training in the following:

Data-based decision making and accountability including a knowledge of various methods of assessment and an ability to use data to make informed decisions about student skills, strengths, needs, and progress;

consultation and collaboration, which encompasses knowledge of various consultation approaches, models, and strategies to assist in collaborating with individuals and groups;

interventions and instructional support to develop academic skills, which includes knowledge of cognitive, learning, and developmental processes, and evidence-based curriculum and instruction;

interventions and mental health services to develop social and life skills, in which school psychologists develop knowledge of various influences on a student's mental health or behavior and how this affects a student's learning and adaptive skills and of evidence-based strategies to develop and maintain positive mental health;

school-wide practices to promote learning, including knowledge of schools as systems, general and special education, and evidence-based practices to foster positive student outcomes in all areas;

preventive and responsive services, which refer to knowledge related to risk and resilience, supports and services to promote prevention, and evidence-based crisis response strategies;

family–school collaboration services, encompassing knowledge of family constellations and how to determine their strengths and areas of need, evidence-based strategies to support families as they support their children's learning and behavioral needs, and to promote collaboration between schools and families;

diversity in development and learning, referring to knowledge of individual differences and diverse characteristics and evidence-based strategies to address and enhance services related to diversity;

research and program evaluation, including knowledge of various research designs, data collection techniques, and program evaluation methods; and

legal, ethical, and professional practice, encompassing knowledge of school psychology history, service models, professional standards, and issues related to effective practice.

Roles of the School Psychologist

Traditionally, the school psychologist’s role primarily consisted of assessment to determine eligibility for special education services. However, as the needs of students, families, and schools have changed, the roles of the school psychologist have expanded.

Roles in which the school psychologist can serve in the social–behavioral realm include psychological counseling, especially as the concerns relate to a child’s performance in school, crisis management, and training to help a child develop and demonstrate appropriate social skills or to manage their anger. Assessment of social–emotional needs and using that data to develop and implement an evidence-based intervention or behavior plan are also possible roles of a school psychologist.

In the academic achievement and learning realm, school psychologists not only conduct assessments but are trained to use these data to assist schools in individualizing instruction, developing evidence-based interventions, and managing classroom behavior. School psychologists are also often involved in assisting the implementation of interventions and collecting treatment integrity data, which entails making sure that the intervention is implemented as intended. Further, they often assist in developing data collection methods to monitor a student’s progress during an intervention and then collecting and analyzing the data when making decisions.

Working with diverse learners, such as those receiving special education services, continues to be an important role for the school psychologist. School psychologists may help teachers and schools to adapt and/or modify their curriculum and instruction; adjust classrooms, routines, and transitions; and develop individualized education programs for those eligible for special develop individualized education programs for those eligible for special education and assist in monitoring their progress. Strengthening family–school partnerships to promote communication and collaboration between home and school is also an important role for the school psychologist.

More recently, school psychologists are involved at the systems level, which allows them to collaborate with district and school personnel. System-level work focuses not only on improving student academic outcomes but also on making schools safer and providing more effective environments in promoting the mental health and social–emotional needs of all students.

School psychologists are involved in and accomplish many of these tasks and activities through their role as a consultant. As a consultant, school psychologists collaborate with teachers, families, and administrators, with the goal of improving the learning and performance of either an individual student or of multiple children at the systems level. Consultation is an indirect service, meaning that the school psychologist (consultant) works directly with the consultee or consultees, who can be the teacher, parent, and/or school administrator, to indirectly benefit the client, that is, the student. Through this indirect service, the consultant is also helping the consultee(s) through the development of knowledge and skills that they may then use in similar situations.

Over the last several decades, school psychologists have had the opportunity to carry out many of these roles within a response to intervention or a Multitiered Systems of Support framework. Response to intervention is often described as a tiered framework that students move through as their level of need in an area(s) intensifies. Corresponding to this increased level of need, instruction also becomes more intensive and individualized. Key components of this framework include a systematic assessment of students’ level and rate of performance and progress within each of the tiers, carefully designed instructional decision-making criteria based upon collected data, and scientifically based core instruction and interventions.

As the need for psychological services for children continues to grow, the demand for school psychologists is expected to increase as well. In 2014, the U.S. Bureau of Labor Statistics reported that the number of school psychologists (as well as clinical and counseling psychologists) was expected to increase by 11% between 2012 and 2022. Although, in 2016, there were reports of a shortage of school psychologists, this trend may change with more awareness of the growing need for school psychologists.

the growing need for school psychologists.

Stephanie Schmitz

See also American Psychological Association; Educational Psychology; Individuals With Disabilities Education Act; Special Education Identification; Special Education Law

Further Readings

American Psychological Association. (n.d.). Postgrad growth area: School psychology. Retrieved from http://www.apa.org/gradpsych/2005/01/schoolpsych.aspx

American Psychological Association. (n.d.). School psychology. Retrieved from http://www.apa.org/ed/graduate/specialize/school.aspx

Careers in Psychology. (n.d.). Employment outlook & career guidance for school psychologists. Retrieved from http://careersinpsychology.org/employment-outlook-guidance-school-psychologists/

Fagan, T. K. (2014). Trends in the history of school psychology in the United States. In P. L. Harrison & A. Smith (Eds.), Best practices in school psychology: Foundations (pp. 383–399). Bethesda, MD: NASP.

Fagan, T. K., & Wise, P. S. (1994). School psychology: Past, present, and future. White Plains, NY: Longman.

National Association of School Psychologists. (2010). Standards for graduate preparation of school psychologists. Bethesda, MD: NASP. Retrieved from https://www.nasponline.org/standards-and-certification

Samuel E. Abrams Samuel E. Abrams Abrams, Samuel E.

School Vouchers

School vouchers

1465

1469

School Vouchers

School vouchers are generally a way for governments to transfer money to parents to allow children to attend private schools, either at no cost to the parents or at a reduced cost. In some cases, the term *voucher* is also been used to describe broader school choice programs that include both private schools and public schools outside of a student's school district. This entry discusses the history of school vouchers, their use in several different countries, and their impact.

The economist Milton Friedman gave the concept of school vouchers its first full exposition. In a 1955 essay entitled *The Role of Government in Education*, Friedman contended that vouchers would amplify educational options for parents dissatisfied with their neighborhood public schools. In particular, Friedman asserted that vouchers would free parents from a governmental monopoly on the delivery of schooling and thus improve variety as well as quality through competition between providers, whether for-profit, nonprofit, or religious.

In Friedman's formulation, the value of vouchers would be the same for all students, regardless of parental income, and redeemable for part or all of tuition at schools satisfying specific minimum standards set by the government. Parents would be free to supplement the value of vouchers with their own money to pay for tuition at more expensive private schools.

The benefits of vouchers, according to Friedman, would include two additional advantages: better pay for teachers, assuming the generation of greater demand for their employment; and passage out of residentially segregated neighborhoods for racial minorities, assuming the wide applicability of vouchers. Regarding the latter matter, Friedman conceded that White proponents of segregation would likewise use vouchers to evade the 1954 Supreme Court decision in *Brown v. Board of Education*, mandating integration of public schools, by sending their children to all-White private schools. But Friedman wrote that while he deplored racism, he considered efforts at persuasion of opponents to integration far preferable to forced integration.

Evolution

At the time of Friedman's essay, school vouchers were already in use, though in more restrictive terms than Friedman articulated. Vermont introduced vouchers in 1869 to allow students in towns without public schools to attend either public or nonreligious private schools in nearby communities within the state or in a neighboring state. The sending town covered the cost of tuition at the recipient school, with the state determining the cost according to a fixed formula. Maine implemented a similar program in 1873. Although only a small percentage of students in Maine and Vermont make use of the vouchers, the system remains essential to many families in remote regions.

The Netherlands introduced its own version of vouchers in 1917. While unified by language, unlike neighboring Belgium, the Netherlands was and would remain divided by religion. The Dutch voucher system addressed this division by allowing parents to send their children to private schools corresponding to their faith or secular pedagogical philosophy rather than to their neighborhood public schools.

The Dutch voucher system was distinctive in four additional regards: Whether public or private, all schools had to comport with national curricular standards established by the Dutch Inspectorate of Education; all private schools were subject to the same teacher salary schedules as public schools; per-pupil funding was the same at all schools, with no allowance for fund-raising activity by parents to supplement individual school budgets (though in the 1980s, a formula was implemented by the government to allocate more money to schools with more underprivileged children); and religious private schools could limit admission to students from families abiding by the school's faith. With approximately 70% of its students at the primary and secondary level using vouchers in 2015 to attend either religious or secular private schools, the

Netherlands exhibited the world’s most robust voucher system.

Although popular with conservatives in the United States, Friedman’s voucher proposal did not translate into policy at home until 1990 and only then in diluted form and on the margin. Yet, Friedman’s proposal did take hold in full force by 1981 in Chile, where many of Friedman’s former students from the University of Chicago assumed policy-making decisions in the government of Augusto Pinochet.

The Chilean version of vouchers, in keeping with Friedman’s recommendation, amounted to a fixed sum granted to parents for use at all government-approved schools, whether for-profit, nonprofit, or religious. If tuition exceeded the value of the voucher, parents had to pay the difference. In the school year before implementation of the reform, 78% of the nation’s schoolchildren attended public schools, 15% attended private schools with government assistance, and 7% attended elite private schools with no such aid. By 1990, 60% attended public schools, 33% attended private schools using vouchers, and 7% continued to attend elite private schools with no such aid. By 2008, the figures were 46%, 47%, and 7%, respectively. In addition, the percentage of students using vouchers to attend for-profit private schools, in particular, had climbed from 18 in 1990 to 31 in 2008.

Despite the increased use of vouchers at for-profit schools in Chile during this period, the presence of these schools generated controversy. Students began protesting en masse in 2006 that for-profit educational management firms diverted desperately needed money from underfunded public schools to investors and proprietors. By 2015, President Michelle Bachelet and the National Congress gave in to the adversaries, approving legislation that would phase out for-profit school management as well as prohibit private schools receiving vouchers from charging more for tuition than the value of the vouchers.

Colombia is another Latin American country that implemented a voucher plan, following in Chile’s path 10 years later. The Colombian program, however, differed in four critical respects: It was meant exclusively for poor secondary students to transfer from overcrowded public schools to underutilized private schools, which amounted to approximately 40% of the nation’s private schools and, as in Chile, excluded elite schools; as demand for vouchers far exceeded supply, lotteries were held to choose recipients; voucher recipients could only continue their studies at private schools if they exhibited academic progress; and for-profit schools were barred from participation in 1996.

for-profit schools were barred from participation in 1996.

For-profit schools nevertheless played a central role in the voucher system introduced in Sweden, which followed in Chile’s path in 1992. In this respect, the Swedish system comported with Friedman’s recommendation. As the private school sector in Sweden was minute when vouchers became policy (with no more than three elite boarding schools, six international day schools, several religious schools, and several schools affiliated with Montessori, Waldorf, and similar pedagogical movements), opening school management to for-profit operators addressed the demand for a greater supply of educational options for parents.

The central impetus for vouchers in Sweden was choice itself. Greater educational opportunity for children in low-income or residentially segregated neighborhoods, a central goal of voucher advocates in the United States, was not an issue, as such disenfranchisement hardly existed in Sweden. With the Social Democrats in control of the government in Sweden for all but 6 years from 1932 to 1991, the conservative Moderate Coalition Party that took over in 1991 was set on changing the way government provided services, from education and health care to transportation and postal delivery, with the expectation that outsourcing to private providers would both increase efficiency and variety. Under the Social Democrats, Sweden functioned as a classic top-down welfare state with steeply graduated income taxes and little private sector involvement in the delivery of public services.

The introduction of vouchers steadily transformed Swedish education. The year before the implementation of the country’s voucher system, only 1% of Sweden’s 1.2 million primary and secondary students attended private schools. In the case of Montessori, Waldorf, and religious schools, the government covered approximately 50% of tuition, and in the case of international schools, approximately 35%. In the 1997–1998 school year, 5 years following implementation, 3% of students attended private schools. By 2010–2011, 15% did so, of which approximately 12% attended schools run by for-profit operators. Admission to schools was determined largely on a first-come, first-served basis, with preference, to facilitate parental convenience, for applicants with older siblings already enrolled.

The critical factor driving the growth of vouchers in Sweden was the governmental decision in 1996 to equate the value of vouchers with per-pupil expenditures in neighboring public schools. At the outset, vouchers were worth

85% of such per-pupil expenditures. When the Social Democrats returned to power in 1994, they decreased the value to 75%. But 2 years later, the Social Democrats, following Dutch precedent, agreed to place the value of vouchers on par with local per-pupil spending with the stipulation that private schools could not charge additional fees or otherwise subsidize individual school budgets with fund-raising activity. This provision brought all Swedish schools into the voucher system, including the international schools and elite boarding schools (with costs for boarding excluded).

With equal funding for all schools and vouchers available to all students, on the one hand, yet a substantial for-profit school management sector, on the other, the Swedish formula accordingly constituted a cross between the Dutch and Chilean systems. In the United States, the high cost of many private schools as well as the considerable opposition to public support of religious education together made implementation of a Dutch or Swedish voucher system impractical.

In contrast to private schools in the Netherlands and Sweden, private schools in the United States often cost far more per pupil than neighboring public schools. Covering the cost of tuition at these private schools with vouchers would thus have been prohibitively expensive. Moreover, the widespread commitment to the separation of church and state in the United States made support of religious schools with publicly funded vouchers, a fraught issue. In addition, many leaders of private schools in the United States placed far greater value on control over admissions than did their Dutch and Swedish counterparts.

Liberals in the United States nevertheless tried to modify Friedman’s conception of vouchers, so that they could be used at a wide variety of private schools. In agreement with Friedman that many urban school systems, in particular, ill-served poor children, the sociologist Christopher Jencks in 1970 proposed a voucher system bearing strong resemblance to the Dutch model: Vouchers would be income adjusted to favor poor children, vouchers would constitute full payment of tuition at participating private schools, and admission to these private schools would be determined by lottery if oversubscribed. Yet, Jencks’s proposal ran into opposition from the left as well as right. In 1979, the same held for a similar proposal of a regulated voucher system made by legal scholars John Coons and Stephen Sugarman.

By 1990, vouchers became a reality in the United States with a pilot program in Milwaukee providing children from low-income homes with modestly valued vouchers to attend secular private schools. In 1995, Cleveland followed suit with a similar program that included religious private schools as well. In 1999, Wisconsin permitted inclusion of religious schools too. As of the 1999–2000 academic year, nearly 8,000 students in Milwaukee employed vouchers at 91 private schools, and approximately 3,500 students in Cleveland used vouchers at 52 private schools. With *Zelman v. Simmons-Harris* in 2002, the Supreme Court upheld the constitutionality of publicly funded vouchers for education at religious schools and thus gave vouchers greater viability.

Although neither the Milwaukee nor Cleveland program fulfilled Friedman's prescription for universal vouchers, they together paved the way to the introduction of similar programs in Florida in 1999, Washington, DC, in 2004, Louisiana in 2008, Indiana in 2011, and North Carolina in 2014. In 2015, Friedman's prescription took hold in Nevada, with the state legislature voting for a universal system conferring US$5,100 in government money per child toward tuition at a private school, whether for-profit, nonprofit, or religious.

Alternate Paths

As vouchers struggled to take flight in the United States, charter schools mushroomed as publicly funded but privately managed schools, with enrollment open to all students and lotteries employed for oversubscribed schools. The movement began with two schools in Minnesota in 1992 and counted nearly 7,000 schools across 41 states and the District of Columbia by 2015. Tuition tax credits, termed "neovouchers" by the education scholar Kevin Welner, have also became popular, with several states allowing parents to take tax deductions on tuition payments at private schools and/or permitting corporations to take tax deductions on donations to private school foundations that could then be repackaged as scholarships for students.

Impact

The effectiveness of vouchers—whether Chilean, Colombian, Dutch, Swedish, or American—has nevertheless remained a subject of contentious debate. For a comprehensive understanding of their effectiveness, the economist Henry M. Levin recommended vouchers be judged according to their impact on choice, efficiency, equity, and social cohesion.

Choice may be understood in the context of both variety of educational options and their accessibility, the latter of which leads to the question of equity, as choice means little if vouchers are not sufficiently funded to cover the cost of tuition at a significant range of private schools. In Chile, Colombia, and the United States, for example, vouchers by these criteria paled in comparison to vouchers in the Netherlands and Sweden. Regarding efficiency, rigorous evaluations have yet to reveal compelling evidence that voucher recipients make more academic progress than their public school peers. Regarding social cohesion, Friedman’s concession in 1955 that vouchers might well lead to more segregation until parents are convinced of the merits of integration has taken on significant meaning. Even in the Netherlands and Sweden, there has been considerable documentation of vouchers leading to segregation.

Samuel E. Abrams

See also Accountability; *Brown v. Board of Education*; Selection Bias

Further Readings

Abrams, S. E. (2016). Education and the commercial mindset. Cambridge, MA: Harvard University Press.

Coons, J., & Sugarman, S. (1978). Education by choice: The case for family control. Berkeley: University of California Press.

Elacqua, G. (2012). The impact of school choice and public policy on segregation: Evidence from Chile. International Journal of Educational Development, 32, 444–453.

Friedman, M. (1955). The role of government in education. In R. Solo (Ed.), Economics and the public interest (pp. 127–134). New Brunswick, NJ: Rutgers University Press.

Ladd, H. F., Fiske, E. B., & Ruijs, N. (2009). Parental choice in the Netherlands: Growing concerns about segregation (Working Paper 182). New York, NY: National Center for the Study of Privatization in Education, Teachers College, Columbia University. Retrieved from http://ncspe.tc.columbia.edu/working-papers/OP-182.pdf

Levin, H. M. (1998). Educational vouchers: Effectiveness, choice, and costs. Journal of Policy Analysis and Management, 17, 373–392.

Levin, H. M. (Ed.). (2001). Privatizing education: Can the marketplace deliver choice, efficiency, equity, and social cohesion? Boulder, CO: Westview Press.

Moe, T. M. (2001). Schools, vouchers, and the American public. Washington, DC: Brookings Institution Press.

Rouse, C. E., & Barrow, L. (2009). School vouchers and student achievement: Recent evidence and remaining questions, Annual Review of Economics, 1, 17–42.

Welner, K. G. (2008). NeoVouchers: Providing public funds for private schools through tuition tax credits. Lanham, MD: Rowman & Littlefield.

Kathleen Lynne Lane Kathleen Lynne Lane Lane, Kathleen Lynne

Emily Cantwell Emily Cantwell Cantwell, Emily

School-Wide Positive Behavior Support Schoolwide positive behavior support

1469

1472

School-Wide Positive Behavior Support

Positive Behavior Interventions and Supports (PBIS) includes a cascade of supports ranging from primary prevention (Tier 1) efforts for all, secondary prevention (Tier 2) efforts for some, and tertiary prevention (Tier 3) efforts for a few. These supports increase in intensity as students move through the continuum of tiered supports. For example, Tier 1 efforts are intended to level the playing field for all students, and the expectation is about 80% of the student body will respond to this school-wide approach to PBIS. Tier 2 supports are for students with common acquisition (can’t do problems), fluency (trouble doing problems), or performance (won’t do problems) deficits. These low-intensity supports can be delivered in small groups (e.g., social skills groups) or individually (e.g., check/in check/out). Tier 3 supports are reserved for students with the most intensive intervention needs. These supports are individualized and often involve families and other related service providers given the highly intensive nature of these interventions (e.g., functional assessment-based interventions). In this tiered system of supports, data are used to determine which students may require assistance beyond Tier 1 efforts and connect them to more intensive research-based supports according to individual students’ needs. Care is taken to ensure each level of support is implemented with integrity as it would not be prudent to suggest a student is not responding to Tier 1 or Tier 2 supports if one is not confident the supports are actually in place as planned (with treatment integrity). School-Wide Positive Behavior Support (SWPBS) refers to a systems change process for a school or district in which an instructional approach to behavior is adopted at Tier 1. SWPBS is a framework intended for all grade levels from preschool through high school. This entry reviews the SWPBS process, from establishing a team to implementing proactive and reactive instructional approaches, and discusses its effectiveness.

SWPBS Team

When a district or school moves forward in establishing SWPBS, the process typically begins with establishing a team that includes an administrator with decision-making authority, general education teachers, special education teachers, and related service individuals (e.g., school counselor, psychologist). Representation from the parent and student communities is encouraged by many researchers. The team identifies three to five positively stated expectations to guide behavioral performance for the entire school. Rather than listing “don’ts,” the goal is to establish expectations to guide desired behavior across all settings. For example, the expectations may include the following: be respectful, be responsible, and give best effort. Then, each expectation is defined for all key areas in a building: classrooms, hallways, cafeteria, buses, and arrival and dismissal.

Some teams use a tool, Student Expectations for Success in School Settings, to guide the development of this expectation matrix with input from all faculty and staff. The Student Expectations for Success in School Settings is completed by all adults working in a given building to determine which behaviors are viewed as critical for success in each area (e.g., listen to instructions, follow instructions the first time). The SWPBS leadership team compiles the responses of the Student Expectations for Success in School Settings and uses it to guide the build of the matrix. This tool is available from the Comprehensive, Integrated Three-Tiered Model (Ci3t) of prevention website.

After expectations are established, the goal is to secure buy-in from all adults, ensuring the majority of faculty and staff view student behavior to be one of the top three priorities for a building. The U.S. Department of Education’s Office of Special Education Program PBIS Technical Assistance Center recommends 80% agreement to the established expectations, as consistency in expectations among adults is critical to successful implementation of SWPBS.

Proactive Approach: Teaching, Practicing, and Reinforcing

After defining expectations for all key settings and securing agreements among adults, expectations are taught to all students and a reinforcement system is established to support all students in learning and meeting expectations. Many schools develop lesson plans to teach expectations for each setting defined in the matrix to support adults in their teaching responsibilities and ensure fidelity of teaching procedures. Oftentimes these setting lessons are taught by all teachers during the first week of the school year and then revisited weekly or monthly throughout the school year. Teaching all students expectations enables teachers to level the playing field for all students by being proactive. Rather than waiting for students to make errors and then giving feedback as to what they did wrong (reactive approach), in SWPBS, an emphasis is placed on teaching expectations, giving students opportunities to practice, and receive reinforcement for meeting expectations. In short, teaching expectations is similar to how one would teach academic content.

As part of the reinforcement structure, the team develops a universal reinforcer (sometimes referred to as a “gotcha”) such as a PBIS ticket that adults give to students paired with behavior-specific praise to acknowledge students who are meeting expectations. It is important that the faculty and staff reinforce malleable factors, meaning things students can change such as effort rather than ability. For example, a teacher might give a student a PBIS ticket and say “Thank you for showing responsibility by getting your homework in on time today” rather than saying “You are so smart! You are the smartest student in the whole fifth grade.” From a behavioral perspective, this contingent introduction of the PBIS ticket and behavior-specific praise increases the likelihood of the desired behaviors occurring in the future. With behavior-specific praise, the student receives positive feedback on the exact expectation demonstrated, enabling the student to know what behaviors are desired and will help the student be successful. By having one universal reinforcer rather than separate systems in each classroom, students are able to receive feedback from a wide range of adults in the school setting: teachers, paraprofessionals, custodians, cafeteria staff, office staff, and administrators. SWPBS does not subscribe to punishment-based procedures. For example, teachers do not take PBIS tickets away when students make mistakes as students’ current mistakes do not “undo” the previous successes they experienced when they met expectations. Collectively, this program for generalizations of the knowledge and skills acquired from the setting lessons creates a positive, productive, safe, and even joyful school climate. In addition, students acquire a base of learned behaviors to help them be successful beyond the school building.

Reactive Approach: Responding to Challenges

In addition to this proactive approach to teaching, practicing, and reinforcing expectations, SWPBS also includes a clearly defined reactive plan for responding to challenges that do arise. The team develops a list of challenging behaviors (e.g., noncompliance, verbal aggression, and physical aggression) that range from minor to major offenses. With input from faculty and staff, the team defines each behavior to support consistency in understanding among administrators, faculty, staff, parents, and students. The team also develops an office discipline referral (ODR) form to document these infractions (e.g., time, date, persons involved, location, and possible reason for the challenge) as well as a plan (often flow chart) illustrating how to respond to minor and major infractions. For example, there are certain behaviors that can be managed in the classroom (often for the first three occurrences) and other behaviors that result in immediate removal from the classroom, sending the student to the office. For example, not completing a homework assignment would likely be managed by the teacher in the classroom, whereas a student who hits another student or the teacher would likely be managed in the office by the principal or vice principal. Many schools elect to use the School-Wide Information System which is a web-based program developed for just this purpose. School-Wide Information System enables teams to quickly make graphs for behavioral challenges occurring per day, per month, time of day, in a given location, and even by individual students. This efficient system supports teams in efficiently and effectively analyzing data.

In addition to ODR data, teams can also select a systematic screening tool for behavior that can be completed by teachers three times per year: fall (4–6 weeks after the school year begins), winter (prior to winter break), and spring (4–6 weeks prior to the end of the school year). School teams work closely with district leaders to select a validated systematic screening tool, with attention to ensuring the tool is reliable, valid, and feasible for use with their student body. There are a range of screening tools, some of which are free access and others that are commercially available. Teachers independently rate each student on their class roster according to the guidelines provided in the screening tool selected. In general, elementary teachers rate their homeroom students, whereas middle and high school teachers rate students in one period (e.g., all second period students are rated).

Screening data are highly predictive of important outcomes for students such as

ODRs earned in a year, number of days suspended, courses failed, and grade point average. Screening data can be used in conjunction with other data collected as part of regular school practices (e.g., academic screening data, attendance, and ODRs) to examine the overall level of risk in a building. For example, graphs showing the percentage of students placing in low-, moderate-, and high-risk categories can be made and examined. If the percentage of students in the low-risk category is below 80%, it would be advisable to focus on refining Tier 1 efforts. These data can also be used to inform teacher-delivered supports. For example, if teachers notice more than 20% of students in their homeroom class are placing in the moderate-or high-risk categories, low-intensity supports such as incorporating instructional choice or increasing student's opportunities to respond are an effective starting point. In addition, screening data can also be used to detect students who might need more than Tier 1 supports have to offer. These students can be connected to Tier 2 (e.g., check/in check/out) and Tier 3 (e.g., functional assessment–based interventions) when Tier 1 efforts are insufficient. At each level, the intent is to provide students with evidence-based strategies, practices, and programs with enough evidence to suggest that if implemented with integrity (as planned), they will yield desired outcomes for students.

Effectiveness

SWPBS focuses on creating positive, productive, and safe environments for all students by subscribing to an instructional approach to behavior that includes proactive and reactive components. In brief, SWPBS focuses on a systemic approach to behavior, grounded in respectful interactions between adults and students. This structure also facilitates resource-efficient structures to support collaboration between general and special education communities, working toward the shared goal of preventing learning and behavior problems from occurring and responding efficiently when challenges do arise. Randomized control trials of SWPBS at the elementary level suggest this systems' change approach is highly effective in not only reducing challenging behaviors but also improving school climate and academic outcomes. Commitment from administrators is a key factor in predicting successful implementation, and ongoing professional learning is also critical.

Kathleen Lynne Lane and Emily Cantwell

See also Data-Driven Decision Making; Response to Intervention

Further Readings

Horner, R. H., & Sugai, G. (2015). Schoolwide PBIS: An example of applied behavior analysis implemented at a scale of social importance. Behavior Analysis in Practice, 8(1), 80–85. doi:10.1007/s40617–015-0045-4

Lane, K. L., Menzies, H. M., Ennis, R. P., & Oakes, W. P. (2015). Supporting behavior for school success: A step-by-step guide to key strategies. New York, NY: Guilford.

Lane, K. L., Menzies, H. M, Oakes, W. P., & Kalberg, J. R. (2012). Systematic screenings of behavior to support instruction: From preschool to high school. New York, NY: Guilford.

Sugai, G., Lewis-Palmer, T., Todd, A., & Horner, R. H. (2005). Schoolwide evaluation tool: Version 2.1. Eugene, OR: University of Oregon, Educational and Community Supports.

Websites

Positive Behavioral and Interventions & Supports: www.pbis.org

Comprehensive Integrated Three-Tiered Model of Prevention: www.ci3t.org

J. E. R. Staddon J. E. R. Staddon Staddon, J. E. R.

Scientific Method

Scientific method

1472

1477

Scientific Method

Science is the modern name for what used to be called *natural philosophy*. Science comprises just those ideas and concepts that can be tested by third parties. The hypothesis that objects of different weights all fall at the same speed is scientific; the idea that there is an afterlife, inaccessible to the living, is not. The many ways in which we can arrive at and evaluate scientific ideas are collectively termed the *scientific method*.

Scientific ideas can come from thoughtful observation or via experiment. Experiments may be designed to test a theory (*hypothetico-deductive*), or they may be simply exploratory "what if?" attempts to satisfy natural curiosity. Hypothetico-deductive experiment involves answering questions of the form "If I do X will I get Y?" Nonexperimental, *inductive*, science infers some general rule from a set of observations: All the swans I know are white, ergo, swans are all white. In practice, these divisions can be arbitrary. Scientific method is not an algorithm; it is not a recipe or a decision tree. There is no "gold standard" that can guarantee scientific advance. This entry further discusses the inductive, deductive, and experimental methods and provides examples illustrating each.

Inductive and Deductive Methods

Inductive reasoning uses specific instances to infer general principles, whereas deductive reasoning derives specific conclusions from one or more premises or axioms. Inductive reasoning takes many forms, but a popular inductive approach comes up with a generalization based on a set of observations. An historical example of this method is the investigation by physician John Snow of the 1854 outbreak of cholera in London, which killed hundreds in a few weeks. Many explanations were offered for this outbreak. No one really understood how diseases spread, as the germ theory of disease had yet to be proposed (that happened after 1860 with Louis Pasteur's study of puerperal fever). The prevailing view was something called the miasma theory, which held that "noxious exhalations" from swamps and like places somehow cause disease.

In those days, there was no domestic water supply. People got their water from hand pumps scattered across the city. The pumps had different sources—local wells or piped from the river Thames or one of its tributaries. Snow did not believe in the miasma theory and sought another explanation for the spread of the disease. He looked at where cases of cholera had occurred. He noticed that almost all of them were clustered within walking distance of a particular hand pump in Broad Street. This allowed him to come up with a *hypothesis*, that the water from the Broad Street pump is contaminated in some way that causes cholera. The hypothesis suggested an obvious experimental test: Remove the handle from the Broad Street pump so that no water can be obtained from it. Snow managed to persuade the local council to disable the pump. His hypothesis was confirmed: The incidence of cholera dropped dramatically, proving (without the aid of statistics) that the pump was the source of the disease.

Snow's experiment is what is called an *AB design*: Two conditions/treatments are applied in succession—handle/no-handle. In a laboratory context, both conditions would normally be repeated, ABAB, just to be sure that B really has the predicted effect. In Snow's case, this was both unnecessary (the effect of removing the handle was large) and unethical (restoring the handle might have caused more deaths)—and in any case, his main purpose was to improve public health rather than advance knowledge.

Snow's discovery is considered to be the beginning of the science of epidemiology. The Broad Street example also illustrates a great and oft-forgotten truth: Epidemiology is a rich source of *hypotheses* but it cannot prove *causation*. Snow found a *correlation* between the incidence of disease and distance from the Broad Street pump. A correlation can come about for many reasons. Snow blamed the pump, which suggested his experimental test, which identified the actual cause. Other interpretations were possible, however, including movement of population away from the area and spontaneous decline of the epidemic (all epidemics eventually cease).

The cautionary message should be clear: *Correlation is not causation*. The inductive method must be combined with the deductive. Induction may yield a hypothesis, but the hypothesis must be tested by experiment to establish its truth. Induction is often wrong—all swans are not in fact white.

Snow’s investigation, like most of science, involved both induction and deduction. Induction, the correlation between disease and distance from the pump, led him to deduction: Disabling the pump should halt the epidemic. But there are many other kinds of induction. The paradigmatic case is Charles Darwin and the theory of evolution by natural selection.

As a young man of age 22 with little formal education, in 1831, Darwin embarked on HMS *Beagle*, a small navy ship charged with hydrographic mapping of South America and whatever other territories its almost equally young Captain Robert FitzRoy could manage. The voyage lasted 5 years and took the little ship all round the world. At every landfall, Darwin went ashore and observed the geology and natural history of the place, collecting plants, animals, and rocks everywhere he went and sending his specimens back to England whenever the ship reached port. His collection methods were opportunistic. He basically tried to get hold of any new thing that struck his fancy. He had no grand theory to guide his collection. His haphazard method turned out to be an asset rather than a liability because it gave him a relatively unbiased sample of the biology and geology of the areas he visited.

He learned many things from his travels. That geography is not fixed: The *Beagle* arrived at the port of Concepción in Chile during an earthquake which had raised the land by several feet in some places. Further proof of topographic change was finding a layer of seashells along mountains hundreds of feet above the sea. He noticed that organisms long isolated from the mainland in the Galapagos Islands seemed to have diverged from the colonizing species and that living species seemed to have similar extinct ancestors. From all of Darwin’s varied observations of geology, zoology, and botany, he inferred first (although this was not original with Darwin) that species evolve, and second (his great contribution) that the process by which they do so is natural selection.

Darwin’s work is perhaps the most famous example of inductive science. But in this case, proof came not so much from experiment as from the ability of Darwin’s theory to make sense of a vast mass of facts or what would now be referred to as *empirical data*.

Experimental Method

In the 21st century, science has become institutionalized. The number of scientists, especially social scientists, has much increased. The pressure to produce results has increased even more as the numbers of scientists have grown faster than research support. All have favored a drift toward what might be called the algorithmic approach to scientific experiment. In biomedicine, for example, the randomized-control-group experiment is often called the “gold standard” of scientific method. But the real advances in science have all followed a less orderly path. There is no single, well-defined method that *guarantees* an advance in understanding.

The experimental method, properly defined, is not any individual experimental procedure or even a list of such procedures. It is not a checklist. It is a *sequence of experiments* that end in some definite, readily testable conclusion about nature. It is the sequence and the conclusion that constitutes the experimental method, not the details of any particular experiment. There is no magic-bullet gold standard that can reveal a great truth in one shot.

When the phenomenon to be studied can easily be repeated, the single-subject ABAB design can be used. This design involves studying one subject, such as an individual person, by first observing the subject, then applying a treatment, then withdrawing the treatment, and finally repeating the treatment. But in studying certain areas, such as the process of learning, the ABAB design has problems. Unlike sensation, learning is not reversible; once something has been learned, it cannot easily be unlearned. So the learning experience cannot be repeated, ABAB fashion, because the response to the second B is likely to be different than to the first. Therefore, most early learning studies, and a majority of experimental studies even in contemporary social science, use the *between-group method*. Subjects, animals, people, and agricultural plots, are randomly assigned to two or more equal groups. The assignment is random, so that the groups shall not differ in any systematic way. One group, the *control* group, is untreated or gets a treatment known to be ineffective. The other, *experimental*, group gets the treatment to be tested, a new drug or training procedure, for example. If the groups differ in some *dependent variable*, such as cure rate or learning rate, the difference is tested statistically. The method has two problems: There is ambiguity going from group results to claims about individuals, and some popular statistical tests have turned out to be flawed.

But in 1938 B. F. Skinner (1904–1990) proposed a different approach to studying the learning process. He was interested in how reward and (to a much lesser extent) punishment might be used to change behavior in socially beneficial ways. As was the custom at the time (the 1930s), he began by studying animals, which was simpler and less ethically problematic than studying human beings. Most learning studies at that time compared groups of animals, giving one group (the *control group*) a “standard” treatment, and giving the treatment to be assessed, such as a different trial spacing, to the other, *experimental group*. Animals were randomly assigned to each group so that the groups, if not the animals, could be considered basically identical.

Skinner’s experimental method allowed for a simpler approach. To study the effects of intermittent reward, he simply trained his hungry animals (usually pigeons) to peck a lighted disk. At first, each peck operated a feeder (via an automatic control circuit) and gave the animals a few seconds access to food. But the birds would continue to peck even if (say) only every 10th peck operated the feeder (called a fixed-ratio schedule of reinforcement), or if only the first peck 60 seconds after the previous reinforcement operated the feeder (a fixed-interval schedule), and so on.

His great discovery was that after sufficient exposure, each of these procedures yields a distinctive pattern of behavior, as revealed by a cumulative record. Moreover, the pattern for each schedule is usually *stable* in the sense that it can be recovered after exposure to a different schedule. This property of *reversibility* meant that the effects of various procedures could be studied in individual animals, with no need for inferential statistics.

The within-subject method (comparing treatment effects successively applied to the same individual) allowed researchers to use the simple ABAB design to investigate the effects of various reinforcement schedules. Many new phenomena were discovered, including the sensory thresholds of animals, the effect of temporal patterns of reward on schedule behavior, and the persistent effects of shock-avoidance procedures. But the contribution of his method that Skinner and his followers thought most important was *control*, control for purposes of education, social melioration, and therapy. They were much less interested in using it as a tool to understand the learning process itself.

The single-subject method, also referred to as the single-case method, is limited because although the behavior of an animal on its second exposure to, say, a fixed-interval schedule looks identical to its behavior on first exposure, the animal is not the same. Or, to put it more technically, the behavior observed may be the same, but the animal subject is not in the same *state* as before. Because it is not in the same state when it first learns, say a fixed-interval schedule, it may respond differently to some other procedure after so learning than it would have if the procedure had been applied before any training at all.

If our animal could be “reset” at the end of the experiment, then the effect of Procedure C would be the same in an experiment that gave the animal the two treatments AC as in one that gave him ABC. We can’t make such a comparison because we can’t “reset” real organisms. Numerous transfer experiments (i.e., Condition A followed by some other Condition B) show that every learning experience has some lasting effect. Animals’ response to C after AB will often be different than their response to C after A. The single-subject method can be used to test some aspects of nonreversible behavior, that is, learning—but to do so requires hypotheses about the effects of transitions from one schedule to another; in other words, it requires theory, something that Skinner strongly discouraged.

The single-subject method cannot answer many questions about learning, such as whether learning occurs faster when practice is spaced over time with rest periods between sessions (spaced practice) compared to when there is no rest between sessions (massed practice). Questions like this usually require comparisons between groups of subjects. Because the performance varies from one individual subject to another, group results often overlap, even when there is a real difference between the two. Assessing the results of such experiments means coming up with a *probability* of some sort–some measure that can tell us how likely it is that the result obtained could have come about by chance.

The first step is to understand how probabilities are measured. Probability theory is mathematically quite simple. But it is also one of the most conceptually difficult parts of applied mathematics. In his 1935 book *Design of Experiments*, R.A. Fisher begins with a deceptively simple example to show how probability can be computed that is now known as the “lady tasting tea” experiment. The experiment involved a lady tasting eight cups of milky tea. For four of the cups, the milk was poured in first, for the other four, the reverse. The lady’s task was to identify which was which.

Assuming that the lady gets more cups right than wrong, how can we judge whether she can *really* tell the difference? Between 1935 and 1971, Fisher wrote that “In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment and to have decided without ambiguity what interpretation shall be placed upon each one of them” (p. 12). In other words, the starting point for *all* inferential statistics is to define precisely all possible outcomes of a given experiment, their probabilities and their meaning in relation to the question being asked. Only if this *sample space* is fully defined can we interpret the results of an experiment correctly. The task is often difficult and sometimes impossible.

What is the sample space for the tea-lady experiment? Well she knows that the probability a given cup is tea-first (call it T) is exactly one half (four out of eight). If the successive choices are independent, then her chance of getting the first one correct is ½, getting the first and second correct is just ½ × ½ = ¼ and so on, so that her probability of getting all eight correct is one over 2^8 = 1/256.

But Fisher’s estimate is just one in 70, so there is something wrong with that simple analysis. The answer of course is that the lady’s choices are *not* independent because she knows that exactly four cups are tea-first. She can ignore all outcomes where tea-first is either more or less than four. The 1/256 answer would of course be correct if the experimenter decided on how to mix the tea by tossing a coin each time. But that procedure would not guarantee exactly four T and four M in the eight cups. In other words, the sample space for that experiment is considerably larger than the one Fisher discusses. Hence, the probability of getting all correct is much smaller.

The result is indeed 70. Because the lady knows there are exactly four of each type, the number of possibilities is less than if all she knew was that the probability of each cup being T is one half. Fisher goes on to ask just how many choices the lady must get correct if we are to believe her claim. This is the thorny issue of *significance level*. Just how improbable must the experimental result be for us to conclude that our hypothesis—the lady has the ability she claims—is true?

But that is the wrong question. No experimental result can tell us whether our hypothesis is *true* or not. What it can tell us is whether our result is likely to be *replicable*: Between 1935 and 1971, Fisher wrote that “a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.” (p. 14). The likelihood that only chance is involved, the criterion significance level we choose, is in fact completely arbitrary. In physics, when statistics are occasionally used in testing a theory, the published levels tend to be exceedingly small. A test of gravitational-wave theory in 2016, for example, reported results significant at a probability of one in 3.5 million—a lottery-level probability that the result could have occurred by chance.

In social science, results that are much more likely than this to arise by chance are routinely accepted as significant. The usual criterion for statistical significance in social science and pharmacology is that the result could have occurred by accident with a probability of just 5% or one in 20. The flaws of this level have recently become apparent and the method of *null hypothesis statistical test* (NHST) is currently undergoing something of a reform if not a revolution.

Despite these uncertainties, NHST is still the method of choice for most social science research. The standard method involves two groups, matched as far as possible on every relevant characteristic (equivalently, subjects are randomly assigned to the two groups). The size of the groups varies from study to study, from as small as 10 to several hundred in large experiments. One of the two groups is randomly chosen as the control group, which does not receive the experimental manipulation, whatever it may be. The other is the experimental group which receives the treatment to be tested. The hypothesis being tested, the *null hypothesis*, is that the two groups are the same—that any measured difference between them could have occurred by chance.

Even if the probabilities yielded by standard statistics are accurate, there is still a serious problem with the NHST method. In a 2005 article entitled “Why Most Published Research Findings Are False,” researcher John Ioannidis concluded that it is more likely for biomedical research claims to be false than true. The failures of the NHST method have shown up most strikingly in drug studies, efforts to find medications to cure disease. In 2011, an analysis of internal efforts by the drug company Bayer to validate new drug target claims indicated that Bayer halted nearly two thirds of these projects because in-house experiments failed to match claims made in the literature.

In retrospect, these failures could have been predicted. Drug companies and academic research labs test thousands of potentially curative chemical compounds. Suppose 1,000 *ineffective* drugs are tested. By chance, the results of about 50 studies, 5%, will surpass the 5% significance level. If some of these tests were done in academic labs, they will be published. Because the drugs in all these hypothetical studies are in fact ineffective, attempts to replicate their effects are likely to fail. Negative results, tests that show ineffective drugs to be ineffective, are unlikely to be published, although efforts to reverse that bias are being made.

The authors of any individual study will not know the whole sample space. They will not know how many drugs have been tried nor how many tests have failed. They will not, therefore, be in a position to assess the likelihood that their own statistically significant one-shot attempt is unreplicable. The NHST method is unlikely to be abandoned by social science, but skepticism toward studies using NHST is warranted until the problems with the method have been satisfactorily resolved.

J. E. R. Staddon

See also Effect Size; Experimental Designs; Positivism; Postpositivism; Qualitative Research Methods; Quantitative Research Methods; Single-Case Research

Further Readings

Fisher, R. A. (1971). The design of experiments. Edinburgh, UK: Oliver & Boyd. (Original work published 1935) Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? The Journal of Experimental Education, 71(1), 83–92.

Ioannidis, J. A. (2005). Why most published research findings are false. PLoS Med, 2(8), e124.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251). doi:10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.J. (2012). Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence?

Perspectives on Psychological Science, 7(6) 528–530.

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspectives on Psychological Science, 11(4), 539–544. doi:10.1177/1745691616646366

Sidman, M. (1960). Tactics of scientific research: Evaluating experimental data in psychology. New York, NY: Basic Books.

Skinner, B. F. (1956). A case history in scientific method. American Psychologist, 11(5), 221–233.

Staddon, J. (2014). Unlucky strike: Private health and the science, law and politics of smoking. Buckingham, UK: University of Buckingham Press.

Neil J. Dorans Neil J. Dorans Dorans, Neil J.

Score Linking

Score linking

1477

1479

Score Linking

Scores are the end products of assessment processes. Scores are used for admissions, placement, diagnosis, and other purposes. Scores from different assessments are often compared. The urge to make comparisons can lead someone who would never consider using height as a measure of weight to fall prey to the temptations of presuming that scores from any one educational assessment can be used as substitutes for scores from another assessment, forgetting that assessments are designed for different purposes. This entry discusses the reasons that score linking is performed, how it is performed, and considerations when linking scores from different assessments.

Even when the purposes of two different assessments are similar, linking scores from these assessments can be a challenge. For example, the SAT and ACT are both used for college admissions. Scores on the tests need to be linked before they are compared. A symmetric score link is a transformation between the scores from one test to those from another such that the path from scores on test Y to test X is the inverse of the path from test X to Test Y. In one table, an ACT score of 28 corresponds to an SAT score of 640; in the inverse table, a 640 corresponds to a 28.

Features of testing situations affect the type of score linking or scale aligning that can be achieved. These features include test content, target test-taker groups, and conditions of measurement. Different types of linking that vary with respect to these dimensions are score equating, linking scores from tests in transition, concordance, and vertical scaling. Inferences that can be made from each of these types of linking scores are constrained by the features of the testing situation.

situation.

Score equating is the highest form of score linking. Its goal is to produce interchangeable scores. Large-scale testing programs develop different editions of the same test from a common blueprint. Equating adjusts for differences in the difficulty of different test editions to produce interchangeable scores. Large representative samples of examinees, sound data collection practices, and appropriate methods are needed to produce equated scores on these test editions. These conditions benefit all types of score linking.

The interchangeability of scores associated with equating is not achieved, however, simply because proper numerical operations have been performed. Only tests that measure the same construct can be equated. A math test can be linked to a reading test, but it cannot be equated to a reading test. Likewise, even though both are measures of size, height cannot be equated to weight. A short test that produces erratic scores cannot be equated to a long test that produces very stable scores. In order to be equated, tests need to produce equally reliable scores. The relationships between scores on two equated tests need to be the same across different subgroups, such as males and females.

Another scenario occurs when there is an interest in linking scores across related but distinct tests. The term *concordance* is used to describe this type of linkage. Typically, the tests measure similar constructs, are administered to similar kinds of examinees, and are used for the same purpose but differ in test specifications. In some cases, one test is a redesign of the other, as with an old and new version of the SAT.

Score linking is more challenging when it involves two different tests, such as the SAT and ACT, that are produced by different assessment companies. Without a concordance, which provides a data-based path between the score scales of the two tests, the temptation to use the norms for each test might prove irresistible. The use of norms tables from different tests, such as the SAT and ACT, however, presumes that the groups on which the percentiles are based are equivalent in ability. That is rarely the case. For example, a particular test taker is more likely to achieve a score at the 75th percentile or above on the ACT than on the SAT because the SAT norms group is more able than the ACT group. A concordance uses data to dispel a misconception of equivalence between groups that are not equivalent. This should lead to fairer treatment of test takers. Unlike equating, which produces the same link between scores across subgroups, concordances between scores from different tests such as the SAT and ACT are subgroup dependent; for example, the linking differs for males and females.

subgroup dependent; for example, the linking differs for males and females. Consequently, concordance tables need to be used with more care.

There is often an interest in comparing performance across tests of different levels of difficulty for a given construct. In the realm of K–12 testing, test scores are often compared across grades even though test content and test-taker groups differ. Linkages of this sort must ensure that the comparisons are meaningful despite the changes in content and examinees that occur with change in grade level. Vertical scaling is the term used to describe this linkage.

The proper interpretation of a score comparison depends on several features of the testing scenario as noted earlier. The interpretability of all of the score comparisons discussed earlier also depends on the design and execution of a sound data collection plan and proper data analysis to achieve the establishment of linking scores between the compared scores. There are a variety of data collection designs and data analysis procedures that can be used to link scores. The use of these designs relies on equivalent (sometimes identical) groups of test takers or the administration of common test material across different groups of test takers.

What happens when someone wants to compare scores from two or more assessments that are built to different specifications and are administered to different test-taker groups under different conditions? For example, the linkage between scores on tests administered in different languages to different language groups are made without common test material or equivalent groups of test takers. Here, untestable assumptions may be made, for example, that a reading item in Chinese is the “same” as its translation in English, to arrive at a presumed linking. Although such conjectural linking scores might satisfy a craving for comparison, they should be viewed as speculative until buttressed by empirical data.

Neil J. Dorans

See also Admissions Tests; Alignment; Cut Scores; Gain Scores, Analysis of

Further Readings

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). (Eds.) Linking and aligning scores and scales. New York, NY: Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187–220). Westport, CT: American Council on Education.

Kolen, M., & Brennan, R. (2014). Test equating, scaling, and linking: Methods and practices (3rd ed.). New York, NY: Springer.

Chad M. Gotch Chad M. Gotch Gotch, Chad M.

Mary Roduta Roberts Mary Roduta Roberts Roberts, Mary Roduta

Score Reporting

Score reporting

1479

1482

Score Reporting

Score-reporting concerns the delivery of test performance information to examinees and other stakeholders. As one of the most visible components of a testing program, score reports fulfill an important role within the testing process, specifically, and within the education system, more generally. Although a simple notion, there exists great variety in the form score reporting may take. Score reports are generated within multiple assessment contexts (e.g., accountability, college admissions, certification/licensure) and may focus on an individual examinee or groups of examinees as a whole. The information reported may be summative, diagnostic, or normative. Intended audiences of the report may vary from a student to parents to policy makers to the general public. Score reports can be paper-based, web-based, or some combination thereof with static or dynamic presentation of information. In some cases, the report may be accompanied by information to help the reader interpret the report. Across these varied contexts and representations, some unifying interpersonal, psychometric, and visual design themes and considerations exist. This entry reviews these aspects of score reporting, examines how technology is influencing score reporting, and notes potential areas of growth within the research base.

Score Reporting as Communication

Score reporting is often cast as an act communication. A concern of score report developers and researchers is the clarity of communication and how well stakeholders understand score report content. Research has shown the recipients of score reports often experience difficulty forming accurate interpretations of score scales, score comparisons, score meanings, statistical significance, and measurement error, for example. Therefore, much attention has been given to interpretive guidance and the presentation of key elements of examinee performance.

Another area of attention in score reporting is the needs of stakeholders. Representative educational stakeholders may be engaged early on in the score report development process to identify effective means for conveying the information they need to perform their personal and professional responsibilities. Also of interest may be the motivations stakeholders bring to the communication exchange and what they need to respond effectively to the score report. Stakeholder engagement may be iterative, with multiple versions of score reports being generated, piloted, evaluated, and refined. In consideration of the central role of score reporting within contemporary notions of test validity, educational measurement professionals have suggested evaluation of score reports focus on how users interpret and use the information that is communicated.

Finally, no instance of communication has a neutral effect on the parties involved. With each act of communication, the roles, understandings, motivations, and goals of the involved parties are extended, reaffirmed, clarified, or disrupted. With this perspective in mind, some have argued to view score reporting not just as an event of information transmission, but rather one that is impactful upon the relationship between the score report recipient and the testing agency. In this approach, design and delivery should take into consideration how authority is conveyed by the testing agency (to potentially both good and ill effect), how responsibility for interpreting and acting upon the report is assigned to various players (e.g., teachers and parents), how interpretive guidance acts to sharpen or soften the claims made about the examinee, and simply how the size and placement of report elements implicitly conveys their relative importance within the report. Rather than there being prescriptive guidance in these considerations, sound score report development will apply the considerations to the particular cultural and contextual aspects of the specific reporting environment.

Test Score Properties

The central focus of a score report is typically the test score. Considerable effort has gone into characterizing the properties of test scores and using this knowledge to inform decisions around content and presentation. Perhaps the score property most often studied in the context of score reporting is reliability, and its related concept, the standard error of measurement. It is common to see the standard error of measurement depicted visually through portrayals such as error bars. One may also find statements of score ranges representing the variation an examinee would likely experience by taking the same test on different hypothetical occasions. In cases of reporting scores with less precision, an examinee’s performance may be represented by a category (e.g., *above standard*) rather than a score to facilitate appropriate user interpretations.

Test structure represents another score consideration, particularly when reports include information about examinee performance across multiple dimensions or content areas. For example, a report may contain sections for how well a student performed in English Language Arts and Math. Those content areas may be broken down further into separate components such as reading, listening and speaking, and writing (English Language Arts) and concepts and procedures, problem solving, communicating reasoning (Math). These subject-or component-specific reports of performance typically follow the test’s table of specifications, representing, at a minimum, some content-based rationale for how examinee performance is reported. Further evidence to support such reporting—and implied use of scores—may be provided through psychometric analysis of score dimensionality. As such, one may find technical documentation of test structure in supporting materials that accompany the score report.

Beginning in the late 2000s, concerted efforts combined considerations of both score reliability and test structure to examine the *value* of reporting scores beyond the total test score. These *subscores* reflect examinee performance on subsets of test items. Regardless of the intended purpose of a test, examinees and other stakeholders often express interest in more fine-grained reports of performance for uses such as diagnosis of strengths and weaknesses, program admission, placement, and planning intervention. The inherent challenge to subscore reporting is that as the test is divided into subsets, the reliability of the resulting subscores is likely to decrease. Thus, reporting subscores may encourage invalid interpretations and uses (e.g., acting in response to small differences between scores).

In 2008, Shelby Haberman introduced a method based on classical test theory for assessing subscore value. The goal was to identify whether or not a subscore can provide a more accurate measure of its intended construct than the total test can provide a more accurate measure of its intended construct than the total test score. This method developed into a broader framework that considers how subscores related to one another and to the total score. Alternative methods for assessing subscore value have been introduced, such as the value added ratio from Richard Feinberg and Harold Wainer. Different perspectives and methods have led to lively debate in the field of educational measurement. Some question whether or not the outputs provided by different methods are practically or clinically significant. Others suggest the purpose of the test (e.g., to discriminate between individuals at one point in time or to assess learning of students longitudinally) should be of foremost consideration. As a whole, the subscore debate ties to both conceptual notions of test validity and practical concerns such as public trust in large-scale testing.

Visual Design Principles

Score reports include visual elements such as numbers, graphs, images, and narrative text. Guidelines for the design of score reports have been informed by universal design, cognitive psychology, and aesthetics. These multiple perspectives are employed to support readability and interpretation. For example, design techniques informed by cognitive psychology build on the premise of supporting the brain’s tendency to actively interpret and make meaning of what is presented. Judicious use of signaling techniques (e.g., color and font) and organizational techniques (e.g., alignment of text and grouping of similar information together) can draw the reader’s attention to important aspects of the report while promoting a visually pleasing and coherent presentation of information.

Visual design choices follow the specific needs of the reporting context. For example, score reports generated for diagnostic purposes could report a profile of scores across multiple skills, using graphical representation and supporting narratives about an examinee’s strengths and weaknesses. In a certification context, comparatively, visual elements of the report may signal an emphasis on a single scaled score number representing the examinee’s performance and portray that score in relation to a passing score.

Technology and Dynamic Reporting Systems

With increasing prevalence of computer-based and online testing, score reports have evolved from their paper-based origins. Web-based environments afford have evolved from their paper-based origins. Web-based environments afford additional flexibility in how score reports can be presented. Online reporting systems can facilitate increased management of and access to student performance data. Web-based tools to support information management could include the use of hyperlinking to additional content, multiple tabular displays, and embedded multimedia formats. Interactive reporting environments have potential to support user-driven interactions with provision of options to customize information and tailor the reports for specific contexts, audiences, and uses.

A common application of technology to score reporting is the creation of an interactive score report with dynamic presentations of information. For example, interpretive information accompanying a score report could be embedded and accessed throughout the report by “mousing over” areas of interest. In addition to interpretive information, online tutorials have been explored with teachers to provide basic information on assessment concepts relevant to the report.

Building the Research Base

To date, score-reporting research has focused largely on methods for communicating technical information to target audiences. This approach has resulted in the production of reporting guidelines based on best practices and communication design principles. Using methods such as interpretive tests, feedback surveys, and think alouds, empirical research efforts have explored the degree to which score reports meet user information needs and the extent to which users comprehend and form accurate interpretations of score reports. These achievements in score-reporting research have been valuable, and the field would benefit from advances in the articulation of robust and diverse theoretical frameworks to inform future empirical research. In particular, opportunities for scholarship exist in examining the social dimension of score reports, more specifically how to include multiple perspectives representing varied backgrounds, attitudes, and histories interacting with assessment systems. In addition, there is more to learn about what score-reporting outcomes are both relevant and meaningful to multiple educational stakeholders. A comprehensive understanding of the audiences of score reports, from administrators to teachers to parents and students, early on in the test development process, will result in score reports that better meet their information needs and support appropriate and accurate interpretations. In turn, score reports can function not only as a medium for communicating test performance but also as a vehicle for advancing learning and instruction.

learning and instruction.

Chad M. Gotch and Mary Roduta Roberts

See also Accountability; Consequential Validity Evidence; Data Visualization Methods; Data-Driven Decision Making; Proficiency Levels in Language; Reliability; Standard Error of Measurement; Standardized Scores; Summative Assessment

Further Readings

Behrens, J. T., DiCerbo, K., Murphy, D., & Robinson, D. (2013, April). Conceptual frameworks for reporting results of assessment activities. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. Applied Measurement in Education, 17, 145–220. doi:10.1207/s15324818ame1702_3

Haberman, S. J. (2008). When can subscores have value? Journal of Educational and Behavioral Statistics, 33(2), 204–229. doi:10.3102/1076998607302636

McCrudden, M. T., Schraw, G., & Buckendahl, C. W. (2015). Use of visual displays in research and testing: Coding, interpreting, and reporting data. Charlotte, NC: Information Age.

Roduta Roberts, M., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. Educational Measurement: Issues and Practice, 29(3), 25–38. doi:10.1111/j.1745–3992.2010.00181.x

Tufte, E. (2001). Visual display of quantitative information (2nd ed.). Cheshire, CT: Graphics Press.

Zenisky, A. L., & Hambleton, R. K. (2015). A model and good practices for score reporting, In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 585–602). New York, NY: Routledge.

David Torres Irribarra David Torres Irribarra Irribarra, David Torres

Scree Plot

Scree plot

1482

1484

Scree Plot

A scree plot is a graphical tool used in the selection of the number of relevant components or factors to be considered in a principal components analysis or a factor analysis. Proposed originally by Raymond Cattell in 1966 in his article *The Scree Test for the Number of Factors*, the scree plot has become a widely used tool to deal with the issue of component and factor selection. Conceptually, the scree plot is a way of visualizing the magnitude of the variability associated with each one of the components extracted in a principal component analysis. This plot allows researchers to examine the pattern of decreasing variability attributable to each successive component in order to inform the selection of how many such components should be considered relevant for interpretation in a principal component analysis or extracted for inclusion in a subsequent factor analysis.

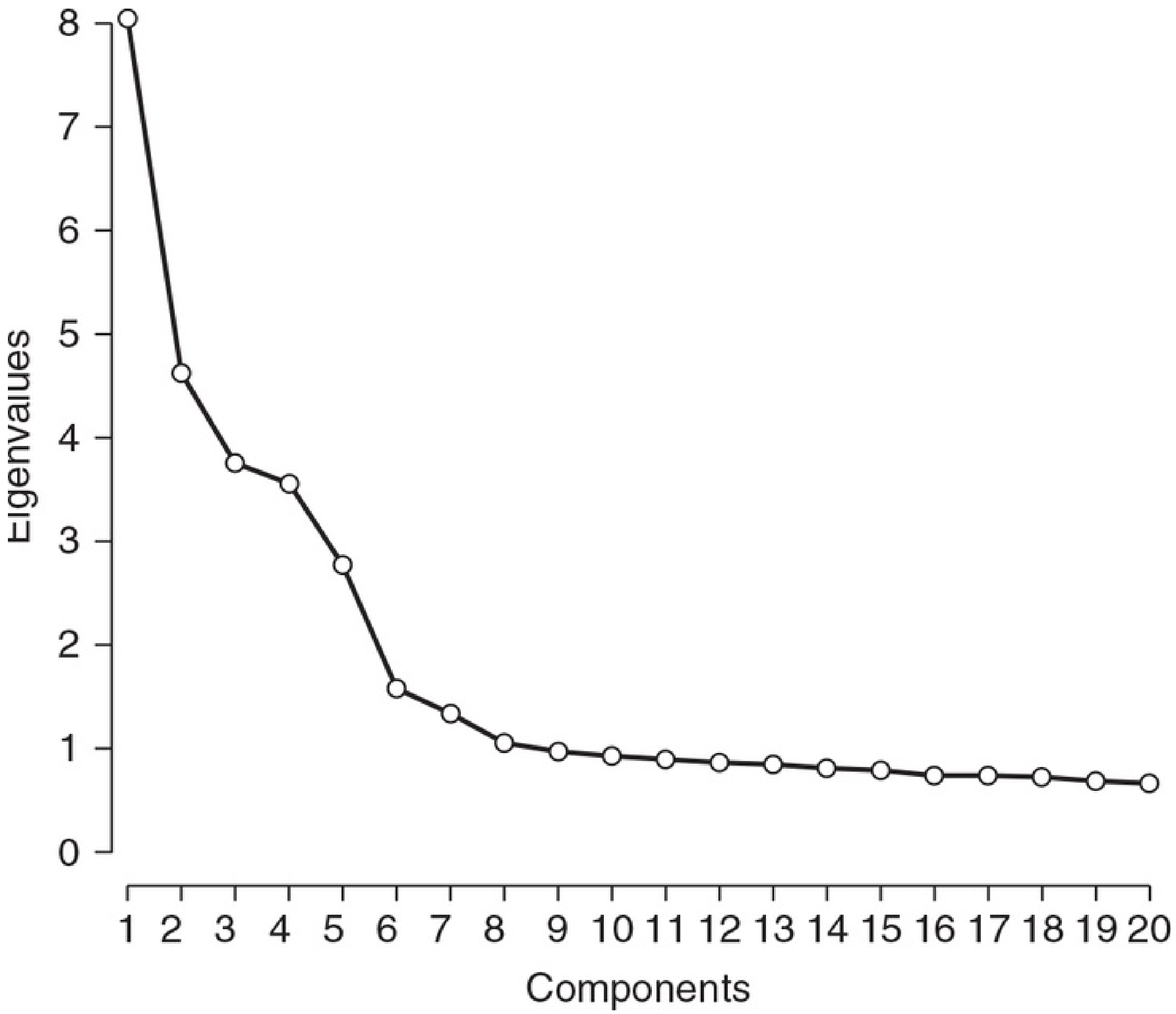
The scree plot takes its name from the characteristic pattern observed in these plots which resembles a mountain side that becomes less and less steep, until it flattens as it reaches the debris and loose stones at its base. In his article, Cattell described the rationale for the name as follows:

> Such a plot falls first in a steep curve but then straightens out in a line which runs with only trivial and irregular deviations from straightness to the nth factor… This straight end portion we began calling the *scree*—from the straight line of rubble and boulders which forms at the pitch of sliding stability at the foot of a mountain. The initial implication was that this scree represents a “rubbish” of small error factors. (1966, p. 249)

The use of a scree plot as a method for component selection relies on the visual judgment of the pattern observed in the plot, specifically the separation between the components that are part of the scree from the relevant components to the left of it, as opposed to alternative component selection methods that do not rely on subjective judgement by using, for instance, a specific cut point (e.g., the Kaiser rule), a statistical test (e.g., Bartlett’s chi-square test), or a computational procedure (e.g., parallel analysis).

Figure 1 shows an example of a scree plot produced through a principal component analysis of a data set of responses to the 50 items of the Big Five Personality Inventory. This example shows the traditional pattern resembling a steep mountainside created in this case by the first five extracted components, followed by the flatter scree produced by the remaining components. It is possible to argue that Figure 1 shows a double scree, the first one occurring between Components 6 and 8 and a larger one starting on Component 9, but considering that the theory behind the instrument that is being analyzed, it is reasonable to adopt the five component solution based on the break in the slope that occurs after the higher scree.

Figure 1 A scree plot example from a principal components analysis of the 50 items of the Big Five Personality Inventory with data from the International Personality Item Pool.



However, it is not uncommon to encounter cases in which there are two or more screes or in which there is no clear discernible break in the slope, which makes the judgement regarding the number of components to consider unclear and highlights the role of subjective judgement involved in the use of the scree plot as a method of component selection. In their review of multiple methods for selection of number of components, Wayne Velicer and colleagues conclude that the scree plot has a mixed track record in studies that evaluate its accuracy of recovery of components, but that the visual examination of the scree plot can prove useful when used in conjunction with other component selection methods.

Building a Scree Plot

The scree plot is created based on the principal component analysis of the correlation matrix of a data set; specifically, this analysis consists of the correlation matrix of a data set; specifically, this analysis consists of the eigendecomposition of the matrix into a set of orthogonal (i.e., independent) eigenvectors and their respective eigenvalues. In the context of a principal component analysis, the eigenvectors correspond to principal components, whereas the eigenvalues associated with each principal component correspond to the variance associated with that component.

The eigendecomposition of a correlation matrix of *N* items will produce *N* eigenvectors and their respective *N* eigenvalues; in other words, the principal component analysis will yield a solution with N principal components, each one of them accounting for a proportion of the total variance. So in general terms, a scree plot is created by listing the extracted principal components from 1 to *N* in the *x*-axis while plotting the eigenvalues associated with each one of the components on the *y*-axis.

In this way, a scree plot will have in principle as many components in the *x*-axis as there are variables in the correlation matrix that is being analyzed, that is to say, an analysis of a set of 50 variables will yield 50 eigenvalues that could be plotted along the *x*-axis. However, researchers may choose to plot fewer components under the assumption that there will be a relatively small number of relevant components before the scree becomes apparent in the plot (as it is the case in Figure 1, where only 20 of the 50 eigenvalues are presented).

Regarding the interpretation of the eigenvalues plotted along the *y*-axis, researchers can take advantage of the fact that the eigendecomposition is being performed over a correlation matrix, whereby the variance of each one of the variables is equal to 1 and the sum of all eigenvalues will add up to *N*. Hence, we can interpret that a principal component accounts for an amount of variance equivalent to the variance of a number of variables equal to its eigenvalue. For example, in Figure 1, we analyze a correlation matrix of a set of responses to 50 items, where the first principal component has an eigenvalue of approximately 8, which can be interpreted as indicating that the first component is accounting for a proportion of variance equivalent to close to 8 of the 50 original variables.

Nongraphical Solutions for the Scree Plot

Because the scree plot relies on the judgment that an observer makes regarding the graphical pattern of the eigenvalues, it is always possible that different observers decide on different numbers of relevant components to interpret or include in a factor analysis. This include in a factor analysis. This subjective component is not present in other methods for selecting the number of relevant components as they rely, for instance, on algebraic solutions or predetermined cut points. However, because of the development of the scree plot as a technique for component selection, researchers have developed complementary nongraphical techniques that could be used to remove observer judgement when interpreting a scree plot. These solutions include Keith Zoski and Stephen Jurs’s linear regression approach and the methods proposed by Gilles Raîche and colleagues that focus on the automatic detection of the scree (the scree test optimal coordinate) and the detection of the break or “elbow” between the scree and the relevant factors (scree test acceleration factor).

David Torres Irribarra

See also Exploratory Factor Analysis

Further Readings

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245–276.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179–185. doi:10.1007/BF02289447

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J. G. (2013). Nongraphical solutions for Cattell’s scree test. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9(1), 23.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy (pp. 41–71). New York, NY: Springer Science+Business Media.

Zoski, K. W., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. Educational and Psychological

Measurement, 56(3), 443–451. Retrieved from https://doi.org/10.1177/0013164496056003006

Theodore J. Christ Theodore J. Christ Christ, Theodore J.

Michael Herriges Michael Herriges Herriges, Michael

Screening Tests

Screening tests

1484

1487

Screening Tests

Screening is an efficient procedure used to identify potential problems or risks for future problems. In this case, a problem is a discrepancy between what is expected and what occurs. Thus, effective screening tests accurately identify those who are likely to not reach critical future outcomes, so that their problems can be remediated through prevention or intervention efforts. Screening tests are common in many disciplines and are used for numerous purposes. This entry discusses the basic principles and technical considerations of screening tests.

In education, screening is the systematic assessment of all students within a classroom, grade, school, or school district. Screening is often done in academic, behavioral, or social–emotional domains. If all of the students within a targeted population are evaluated, then the screening is *universal*. Screening tests may assume many formats, such as brief timed batteries of key academic competencies, computer adaptive tests of broad scholastic achievement, teacher ratings of student behavior, self-report measures of psychological symptoms, or many additional forms. Regardless of the test format, screening informs teachers and other educators in the process of data-based decision making. Their purpose, then, is not just to identify problems but also to provide structured guidance for altering service delivery and resource allocation within schools and school systems.

Historically, screening has been used to highlight existing skill deficits or other shortcomings inherent to individuals (e.g., learning disabilities, mental health issues, behavior problems). However, the current perspective proposes screening as a method to determine the extent to which general school services, such as core academic curricula or behavior intervention systems, are meeting the needs of all students in addition to identifying individual students not on track for success. In this context, screening tests can be compared to thermometers, in that they are holistic indicators of the “health” or “wellness” of an educational system, and those individuals who comprise that system.

In contemporary education, universal screening and effective screening tests are an essential component of multitiered systems of support, wherein intervention services of varying intensity are allocated based on needs that are often identified by screening. Multitiered systems of support are based on the assumption that core preventive services and the early identification of educational problems help educators to address student needs, so that their struggles are not later exacerbated. Effective screening tests permit this early identification and intervention, which subsequently improves the efficiency of service delivery and resource allocation for schools and school systems.

Basic Principles

The process of universal screening in education involves administering a test to all students in a population, scoring the tests and recording student performance, analyzing the data, and interpreting the results to inform decision making. This is a complex process that necessitates the cooperation and coordination of multiple key school personnel. The formation of a school leadership team that is responsible and accountable for all steps in this process has been recommended in the professional literature. To ensure that screening tests are employed effectively and efficiently, some basic principles and effective practices should be identified.

Test Selection

Screening tests are used to assess the performance of individual students as well as the presence and severity of problems in educational systems. These tests are used for decision making and allocating educational services for students based on the intensity of their needs. Therefore, screening tests should be selected based on the intended uses and interpretations of test scores.

Screening tests commonly assess student ability and performance with regard to core academic and functional domains (e.g., reading, math, behavior). It is advised that tests be aligned with core curricula and instruction and linked to academic or behavioral goals and standards. Greater alignment between test content and these criteria will result in more accurate decisions about student performance and educational need. Screening tests that are misaligned with regard to schoolwide standards and objectives will be less effective at identifying the academic needs of students.

Another important consideration for test selection involves a review of the psychometric evidence supporting the screening measure. A qualified member of the school leadership team may assume responsibility for reviewing the technical documentation for a screening test and synthesizing this information to make an appropriate selection.

Administration and Scoring

Screening tests must be administered with sufficient frequency to inform educational decision making. However, there is also a trade-off between the amount and frequency of test administration, and the time and resources spent to administer a test. Typically, schools administer screening tests at the beginning, middle, and end of each school year (e.g., fall, winter, spring) for all students. Administering screening tests at equal intervals throughout the school year allows educators to evaluate student performance at each screening period as well as growth across those time frames.

Different screening tests require different formats for administration, which impacts the efficiency of data collection. Some screening tests are administered to students on an individual basis and require manual scoring. More recently, test developers have created computer-based screening tests that can be group administered and feature automated scoring. These computer-based measures allow for more efficient screening procedures and often offer information for test interpretation. Aspects of test administration and scoring depend on the flexibility of school personnel, time and space resources, and technological tools available. Any approach for conducting universal screening should be planned and documented, so that the process runs smoothly.

Analysis

The administration, scoring, and analysis of screening test data should be

The administration, scoring, and analysis of screening test data should be completed in a timely manner so that decisions are made promptly and teachers can focus their efforts on instruction. Past recommendations suggest that the entire process of universal screening should take no more than 2 weeks. Once all screening data are scored and recorded, class reports with student scores should be generated. Teachers and other educators may form grade-level teams to determine whether core instruction is meeting student needs and identify individual students who may require supplemental services. The use of benchmarks and decision rules can facilitate data analysis.

Within a response to intervention framework, it is a common assumption that approximately 80% of students should be responding to core instruction. Therefore, approximately 80% of students should be scoring on par or better than benchmarks established for screening tests. If a greater proportion of students are not meeting expectations, then a grade-level or schoolwide problem exists, and changes should be made to the core instructional program. Conversely, if screening test data indicate that core instruction is acceptable for at least 80% of students, decision rules can be applied to identify individuals in need of intervention. Students should receive small-group or individualized intervention services that match the nature and intensity of their problem.

Technical Considerations of Screening Tests

Screening tests are used to determine which students are at risk for educational failure and to inform decision making about additional instructional needs. Screening tests aid this determination by producing data, or observations of behavior within an identified skill area. These data represent a partial sample of behavioral information that is used to generate an inference about ability or future performance. This inference is the conclusion that results from the interpretation of an incomplete set of information.

Because limited samples of behavior cannot perfectly predict future performance, some of the inferences derived from screening data will be incorrect. However, although all screening tests are imperfect, the tests that yield higher quality data will also help educators make more accurate inferences. Certain technical information and psychometric properties of screening tests indicate the relative strengths and weaknesses of screening tests. Educators responsible for evaluating and selecting screening tests should be familiar with these concepts to ensure that the best tests available are selected based on their intended

intended use.

General Information

Screening tests should come with a technical manual that organizes information relevant for the application and interpretation of the test. A strong technical manual will begin with a clearly articulated overview and purpose of the test that includes a theoretical rationale and justification for its application as a screening tool. The manual should also present information regarding the expected qualifications of test users and outline the training required to administer the screening test with competence. Relatedly, the eligible population of examinees should be described.

Screening tests should also present applied information concerning the development and administration procedures of the test. Information about how the screening test was developed may come in the form of a test blueprint, wherein multiple test characteristics are defined. These include item format, content balance, the number of items per test, stimuli presentation methods, item scoring procedures, and information about test score interpretation (e.g., norm or criterion referenced). There should also be clarification about item writing guidelines and how the items were assembled into actual test forms. A final, essential component of general test information is a description of procedures for a standardized test administration. Ensuring a screening test is administered in accordance with the intended procedures is a critical aspect of obtaining reliable and valid measurements.

Psychometric Characteristics

Before a screening test is published and distributed, it should go through a rigorous psychometric evaluation. The results of this evaluation should be shared in the test manual, so consumers can determine whether the test is appropriate for their needs. While the scope of this chapter does not involve a nuanced account of each psychometric detail, test reviewers should understand and seek this information to be confident that a screening test is sufficiently empirically supported.

A test's technical manual should describe the experimental design used for field-testing the screening tool. A thorough report of the population that participated in field-testing should be included, with information disaggregated by relevant in field-testing should be included, with information disaggregated by relevant demographic categories (e.g., age, sex, racial/ethnic category, special education status). In addition, the test theory used for analyses (e.g., classical test theory, item response theory) should be described and justified. Finally, individual item statistics should be reported, including item difficulty parameters, item discrimination, and other information as needed.

A technical manual must also present psychometric information about the reliability of a screening test. In general, reliability refers to the overall consistency of a test. A screening test with high evidence of reliability will yield similar results across repeated administrations. There are multiple dimensions of reliability that can be used to support a screening test, including test–retest, alternate form, and inter-rater reliability. Most estimates of reliability should also be presented with a standard error of measurement, which estimates the precision of measurement for a screening test. The standard error of measurement is closely related to the concept of reliability, in that a test with a low standard error of measurement will also show high estimates of reliability.

Finally, a technical manual should also present information about the validity of screening tests. According to the *Standards for Educational and Psychological Testing*, developed jointly by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education in 1999, validity refers to “the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test” (p. 184).

There are multiple perspectives and dimensions of validity, which cannot be addressed fully in the limited space of this entry. A technical manual must report evidence for content validity, which determines the extent to which a screening test is adequately measuring the trait that it reports to measure. It should also present evidence for criterion validity, which evaluates the extent to which the screening test yields measurements that correspond with other theoretically similar assessment tools, and is useful in predicting performance on important indicators of future success. Depending on the purpose of a screening test, additional validity evidence may be warranted.

Theodore J. Christ and Michael Herriges

See also Alignment; Content Validity Ratio; Criterion-Based Validity Evidence; Curriculum-Based Assessment; Curriculum-Based Measurement; Formative

Assessment; Formative Evaluation; Progress Monitoring; Response to Intervention; Validity

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). Universal screening in educational settings: Evidence-based decision making for schools. Washington, DC: American Psychological Association.

Alison L. Bailey Alison L. Bailey Bailey, Alison L.

Second Language Learners, Assessment of Second language learners, assessment of

1487

1489

Second Language Learners, Assessment of

Assessment of second language (L2) learners can have a variety of purposes including monitoring progress toward the acquisition of an L2 and determining attainment levels for higher education admission and occupational and social prerequisites (e.g., entry into health care and legal professions; requirement for citizenship eligibility). This entry focuses on L2 assessment in educational settings used to inform educators and students about learner language and literacy competencies. Language constructs, assessment types, and contemporary L2 assessment themes (e.g., validity of inferences, dimensionality, influences of age, and uses of technology) are briefly addressed.

L2 Constructs

L2 assessments measure learner's receptive and expressive language abilities in both oral and written modalities. These abilities and modalities are traditionally described as the domains of listening (receptive/oral), speaking (expressive/oral), reading (receptive/written), and writing (expressive/written). Domains include specific constructs, and two or more domains may be integrated within a single assessment format. For example, speaking includes constructs such as pronunciation, fluency (i.e., the rate of speech and degree of pausing and dysfluency), productive vocabulary, and discourse organization. Reading includes the constructs of phonological processing, vocabulary and grammatical knowledge, and understanding of extended discourse; and writing includes orthographic knowledge, vocabulary and grammatical usage, and organization of text. Advanced statistical modeling such as multidimensional scaling analysis is used to examine the underlying factor structure of these constructs, revealing underlying subskills and whether assessments measure a unitary construct of L2 proficiency. Integrated assessment expressly targets two or more constructs and reflects the authentic ways in which different domains of language are simultaneously processed and used (e.g., a measure integrating constructs within both the listening and speaking domains).

Beyond these fundamental language constructs are additional constructs assessed to further characterize language. Knowledge of linguistic forms and features alone does not result in successful meaning-making with interlocutors; rather, this requires communicative competence. Such knowledge of the functions of language in diverse contexts, while typically neglected in assessment, is an important part of the L2 construct. Pragmatic abilities, for instance, focus on the socially appropriate uses of language such as culturally prescribed politeness routines and knowledge of the registers or conventions for language usage in particular contexts (e.g., the language of academic settings).

Special considerations for the assessment of academic content (e.g., mathematics and science) of L2 learners are outside the scope of the current entry, but the intertwined nature of language and content and related issues of language as a source of measurement error, construct irrelevance, and modifications (i.e., accommodations) to simplify language demands on academic achievement tests have been widely studied.

Assessing Characteristics of L2 Learners

There are also constructs relevant for predicting language-learning outcomes such as L2 learner characteristics. Learner characteristics have frequently been measured because of their value in making suitable placements for instruction and for predicting attainment outcomes. Aptitude for L2 language learning, attitudes, and motivation have long been operationalized and measured in the L2 arena.

More recently, self-regulation (i.e., consciously identifying and planning goals) has been found to predict vocabulary outcomes with L2 learners. An important consideration in the operationalization and measurement of L2 is its distinction from foreign language assessment; learners acquire an L2 in settings where the L2 is the dominant societal language, whereas a foreign language is typically acquired as an academic subject in societal settings where it is not dominant.

Language Acquisition Theory and Assessment Types

If the overarching purpose of L2 language and literacy assessment is to capture language growth and attainment in immersive settings, assessment development should be guided by theories of *how* language progresses and *what* aspects of language progresses in such settings. Moreover, attention to modern assessment theory in the L2 field has led to the adoption of a framework of validity arguments that focus on the strength of claims about interpretations and uses of language assessment. L2 assessments may measure skills with a discrete point approach (e.g., testing the accuracy of individual grammatical affixations used to mark tense, gender, or number on verb forms), or they may be guided by theories of communicative language ability that align with authentic language use.

Large-scale (direct) summative assessments of language proficiency are used for educational accountability purposes and have traditionally measured discrete skills. However, with the advent of computer-based assessment, item types can more effectively integrate language skills into tasks that mirror the language demands of K–16 classrooms, such as working with an on-screen avatar to cocreate an explanation that approximates the processes involved in working with a real-life classmate.

L2 assessment types that have traditionally been widely used include oral interview techniques (e.g., the Oral Proficiency Interview) that simulate conversation, cloze tests (fill in the blank), and elicited imitation tasks. However, increasingly in adult English as a Second Language instructional environments, there has been a focus on learning-oriented assessment, analogous to adoption of formative assessment for learning in K–12 classrooms, that focus on teacher's feedback and peer and self-assessment. Additional technology integration in L2 assessment includes the use of automatic speech recognition for evaluating spoken language.

Assessment of Young L2 Learners

With maturation and/or language instruction, L2 learners build on their repertoire of skills to increase proficiency, and assessments need to be designed to capture this development in age-appropriate ways. Unlike the assessment of many other knowledge domains that are sequenced by age, L2 development may begin at any age and even very young students can become highly proficient L2 speakers. However, L2 assessment with young language learners needs to be designed to take into account children's limited attention spans, greater fatigue, and lack of testing familiarity, as well as cultural and curricular influences.

A key concern for assessment with school-age L2 learners in particular is the need to guard against misidentifying phases of L2 development as language disability. It may even be necessary to assess both the L1 and L2 of learners to get an accurate understanding of their complementary language abilities for appropriate intervention decisions. Degree of accuracy is of course of great consequence in any L2 assessment situation. The technical quality of assessments for L2 learners (e.g., strong evidence of validity and reliability) is important in no small part because of the influence assessment results may have on a learner's future instructional experiences.

Alison L. Bailey

See also Accountability; Classroom Assessment; Computer-Based Testing; English Language Proficiency Assessment; Formative Assessment; Literacy; Multidimensional Scaling; Reading Comprehension Assessments; Self-Regulation; Speech-Language Pathology; Student Self-Assessment; Summative Assessment; Technology-Enhanced Items; Written Language Assessment

Further Readings

Abedi, J. (2010). Linguistic factors in the assessment of English language learners. In The Sage handbook of measurement (pp. 129–150).

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford, UK: Oxford University Press.

Bailey, A. L. (n.d.). Assessing the language of young learners. In Shohamy & N. H. Hornberger (Eds.), Encyclopedia of language and education, Vol. 7: Language testing and assessment (3rd ed.). Berlin, Germany: Springer.

Bailey, A. L., Heritage, M., & Butler, F. A. (2014). Developmental considerations and curricular contexts in the assessment of young language learners. In A. J. Kunnan (Ed.), The companion to language assessment. Hoboken, NJ: Wiley.

InbarLourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), The age factor and early language learning (pp. 8–96). Berlin, Germany: Mouton de Gruyter.

Morrow, K. (2012). Communicative language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), The Cambridge guide to second language assessment. Cambridge, UK: Cambridge University Press.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. Journal of Speech, Language, and Hearing Research, 57(6), 2208–2220.

Röver, C. (2014). Testing ESL pragmatics: Development and validation of a web-based assessment battery. Frankfurt, Germany: Peter Lang.

Tseng, W. T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. Applied Linguistics, 27(1), 78–102.

Turner, C., & Purpura, J. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), Handbook of second language assessment (Vol. 12). Berlin, Germany: Walter de Gruyter, GmbH & Co KG.

Xie, S., Evanini, K., & Zechner, K. (2012, June). Exploring content features for automated speech scoring. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 103–111). Association for Computational

Linguistics.

Kyra N. Fritz Kyra N. Fritz Fritz, Kyra N.

Nicholas K. Lim Nicholas K. Lim Lim, Nicholas K.

Selection Bias

Selection bias

1489

1491

Selection Bias

Selection bias, also known as sampling bias, usually refers to groups (e.g., experimental, control) that are systematically different prior to experimental manipulation or intervention due to the assignment of participants to groups. In other words, variations detected during a study are attributable to group differences due to selection bias or the independent variable (e.g., manipulated variable). Selection bias can occur during participant selection, assignment, and/or during the study. The bias that occurs during participant selection is generally identified as a threat to external validity, whereas bias that occurs during assignment is known as a threat to internal validity. During a study, if a significant number of participants withdraw without completing the study, selection bias can also occur. This entry examines the context in which selection bias may arise, how to avoid selection bias, and the limitations of ensuring selection bias.

Context in Which Selection Bias May Arise

Selection of a sample for a study leading to selection bias may occur when the researchers attempt to generalize their observations beyond the sample to other populations. Participants assigned based on cost, convenience, or to conditions in a manner intended to disperse participant characteristics (e.g., demographics and diagnosis) as opposed to an unbiased selection process may introduce selection bias. For example, researchers who are interested in evaluating a national school-based program may only solicit participants from their region due to cost considerations. The question then becomes whether the results obtained in the study can generalize to participants attending schools in other geographic regions.

Nonrandom assignment of participants in experimental designs (quasi-experimental) are likely to contribute to outcome differences that are not due to the intervention effect; but rather, certain characteristics of the groups being compared. This design is likely to occur especially when random assignment is unavailable to the researchers.

Even if there is a random assignment prior to the beginning of the study, participant attrition, especially in a nonrandom fashion, can end up with a selection bias problem. Participants may withdraw for various reasons that often times are unknown to the researchers. One should be wary when participants withdraw from a study.

How to Avoid Selection Bias

Random assignment of participants to groups is a commonly used procedure to guard against selection bias (as a threat to internal validity). Random assignment minimizes the likelihood that groups will be systematically different prior to introducing the independent variable. With random assignment and when sample size is sufficiently large, it is more likely to produce group equivalence before the independent variable is applied. Another method that may reduce selection bias is random assignment of participants in matched sets to ensure groups are equivalent based on key variables (e.g., age, income, gender, and geographic region). Pretesting participants may also be employed, which provides the opportunity for researchers to evaluate the presence, possible size, and direction of bias. However, even if no pretest differences are found, it does not guarantee the absence of selection bias.

Researchers are generally cautioned against overgeneralizing their conclusions in the face of promising results, even if they are confident in having sufficiently addressed the threat to internal validity, as doing so can still pose a threat to external validity. Ensuring all participants complete the entire study may be impossible without compromising the ethical treatment of participants. Researchers can try to generate high interest and increase motivation of participants to complete the study, while keeping in mind that there is nothing to stop a participant from withdrawing once the study begins. Researchers can employ blind or double-blind designs and provide detailed debriefing after the study. Blinded designs can aid in establishing strong baseline measures and reduce dropouts, but researchers should be wary of potential ethical issues.

Limitations of Ensuring Selection Bias

Although random assignment is often proposed to reduce the likelihood of selection bias, it is not always a practical solution. Random assignment may not always be possible. Returning to the example about examining a national school-based program, it may not be feasible to employ random assignment for classrooms and schools due to school proximity, permission to implement the intervention program in school, or inability to obtain informed consent, for example. In addition, it is certainly not ethical to randomly assign students to various classrooms or schools just for the sake of the study. In this regard, groups are generally preestablished for research purposes. Because randomization is not likely to occur, researchers must make implausible that selection may account for group differences. In other words, researchers must try to control for or hold important extraneous variables constant to ensure that meaningful comparisons between groups can be made. Another possible limitation is the sample size. Small sample sizes can lead to extraneous variables not being well controlled.

Kyra N. Fritz and Nicholas K. Lim

See also External Validity; Internal Validity; Random Assignment; Sample Size

Further Readings

Gravetter, F. J., & Forzano, L. B. (2012). Research methods for the behavioral sciences (4th ed.). Belmont, CA: Wadsworth.

Kazdin, A. E. (Ed.). (2003). Methodological issues & strategies in clinical research (3rd ed.). Washington, DC: American Psychological Association.

Kazdin, A. E. (2003). Research design in clinical psychology (4th ed.). Boston, MA: Allyn & Bacon.

Kazdin, A. E. (Ed.). (2016). Methodological issues & strategies in clinical research (4th ed.). Washington, DC: American Psychological Association.

Leary, M. R. (2012). Introduction to behavioral research methods (6th ed.). Upper Saddle River, NJ: Pearson Education.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Harcourt.

Ryan J. Kettler Ryan J. Kettler Kettler, Ryan J.

Leah Dembitzer Leah Dembitzer Dembitzer, Leah

Selection Items

Selection items

1491

1493

Selection Items

Selection items (or selected response items) are test items on which the examinee selects one of a set of choices, rather than generating an original response. Examples of selection items include conventional and complex multiple-choice items, context-dependent item sets, single and multiple true-false items, alternate-choice items, and matching items. A selection item from a set that is based on a single premise or idea is context dependent; an item alone is context independent.

A primary advantage of selection items is that they are objectively scored based on match to a preexisting key. Thus, scores from selection items tend to be relatively reliable because interscorer agreement is virtually 100%, removing variation based on scorer as one source of construct-irrelevant variance. Also due to objective scoring, selection items can easily be transferred to a computer-based testing medium. Another advantage of selection items is that it is possible for examinees to respond without a certain set of access skills. Constructed-response items that require short, medium, or long answers need examinees to have some ability to formulate thoughts and sentences, as well as to either handwrite or type answers. Examinees who have impairments in these areas may attain depressed scores on tests that are not intended to measure writing or fine motor skills; this is another example of construct-irrelevant variance against which selection items are robust.

A conventional, multiple-choice item typically features a stimulus, an item stem or question, a correct answer choice, and a set of incorrect choices. Although the simplest multiple-choice items offer only one answer choice that correctly satisfies the stem, some include no correct answers and some include multiple correct answers followed by choices of *none of the above*, *all of the above*, or some specific combination of the above (e.g., A and B, but not C). A complex, multiple-choice item may contain a question or stem with answers provided, followed by answer choices regarding whether the provided answers are correct. Multiple-choice items can be written such that each incorrect answer choice selected provides information about the response process of the examinee. The following is a conventional multiple-choice item from a mathematics test:

1. Sarah has $5.00. A hot dog costs $1.50 and a soda costs $1.00. Sarah buys two hot dogs and a soda. How much money does Sarah have left?

   A. $0.00

   B. $1.00

   C. $1.50

   D. $2.50

The correct choice is answer B. Examinees might choose D if they do not multiply anything by 2. They might choose C if they incorrectly multiply $1.00 by 2, instead of multiplying $1.50 by 2. Examinees might choose A if they add incorrectly, or if they round $1.50 to $2.00 prior to multiplying.

A true-false item typically contains a statement that the examinee codes as either entirely true or not entirely true. If any part of the statement is false, then the correct answer is “false.” The following is a true-false item from a social studies test:

> 2. The first capital city of the United States, Washington, DC, was named for the country’s first president.

The correct answer is “false” because the first capital city of the United States was Philadelphia, PA. Even though the other part of the statement is true, the item is false because it is not entirely true.

A similar form selection item is an alternate choice. Alternate-choice items have two response options. The options are different than true or false (e.g., fact or two response options. The options are different than true or false (e.g., fact or opinion, agree or disagree, and any two options). True-false items can also be in sets, known as multiple true-false, or a cluster. Within true-false clusters, a single-item stem contains a set of statements that each require their own true-or-false response.

Matching items are typically in sets that share options from which correct answers are chosen. Often these sets contain the same number of answer choices as items, and the answers are chosen without replacement, such that each item has one answer and each answer has one item. The following items are matching items from a science test.

3. Describes an animal that eats only meat
4. Describes an animal that eats only plants
5. Describes an animal that eats both meat and plants

Answer choices

herbivore

omnivore

carnivore

The answers to items #3, #4, and #5 are carnivore, herbivore, and omnivore, respectively. Matching items are more complex if the number of items and the number of choices are unequal, or if the instructions allow for one choice to be used zero or multiple times.

Criticisms of selection items include that they are susceptible to guessing, can seem tricky to examinees, and may be more suitable for demonstrating knowledge at lower levels of complexity. A multiple-choice item that has one correct option and three incorrect options can be answered correctly in 25% of instances by an examinee who does not even read the item or attempt to process it. For a true-false item, this rate is 50%; and for a matching item, this rate depends on the number of items and choices in the set, as well as the rules about replacement of choices. Selection items can seem tricky to some examinees who may wonder whether the correct answer is available among the choices, whether one word changes an otherwise true statement to false, or whether one answer choice can be the best match for two different items. That being the case, selection items may be susceptible to test-taking strategies such as eliminating implausible choices and selecting the best remaining choice, responding false to any true-or-false item that contains an absolute, or matching items to choices about which one is certain before revisiting choices about which one is uncertain. Selection items seem well suited to remembering basic facts and concepts, the lowest level of Bloom's taxonomy; it is more difficult for selection items to capture levels matched to greater complexity of thought.

Ryan J. Kettler and Leah Dembitzer

See also Constructed-Response Items; Fill-in-the-Blank Items; Matching Items; Multiple-Choice Items; True-False Items

Further Readings

Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, T. M. Haladyna, & M. Raymond (Eds.), Handbook of test development (2nd ed., pp. 293–409). New York, NY: Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309–334. Retrieved from http://dx.doi.org/10.1207/S15324818AME1503_5

Kettler, R. J., Braden, J. P., & Beddow, P. A. (2011). Test-taking skills and their impact on accessibility for all students. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.). Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy (pp. 147–162). New York, NY: Springer.

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. Peabody Journal of Education, 84, 529–551.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Technology & Society, 10(4), 95–109.

Rodriguez, M. C. (2011). Item writing practice and evidence. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy (pp. 201–216). New York, NY: Springer.

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, T. M. Haladyna, & M. Raymond (Eds.), Handbook of test development (2nd ed., pp. 259–273). New York, NY: Routledge.

Schmeiser, C. B., & Stone, C. A. (2006). Test development. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 307–354). Westport, CT: Praeger.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Minkyoung Kim Minkyoung Kim Kim, Minkyoung

Self-Directed Learning Self-Directed learning

1493

1495

Self-Directed Learning

The term *self-directed learning* describes a process whereby individuals take the initiative, with or without assistance, in perceiving their learning needs, setting up learning goals, identifying human and nonhuman resources for learning, selecting and implementing appropriate learning strategies, and evaluating their learning outcomes. It refers to the degree of choice that learners have within an instructional situation. It was originally derived from adult learning and andragogy and now is a core theoretical construct as a field of study distinguished from adult education. This entry describes the nature of self-directed learning, including its importance and benefits in learning, along with theoretical support. It also describes ways to increase self-directedness and facilitate self-directed learning.

Nature of Self-Directed Learning

The philosophical orientation underlying self-directed learning is humanistic in nature. From this perspective, the focus of learning is on the individual and self-development, with learners expected to assume primary responsibility for their own learning with autonomy. Self-directed learning views learners as responsible owners and managers of their own learning process. Self-directed learning assumes that learners are motivated by internal incentives, such as the desire to grow.

There are several related concepts often used interchangeably or in similar ways. Some examples include self-regulated learning, self-planned learning, and autonomous learning. Self-regulated learning is a more process-oriented concept, whereby learners control and evaluate their own learning and behavior to achieve their goals, while self-directed learning focuses more on learners’ initiation of learning. Self-planned learning is defined as a learner’s deliberate attempt to learn some specific knowledge and/or skill, wherein the learner is responsible for the detailed decisions and arrangements regarding the learning activities. Autonomous learning, sometimes called student-centered learning, relates to the change in focus in the classroom from teaching to learning where students are actively involved in a process of knowledge construction through various learning activities and using their prior knowledge. Although all these concepts emphasize student autonomy, they do not mean the learners conduct all their activities on an entirely independent basis.

To better understand self-directed learning, it is important to understand how it differs from teacher-directed learning. The latter assumes the learner is essentially dependent on the teacher who takes full responsibility for what the learner should be taught. In contrast, self-directed learning assumes people become increasingly self-directing as an essential component of maturing. Similarly, in the teacher-directed learning perspective, the learner’s experience is less valued than the teacher’s or teacher surrogate’s, whereas the self-directed learning perspective views the learner’s experiences as an increasingly rich resource for learning. Moreover, teacher-directed learning is usually subject centered, while self-directed learning is typically centered on a task or problem.

Importance of Self-Directed Learning

Learning is ideally a lifelong process in the pursuit of knowledge for either personal or professional reasons. Self-directed learning helps learners to become more effective learners and to develop their own effective learning patterns depending on their learning styles, pace of learning, interests, goals, among other factors. There is growing evidence that individuals who take initiative in learning learn more things better and deeper than those who are passive and dependent in their learning. In particular, the emergence of web-based forms of learning has enabled learners to easily access useful, free, and open learning content, giving learners more power over decisions about what to learn, when to learn, and how much to learn. Self-directed learning opportunities have multiplied in this age of open educational resources, and it is no longer realistic to define the purpose of education as delivering prepackaged knowledge.

The need for a paradigm shift from teacher-centered to learner-centered education has grown stronger as society evolves deeper into the information age. Learner-centered education focuses on how students learn instead of how teachers teach and requires students to take ownership of their learning, which increases the importance of self-directedness.

Theoretical Model of Self-Directed Learning

A comprehensive theoretical model of self-directed learning includes three dimensions that are closely connected and overlapping to some extent: self-management, self-monitoring, and motivation.

Self-management is concerned with the control of the contextual conditions. It focuses on the social and behavioral implementation of learning intentions, as the individual does not construct meaning in isolation. Self-management of learning is a collaborative experience, wherein learners use instructional support, learning materials, and communication with others. Increased learner control brings increased learner responsibilities regarding the learning process and the construction of meaning.

Self-monitoring addresses cognitive and metacognitive processes. Responsibility for self-monitoring reflects a commitment and obligation for the learner to construct meaning through critical reflection and a collaborative confirmation process. Self-monitoring is dependent on both internal and external feedback.

Motivation plays a significant role in the initiation and maintenance of effort to learn and the achievement of cognitive goals. There are two different kinds of motivation: entering motivation and task motivation. Entering motivation is related to the process of deciding to participate, which is concerned with selecting learning goals. Task motivation is related to the effort to stay on task and persist, which is integrally connected to self-management and also closely associated with the issue of volition.

Ways of Facilitating Self-Directed Learning

The ability to be self-directed is situational. One may be self-directed in one subject, yet a dependent learner in another. However, once a learner has developed self-directed learning skills, certain features of self-direction are transferrable to new learning situations. The learning process centers on learner needs rather than content. Therefore, the role of educators is more as a “guide on the side” than a “sage on the stage.” The main role is to act as facilitators and guides rather than to deliver content as experts.

There are several ways to facilitate self-directed learning. First, educators can raise learners’ awareness of their role in the learning process—a personal responsibility orientation—and encourage learners to be involved in decision making for their own learning. Second, educators can provide support matched to the student’s stage of self-direction. Scholars defined four stages of self-direction: (1) dependent, (2) interested, (3) involved, and (4) self-directed. For dependent students, coaching with immediate feedback and informational lecture are good methods. However, educators should be careful to not control too much, which would hinder learner initiative and enhance dependency. For learners who are in Stage 2, interested, educators should take a motivator role, providing inspiring lectures and guided discussions. Involved learners (Stage 3) need facilitators who participate in discussion as equals and facilitate collaborative, small-group activities. Lastly, for learners who are self-directed, educators should assume a mentor role providing mentorship. However, if they withdraw too much, then the educators might end up losing touch and fail to monitor learner progress.

Minkyoung Kim

See also Active Learning; Andragogy; Constructivist Approach; Cooperative Learning; Instructional Theory; Individualized Education Program; Learning Progressions; Metacognition; Motivation; Self-Regulation

Further Readings

Brockett, R. G., & Hiemstra, R. (1991). Self-direction in adult learning: Perspectives on theory, approach and practice. London, England: Routledge.

Brookfield, S. (1984). Self-directed adult learning: A critical paradigm. Adult Education Quarterly, 35(2), 59–71. Retrieved from https://doi.org/10.1177/0001848184035002001

Brookfield, S. (1985). Self-directed learning: From theory to practice (No. 25). San Francisco, CA: Jossey-Bass.

Candy, P. C. (1991). Self-direction for lifelong learning: A comprehensive guide to theory and practice. San Francisco, CA: Jossey-Bass.

Garrison, D. R. (1992). Critical thinking and self-directed learning in adult education: An analysis of responsibility and control issues. Adult Education Quarterly, 42(3), 136–148.

Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. Adult Education Quarterly, 48(1), 18–33. Retrieved from https://doi.org/10.1177/074171369704800103

Grow, G. O. (1991). Teaching learners to be self-directed. Adult Education Quarterly, 41(3), 125–149. Retrieved from https://doi.org/10.1177/0001848191041003001

Knowles, M. S. (1975). Self-directed learning. New York, NY: Association Press.

Merriam, S. B. (2001). Andragogy and self-directed learning: Pillars of adult learning theory. New Directions for Adult and Continuing Education, 2001(89), 3–14. doi:10.1002/ace.3

Shane D. Blair Shane D. Blair Blair, Shane D.

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Self-Efficacy

Self-Efficacy

1495

1497

Self-Efficacy

Self-efficacy is confidence in one’s ability to succeed at a task and is a primary influence in motivating a person. This construct determines various aspects of an individual’s behavior toward a task, including the individual’s thoughts, motivations, and overall performance, especially when the individual faces a difficult task. Depending on an individual’s self-efficacy beliefs, the individual can endure the hardships brought forth by a task in order to attain satisfactory results or can cease any effort toward the completion of the task. Establishing a positive sense of self-efficacy can lead to positive life outcomes. This entry reviews the theories, sources, and correlates of self-efficacy.

Theoretical Views

There are different theoretical views to explain self-efficacy. Three of which are behavioral theory, social learning theory, and attribution theory. Behaviorists theorize that self-efficacy is created and maintained by reinforcement. Someone who receives verbal praise or other reinforcers for completing a task is more likely to seek mastery of that task in order to receive more reinforcers. If the task is naturally enjoyable to the individual, then the task itself acts as a reinforcer, and the individual will seek mastery for pleasure.

Social learning theory focuses more on the perception one has of how important a task is to society. In interacting with a social group, an individual will learn which tasks are most valued, will put more effort into those tasks, and will seek to become competent in those tasks through observation, engaging in the activity with already skilled models, and imitation.

Attribution theory describes how people credit the consequences of an outcome along three components: locus, whether the outcome is attributed to an internal or external cause; stability, whether the outcome will change or remain consistent over time; and control, whether one believes he or she can alter the outcome. Attributing the consequences to one of these three components has different effects on self-efficacy beliefs. For instance, students who attribute failure on a history test to an internal cause would blame their own ability for the poor grade. Because of the belief that one has low ability regarding the subject of history, self-efficacy beliefs concerning history will suffer. In contrast, attributing failure on a test to an external cause could result in the students blaming their studying habits in preparation for the test, an illness, or the teacher developing a difficult test. This would retain the student’s current self-efficacy beliefs on the subject of history because there were unique circumstances that prevented the student from succeeding.

Sources of Self-Efficacy

Albert Bandura, a major investigator in the area of self-efficacy, described four sources of self-efficacious ideology: mastery experiences, vicarious experiences, social persuasion, and physiological/affective states.

Mastery experiences are derived from the individual’s history of achievement and praise. Imagine a student who must answer a number of math questions in addition to finishing an art project for school. If the student is repeatedly successful at completing math problems while also being praised for good work, the student is likely to have more positive self-efficacy beliefs when completing math problems in the future. On the other hand, the same student would have poor self-efficacy beliefs if the student received no praise on his or her artwork while not performing to self-perceived satisfactory standards.

A vicarious experience comes into play when the teacher incorporates feedback and modeling into a lesson. This would come in the form of a teacher providing instant feedback regarding how the student is performing each step necessary to complete a task while including models who would demonstrate the task for the student.

Social persuasion strengthens self-efficacy beliefs through words of encouragement, meaningful expectations, or other social actions. Simply convincing someone performing a task that he or she is more or less capable of achieving a goal will affect his or her motivation to achieve.

Finally, physiological and affective states contribute to self-efficacy. Affective states describe emotions, such as joy or sadness, that affect motivation to complete a task, whereas physiological states describe the physical experiences that occur with affective states. Depending on the emotion, one can have more or less motivation. For example, when giving a speech, one may become anxious and experience distracting thoughts and uncontrolled worries along with a fast heartbeat, clammy palms, and a dry mouth. Experiencing both these affective and physiological states during the task may decrease the likelihood that one will choose adaptive coping strategies in order to halt the diminishing of one’s self-efficacy beliefs.

Correlates of Self-Efficacy

Self-efficacy is related to many performance-based outcomes in both the academic and professional worlds. Having either positive or negative views of one’s self-efficacy is indicative of how willing a person will persevere through a difficult task. An individual who has positive self-efficacious beliefs will believe one has more control over a situation and is more optimistic when thinking about the outcome of one’s work. If an individual believes that one is able to solve a problem, the individual will invest extra time and effort into the task, and the outcome is more likely to have favorable results. When individuals fail a task but manage to preserve their standards of self-efficacy, it is probable that the person will attempt to complete the task again.

Maintenance of positive self-efficacy beliefs gives an individual a sense of control over events that occur in life and encourages the individual to feel capable of handling more tasks. This maintenance can produce positive life outcomes. Conversely, individuals who believe that they cannot succeed harbors negative emotions that decrease the possibility of accomplishment. These negative self-efficacy beliefs function as sources of frustration that can adversely impact an individual’s mental health.

Vulnerability to stress is subject to increase depending on one’s self-efficacy beliefs and how important a task is perceived to be by the individual. If individuals believe that one cannot accomplish a difficult task that could potentially impact their life, then the individual will evaluate the task as a threat as opposed to a challenge and feel stressed. If one does not learn how to properly cope with the stress a task produces, self-efficacy beliefs will diminish and the individual is less likely to persist on the task in the future.

Shane D. Blair and Patricia A. Lowe

See also Anxiety; Attribution Theory; Behaviorism; Motivation; Reinforcement; Social Learning

Further Readings

Bandura, A. (1982). Self-efficacy mechanism in human agency. American Psychologist, 37, 122–147.

Bandura, A. (1997). Self-efficacy: The exercise of control. New York, NY: W. H. Freeman.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self efficacy beliefs to academic outcomes: A meta-analytic investigation. Journal of Counseling Psychology, 38, 30–38.

Gregory L. Callan Gregory L. Callan Callan, Gregory L.

Self-Regulation

Self-Regulation

Self-Regulation

The term *self-regulation* is used by many professionals across a diverse set of fields, such as psychology, education, athletics, and musicianship, to describe a number of related yet distinct phenomena. Moreover, there are several subfields that address self-regulation of behaviors, cognitions, motivation, or emotions. Although self-regulation may have many connotations, a broad theme is that it refers to adaptation to one's environment. Usually, this adaptation is viewed as a cyclical feedback loop in which one identifies a need, selects actions to address that need, acts, and then evaluates the effectiveness of the selected actions. In addition, self-regulation implies that the individual plays an active role and is the primary agent of change in this cyclical loop as opposed to external forces or persons. Relatedly, because the individual is the primary source of change, self-regulation is usually considered to be a means to attain goals that are valued by that person.

There are several subfields of self-regulation including behavioral self-regulation, emotional self-regulation, and self-regulated learning (SRL). Behavioral self-regulation refers most specifically to aspects such as staying seated, completing work, or waiting one's turn to speak. In contrast, emotional self-regulation deals primarily with the management of feelings and subsequent behaviors. The majority of this entry focuses on SRL, which deals with the application of self-regulation to learning.

Development and Theories of Self-Regulation

Although the concept of self-regulation has roots in many fields and with many theorists, Albert Bandura is often recognized as a primary contributor to the early theoretical basis. For example, Bandura proposed a model of self-regulation consisting of three primary processes: self-monitoring, self-judgment, and self-reaction. Since that time, a number of theorists have expanded this model to incorporate additional psychological constructs and processes. Although there are many models, one prominent model within the field of SRL has been proposed by Barry Zimmerman. This three-phase model suggests that SRL entails three interrelated phases of forethought, performance, and self-reflection, which respectively describe what an individual does before a task (i.e., forethought), during a task (i.e., performance), and after a task (i.e., self-reflection). Some researchers within the field of emotional self-regulation have also adopted this three-phase model of self-regulation. Rather than comprehensively review all models of self-regulation, this entry describes the three-phase model in detail. The Further Readings provide resources for the other prominent models.

Three-Phase Model of SRL

During the forethought phase, regulated individuals tend to set goals and create plans of actions to facilitate the attainment of those goals. Setting goals is important because it focuses one’s energy toward an important outcome and serves as a benchmark against which performances can be evaluated. On the other hand, planning facilitates the selection of efficient and effective ways to complete a task and bolsters motivation. Within the three-phase model, motivation is perceived to be important during the forethought phase because setting goals and planning require effort that individuals may not choose to expend if they are not motivated. Some prominent motivational variables that have been linked to self-regulation include beliefs in one’s capability to succeed (i.e., self-efficacy), interest in the target task (i.e., task-interest), or personal value of the task to an individual (i.e., value). Setting goals, planning, and having adequate motivation support the next phase: performance.

The performance phase of SRL describes what an individual may do during actual engagement in the task. At this time, there are several regulatory processes that are believed to be important. The first is self-control, which refers to how an individual manages the demands of a task. For example, when a person is taking a test or learning a difficult musical piece, the person may need to manage cognitive resources, emotions, or motivation to optimize performance. If a musician does not manage his emotions and becomes very anxious, the musician may begin to play too quickly or fumble through the notes. In contrast, students completing mathematical word problems may need to manage their cognitive resources to optimize performance. To do this, students may use cognitive strategies (e.g., draw a picture) to help them succeed.

Although learners may also receive feedback from others such as teachers, parents, or coaches, these individuals are not always available and do not always provide useful feedback. Another important process within the performance phase is self-observation, which describes two means for collecting information. Frist, recording or keeping track of (possibly by graphing) one’s attainment or performance (i.e., self-recording). The second process entails individuals’ awareness of how well they are currently performing (i.e., self-monitoring). Self-observation is crucial because it is an internal source of data that can be compared against one’s goal and may drive future adaptations within the third phase of SRL. Self-observation is the metaphorical thermometer in a furnace. When the air temperature does not match the “set temperature,” the furnace is engaged to raise or lower the temperature. Without this thermometer, the furnace would not be able to regulate the temperature without external intervention.

The self-reflection phase occurs following task completion and after performance feedback is received. During self-reflection, regulated individuals determine whether the goal was satisfactorily met (i.e., satisfaction), why the goal was or was not met (i.e., attributions), and what needs to change for the next performance (i.e., adaptive inferences). Self-reflection is crucial because it bridges prior learning to future learning experiences. In the absence of self-reflection, learners may make the same mistakes repeatedly and thus stagnate in their skill development or learning.

Gregory L. Callan

See also Metacognition; Motivation; Social Cognitive Theory

Further Readings

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 451–502). San Diego, CA: Academic Press.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), Metacognition in educational theory and practice (pp. 279–306). Hillsdale, NJ: Erlbaum.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation: Theory, research, and applications (pp. 13–39). San Diego, CA: Academic Press.

Ludmila N. Praslova Ludmila N. Praslova Praslova, Ludmila N.

Self-Report Inventories Self-Report inventories

1498

1503

Self-Report Inventories

The term *self-report inventory* typically refers to a structured set of written questions, which are formatted in a consistent way and ask respondents to evaluate themselves in yes/no, true/false, or a rating scale format. Occasionally, inventories might include open-ended questions. Critics of self-report inventories often challenge their validity. Nevertheless, self-report inventory methodology is widely used in educational, diagnostic, organizational, and other relevant contexts. Examples of well-known inventories are psychological diagnostic instruments such as the Minnesota Multiphasic Personality Inventory (MMPI) or the Beck Depression Inventory, learning style assessment instruments such as Kolb’s Learning Style Inventory, and workplace instruments such as the Maslach Burnout Inventory. This entry describes self-report inventories and their common applications, advantages and disadvantages of their use for research and assessment, and methodological recommendations for minimizing disadvantages. The entry concludes with a list of resources relevant to appropriate use of self-report inventories.

Self-Report Inventories

Research and individual assessment methods that ask individuals to provide information about themselves are known as self-report methods. Self-report (including oral interviews and written questionnaires) is the most commonly used methodology in social sciences. The term *inventory* is occasionally used interchangeably with *questionnaire* and *survey*, but the more narrow definition describes an inventory as a type of questionnaire in which questions are presented in a consistent format and respondents are typically asked to respond in a yes/no, true/false, or rating scale format. Self-report inventories are popular among both researchers and practitioners. This methodology has strong roots in clinical and personality psychology, with MMPI (later in revision, MMPI-2, as well as MMPI-A for adolescents) in use since 1943 but is also very popular in educational applications, such as learning styles and learning differences assessment, and organizational applications, such as personnel selection, leadership development, and organizational development. The use of normal personality self-report inventories, such as the NEO Personality Inventory-Revised, spans multiple contexts.

Advantages of Self-Report Inventories

Popularity of self-report inventories is due to several characteristic researchers and practitioners find appealing. Advantages of self-report-inventories include advantages of self-report and advantages of written inventory format.

Advantages of Self-Report

Self-report provides direct insight into respondents' point of view and subjective experience. Many have argued that learning about people is not possible without tapping into their own views and perceptions directly. For example, the subjective experience of individuals dealing with math anxiety, social anxiety, depression, or job burnout is truly only accessible to these individuals.

Advantages of Inventories

Subjective experience could be communicated in unstructured ways. However, using a structured format, such as that of an inventory, allows researchers to some extent to quantify individuals' subjective experience and compare the reported intensity of the experience to normative data. Although structured interviews could also allow quantifying and comparing of subjective experience, written inventory format has several advantages over structured interviews. Privacy and confidentiality can be very important to participants, and inventories can be filled out privately, which can in turn help to ensure confidentiality. An additional advantage is that participants may disclose information they might not have disclosed in an interview.

Finally, practicality and cost are important advantages of self-report inventories.

Inventories are useful for addressing many research questions in a practical and efficient manner and allow for research on topics not accessible to observation and experimentation. Practitioners may also find self-report inventories to be convenient and effective. Compared to interviews, inventories are much cheaper to administer and they can be administered to a large number of individuals simultaneously. With the transition of some inventories from paper to digital format, advantages of paperless administration and instant scoring have further contributed to the practicality of inventories. Such practicality, coupled with the advantage of tapping into subjective experience in ways that observation or experimental manipulation could never do, makes self-report inventory a preferred method of many researchers and practitioners.

Disadvantages of Self-Report Inventories

Despite their popularity, self-report inventories have many critics who point out multiple potential validity threats. Similar to advantages, disadvantages can be seen as specifically stemming from the nature of self-report and as stemming from the inventory format.

Disadvantages of Self-Report

Self-reports have several disadvantages that present potential threats to validity of individual scores or group data. Sometimes these threats stem from motivated behavior of participants; other times they are due to automatic processes involved in human cognition.

Constraints of Self-Knowledge

Even if participants are motivated to be honest, they may lack self-knowledge and introspective ability, and their self-perceptions might be very different from the objective reality. Therefore, self-report information may be “incorrect” despite participants’ desire to be accurate. From the perspective of social cognition, people often do not know what influences their behavior, how it compares to behaviors of others, or even how often they engage in specific behaviors, especially undesirable ones. For example, it is well-documented that individuals are in general implicitly motivated to see themselves in a positive light, take credit for success and deny responsibility for failure, and see themselves as “above average.” Thus, honest self-reporting may result in scores that are unrealistically high. However, there are exceptions to this general positivity bias—moderately depressed individuals tend to see themselves more accurately than either nondepressed individuals, who are unrealistically positive, or severely depressed individuals, who are unrealistically negative. In addition to self-serving and self-enhancing biases, self-knowledge can be constrained by defensiveness or denial, which may operate outside of awareness. Constraints on self-knowledge might be especially pronounced for specific populations, such as children, adolescents, and certain clinical and subclinical populations (i.e., individuals affected by psychiatric disorders or certain personality tendencies).

Additional threat to validity of self-reports in educational settings is the limited ability of individuals to self-report their own levels of skills and knowledge. Specifically, the Dunning–Kruger effect, or "incompetent and unaware of it" effect, is the tendency of individuals, especially those low in competence, to significantly overestimate their level of competence. In general, more competent individuals tend to be more accurate or modest in their self-evaluation.

Intentional Self-Presentation

Self-report can be significantly influenced by various forms of impression management, including exaggeration, faking, and lying. Individuals can be motivated to engage in socially desirable responding or "faking good" by trying to present themselves as more conscientious, capable, or culturally sensitive in order to fit in or to get a reward. Negative self-presentation, also referred to as "faking bad" or "malingering," may be driven by the desire to present oneself as less emotionally stable, competent, or aware in order to avoid responsibility or obtain help.

Disadvantages of Inventory Format

In addition to limitations of self-report in general, the format of self-report inventories is also associated with several threats to validity.

Limitations of Understanding

Understanding written questions and rating scales requires foundational skills of reading comprehension and basic anchoring of responses. In educational contexts, understanding can be limited by skill development in children and adolescents as well as by learning differences or learning disabilities.

adolescents as well as by learning differences or learning disabilities.

Acquiescent Responding and Reactant Responding

Acquiescent responding, or acquiescence, is the tendency to agree with statements without regard to their content. The opposite tendency, indiscriminant disagreement, is referred to as reactant responding. Dichotomous response formats (yes/no, true/false) may produce especially dramatic differences in the proportion of “yes” or “no” answers selected by participants. Researchers disagree on how practically important are the effects of acquiescent and reactant responding. Some believe that effects are negligible, whereas others consider potential effects highly concerning. Threat to validity could be especially pronounced in situations in which it is difficult to determine whether answers are due to responding bias or to actual differences in the construct being measured. For example, if someone reports high anxiety in the presence of others, in testing situations, and in solving mathematical problems, is this a case of a highly anxious individual or acquiescent responding?

Extreme Responding and Midpoint Responding

Extreme responding is the tendency to disproportionately endorse the extreme rating scale choices (e.g., 1’s and 5’s on a 5-point scale). Although extreme responding may create validity threat, the low end of the extreme responding continuum, the tendency to use the midpoint, is also problematic. In some cases, individual differences play a role in extreme responding across time and instruments. Situational factors (ambiguity of the situation, stress, mood, rapid responding, and providing responses whether one is motivated to do so or not) may also play a role in extreme responding. In educational settings, student ratings of instruction often reflect “love” (high), “hate” (low), or “meh” (midpoint) ratings across all items, which contributes to validity concerns with the use of such ratings.

Addressing Disadvantages of Self-Report Inventories

Although limitations of self-report inventories are important to consider, most researchers and practitioners agree these inventories are still useful. Several approaches have been developed to address and ameliorate disadvantages of self-report inventories. Careful inventory design, supplementing self-report with additional data sources, and thoughtful analysis and interpretation of data facilitate effective use of self-report inventories.

facilitate effective use of self-report inventories.

Validity Threats Resulting From Self-Knowledge Limitations

Limitations of self-knowledge can be ameliorated in several ways. In some cases, self-report inventories can be supplemented by using reports of others—parents, teachers, colleagues—in evaluating the individual. Although reports of others may not necessarily reflect the ultimate objective reality either, comparisons between multiple data points provide rich information for analysis and further consideration. In some cases, tapping into automatic responding is a valuable supplement to self-report. For example, in developing intercultural awareness, supplementing scores on self-report inventories with obtaining and discussing scores on implicit attitudes tests designed to tap into automatic processing may enrich self-understanding and understanding of implicit, automatic biases. Finally, in addressing the Dunning–Kruger effect, training in relevant skills typically improves one’s ability to evaluate one's own competence.

Validity Threats Resulting From Intentional Self-Presentation Responding

Validity threats stemming from intentional self-presentation can be addressed in part by the same safeguards as those helpful in addressing limitations of self-knowledge. There are additional methodologies that address self-presentation concerns specifically.

Intentional self-presentation responding, such as socially desirable responding or malingering, can be driven by both situational factors (e.g., high-stakes decisions based on scores) and personality differences/dispositional propensities toward impression management. To capture this trait-like form of responding, some inventories include “lie,” “faking,” or “malingering” subscales. Although somewhat effective, these scales can also produce false-positive or false-negative results. Some respondents are able to cheat those scales, whereas others might be misclassified due to cultural response sets, such as modesty bias documented in Asian American populations. This, obtaining information from multiple sources, as well as careful interpretation of data, provides validity support above and beyond the use of responding–measuring subscales support above and beyond the use of responding "measuring subscales.

Another way to ameliorate issues related to intentional self-presentation is to use items that do not have an obvious “correct” response. Although participants may feel upset by the apparent lack of face validity, and that it is difficult to write items that appear neutral in social desirability yet provide valid and reliable measurement, using such items provides significant benefits.

In addition to addressing participant’s ability to respond in ways influenced by impression management, sometimes it is helpful to reduce their motivation to do so. Motivation to respond in a socially desirable manner can be reduced by maximizing the anonymity and confidentiality of respondents.

Validity Threats Resulting From Limitations of Understanding

In some cases, limitations of understanding can be addressed by providing some training on how to use rating scales and by ensuring that the language of the inventory is appropriate to the reading level of respondents. For example, inventories or versions of inventories can be developed specifically for children. In other cases, using reports of others—such as parents or teachers—might be most helpful.

Validity Threats Resulting From Acquiescent and Reactant Responding

Validity threats stemming from acquiescent and reactant responding are typically addressed by inventory design in which half the items are written as true key (a high rating indicates a high level of the characteristic being measured) and half the items are false key (also known as reverse scored items; high rating indicates a low level of the characteristic being measured). However, this also leads to reduction in the α reliability of the instrument, and the tendency of factor analyses to produce two factors—one for the true-keyed items and one for the false-keyed items due to the tendency of true and false keyed items to correlate with each other.

Validity Threats Resulting From Extreme and Midpoint Responding

Midpoint Responding

Extreme responding and midpoint responding bias cannot be addressed just by balancing the key or introducing reversed-scoring items because extremity works in both directions and midpoint. In some situations, using dichotomous formats (true/false, yes/no) is recommended because "yes" and "no" responses are equally extreme. Another approach is using fixed/forced distributions. For example, the respondent may be asked to rank response options from most to least preferred using the entire range of options. Specific variation of fixed distribution is Q-sort, which requires ranking responses relative to other responses in the set. Midpoint responding can also be addressed by a "softer" form of forcing the answer via eliminating the "middle" option on the scale and using an even number of scale points (e.g., eliminating the *neither agree nor disagree* option).

In those cases in which extreme and midpoint responding are caused by participant lack of motivation, designing inventories that are as short as possible and as easy to use as possible while still providing valid data might be the most appropriate solution. In addition, making the assessment situation as appealing as possible, providing explanation of the importance of one's responses, and avoiding "survey fatigue" by limiting the amount of information individuals are asked to provide also tend to improve participant motivation. This also helps to somewhat alleviate random or careless responding.

Self-response inventories, as all forms of measurement, are not perfect. However, they are and will likely remain popular due to their practicality. Carefully designing items and scales helps alleviate threats to validity. Scales designed to measure systematic response biases along with statistical techniques for separating out extraneous variance are also used to improve accuracy of self-report. It is also often recommended to supplement data from self-report inventories with data from other sources and to triangulate findings using multiple methods.

Future Directions

Self-report inventories are becoming increasingly used in digital, often Internet-based formats. Technology allows researchers and practitioners to quickly receive inventory scores as well as a wealth of comparative data. Advances in computer-adaptive methodology should help further refine self-report inventories and help develop shorter, yet valid and reliable inventories. At the same time, increasing use of web-based inventories creates new threats, such as potential hacking or leaking of scoring keys. In addition, proliferation of poor quality, easily accessible “inventories” distributed via social media and multiple websites contributes to survey fatigue, misinformation, and confusion. This may call for increasing measurement literacy, such as understanding of validity and the importance of using high-quality instruments, especially for diagnostic purposes, among the general population.

Ludmila N. Praslova

See also Reliability; Scales; Social Desirability; Survey Methods; Triangulation; Validity

Further Readings

Costa, P. T., & McCrae, R. R. (1989). Manual for the NEO personality inventory: Five factor inventory/NEO-FFI. Odessa, FL: PAR.

Donaldson, S. I., & Grant-Vallone, E. (2002). Understanding self-report bias in organizational behavior research. Journal of Business and Psychology, 17(2), 245.

Kolb, D. A., Boyatzis, R., & Mainemelis, C. (2001). Experiential learning theory: Previous research and new directions. In R. Sternberg & L. Zhang, (Eds.) Perspectives on cognitive learning, and thinking styles. Mahwah, NJ: Lawrence Erlbaum Associates.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. doi:10.1037/0022–3514.77.6.1121

Maslach, C., Jackson, S. E., & Leiter, M. P. (1996). Maslach burnout inventory (3rd ed.). Palo Alto, CA: Consulting Psychologists.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 224–239). New York, NY: Guilford.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. Journal of Personality and Social Psychology, 78, 582–593.

Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys. New York, NY: Academic Press.

Benjamin D. Rosenberg Benjamin D. Rosenberg Rosenberg, Benjamin D.

Mario A. Navarro Mario A. Navarro Navarro, Mario A.

Semantic Differential Scaling Semantic differential scaling

1503

1507

Semantic Differential Scaling

Pioneered by Charles Osgood in 1952, semantic differential scales are a popular technique for measuring people’s attitudes toward nearly anything. Semantic differential scales use a standardized set of bipolar adjectives (see Figure 1) on which research participants rate an issue or object. This simple procedure confers a variety of benefits, both for researchers and study participants.

Through a series of statistical analyses, Osgood identified three recurring, stable dimensions on which people can judge nearly anything (see Table 1): (a) *evaluative*, focused on the value of the object (e.g., good/bad); (2) *potency* or power of an object (e.g., strong/weak); and (3) *activity* or movement of an object (e.g., slow/fast). To use a semantic differential scale, research participants respond to several bipolar adjectives designed to measure each dimension of a specific object or issue by placing a mark on one of the seven blanks between the two adjectives.

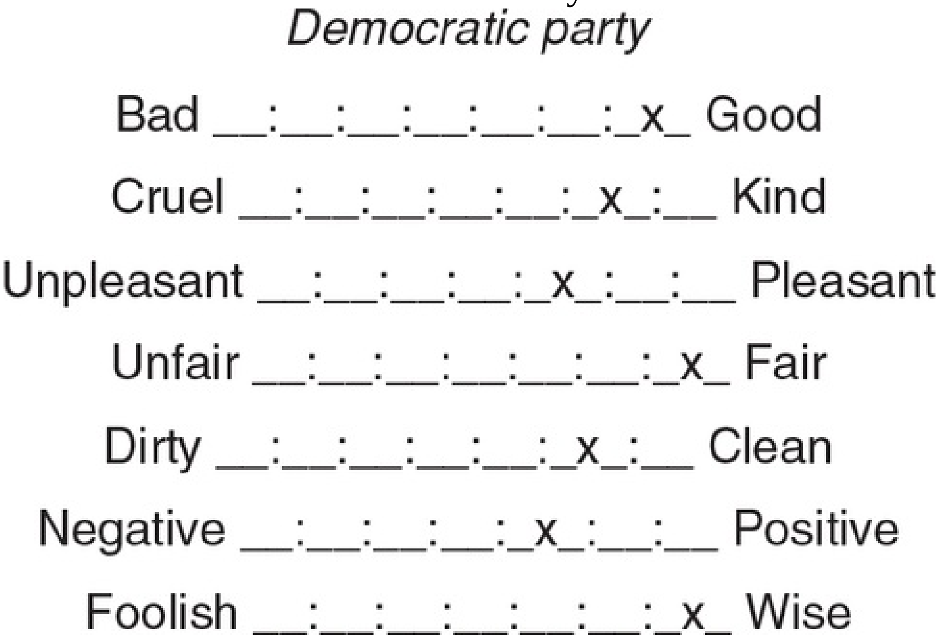
This entry provides a discussion on the creation and history of semantic differentials, their current uses, some of the pros and cons of use, and future directions for semantic differential research.

A Brief History of Semantic Differential Scaling

Initial Research

As psychology developed as a science, researchers began to use rating scales to measure people’s attitudes and beliefs. Thus, to place Osgood’s development of the semantic differential in historical context, it is instructive to briefly consider other prominent measurement techniques that were proposed in the 1920s to 1940s. Most notably, the Thurstone, Guttman, and Likert methods assess people’s level of agreement with a series of unique questions about a particular topic. For example, on a Likert-type scale, research participants could be required to answer numerous, differently worded questions about various aspects of the Democratic Party on a *strongly disagree* to *strongly agree* scale.

Figure 1 Example of a series of semantic differential items used to evaluate attitudes toward the Democratic Party



Democratic party

Bad __:__:__:__:__:__:_x_ Good

Cruel __:__:__:__:__:_x_:__ Kind

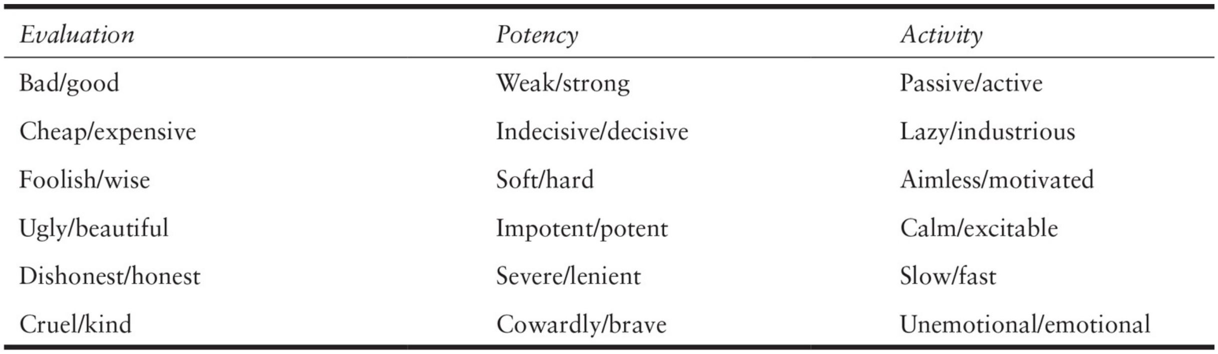
Unpleasant __:__:__:__:_x_:__:__ Pleasant

Unfair __:__:__:__:__:__:_x_ Fair

Dirty __:__:__:__:__:_x_:__ Clean

Negative __:__:__:__:_x_:__:__ Positive

Foolish __:__:__:__:__:__:_x_ Wise



Evaluation	*Potency*	*Activity*
Bad/good	Weak/strong	Passive/active
Cheap/expensive	Indecisive/decisive	Lazy/industrious
Foolish/wise	Soft/hard	Aimless/motivated
Ugly/beautiful	Impotent/potent	Calm/excitable
Dishonest/honest	Severe/lenient	Slow/fast
Cruel/kind	Cowardly/brave	Unemotional/emotional

In the early 1950s, Osgood and several other scholars proposed the semantic differential as a technique for measuring the meaning that people place on particular concepts. In direct contrast to the contemporary measurement techniques mentioned earlier, Osgood’s semantic differential relied on a standardized set of items, on which participants can rate nearly anything. Thus, a standardized set of items, on which participants can rate nearly anything. Thus, a key benefit of semantic differentials is that they are less taxing on research participants and researchers than other methods.

Several early investigations found that semantic differentials could efficiently capture the changing nature of social stereotypes. These initial studies also revealed that people's judgments fall into the three dimensions noted earlier—evaluative, potency, and activity. Moreover, the findings from Osgood's work provided initial evidence that semantic differential scales were relatively objective, reliable, and valid ways of measuring a wide range of concepts.

Updates and Current Uses

A key result of early research on semantic differential scaling was that compared to the potency and activity dimensions, evaluative questions revealed the most about people's overall assessment of an object or person. A great deal of research confirms this result, as evidenced by the publication of over 500 studies, on a range of topics, from all over the globe. Thus, for researchers interested in people's attitudes, of which evaluation is a major component, using semantic differentials is a key measurement tool.

Attitude researchers often take the sum or average of people's responses to a set of semantic differential items to get an idea of their overall attitude toward a topic. For instance, on the scale in Figure 1, scores of 7, 6, 5, 7, 6, 5, 7 result in a sum of 43 or an average of 6.1—both of which indicate quite positive attitudes toward the Democratic Party.

Creating a Semantic Differential Scale

Selecting Adjective Pairs

Some scholars continue to use questions aligned with Osgood's original evaluative component to assess attitudes toward a range of topics, from political issues to workplace policies to advertisements. Yet other researchers have found that Osgood's adjective pairs do not fit every topic and instead choose their own adjective pairs.

Thus, for these scholars, a central consideration is the way in which they select the specific pairs of adjectives in a given study. Even though it may be tempting for researchers to choose adjective pairs subjectively, the method of selecting adjectives has implications for the quality of data collected. As such, some scholars have suggested following a four-step method for creating the semantic differential scale.

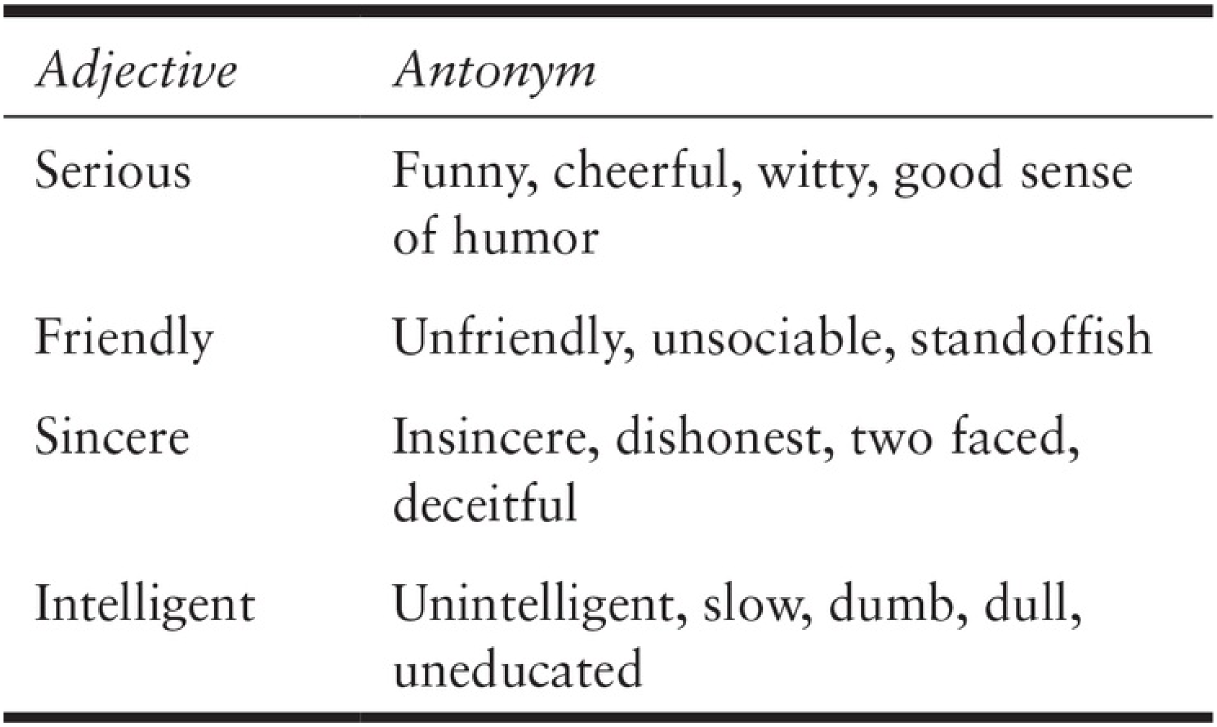
In Step 1, researchers ask a group of participants to provide descriptive adjectives for the concept or set of concepts in which they are interested. For example, in studying people’s attitudes toward the Democratic Party, researchers in this step might ask participants to come up with adjectives to describe concepts like *democrat, liberal, progressive,* and *prochoice*. Then, in Step 2, researchers would use the list of adjectives that the sample group produced to create a prototype semantic differential scale; this scale would then be tested on a separate sample group in Step 3. Finally, researchers would subject people’s responses to the prototype scale to statistical analyses, the results of which would be used to form the final semantic differential.

A second, somewhat less systematic procedure for selecting adjective pairs is to rely on those used in prior studies in the same topic area or on those from Osgood’s original research. For instance, researchers could examine previously published work to see what adjective pairs other scholars have used to measure attitudes toward the Democratic Party; they could then use a similar set of words in their study. A potential benefit of using these previously vetted sets is that the word pairs are more likely to represent each aspect of judgment (i.e., evaluative, potency, and activity).

The examples provided in Table 1 do not cover every possible adjective pair but various configurations that could be used to measure a wide range of attitudes.

Selecting Antonyms

A consideration in the process of selecting adjective pairs concerns the opposite, or negative, end of the scale (see Table 2). More specifically, there are a few options for listing the antonyms of the adjectives determined using the aforementioned process. For many adjectives, researchers can use a complementary opposite at the other end of the semantic. This process is particularly easy when a negative morpheme can be added to the beginning of the adjective (e.g., *successful—unsuccessful, honest—dishonest, patient—impatient*).



Adjective	*Antonym*
Serious	Funny, cheerful, witty, good sense of humor
Friendly	Unfriendly, unsociable, standoffish
Sincere	Insincere, dishonest, two faced, deceitful
Intelligent	Unintelligent, slow, dumb, dull, uneducated

As another option, researchers can simply add *not* to the adjective (e.g., *rich—not rich*, *generous—not generous*); the benefit of this approach is that it works with just about any descriptor. This technique is less satisfactory, however, because just adding *not* does not necessarily connote the exact opposite of an adjective. Case in point, *not generous* is not the opposite of *generous*—indeed, it does not necessarily mean *stingy*. In cases like these, where adding a negative morpheme (e.g., *un-*, *dis-*) or *not* does not produce a complementary antonym, researchers can select a gradable antonym. Using *stingy* as an antonym for *generous* is probably the easiest for participants to understand and will likely produce the best data for researchers.

Some adjectives have many possible antonyms, so a difficult task for researchers can be to select the most appropriate one. In certain cases, it is better to use complementary antonyms, while in others, it is better to use gradable ones; thus, researchers often combine each type in a semantic differential scale.

Number of Adjective Pairs Used in a Scale

Semantic differential scales most frequently contain between 8 and 12 adjective pairs, but this is by no means a steadfast rule. Researchers may use as few as 4, and as many as 20, adjective pairs to assess the same concept. As a general rule, researchers must balance comprehensiveness (i.e., measuring every component of the attitude object) and practicality (i.e., asking participants a reasonable number of questions). The way in which researchers weight comprehensiveness and practicality depends on the aims of a particular study and the method that previous researchers used.

Using any of these procedures to form a semantic differential scale does not guarantee that it will be a valid measurement tool for a given study. However, these methods represent a vast improvement over researchers selecting items haphazardly.

Additional Considerations

In addition to the specific method used to select adjectives, researchers must consider the relevance of the adjectives to the *group* and *topic* that they are studying; they should think about people’s general positivity or negativity toward the topic.

Relevance of the Scale

First, researchers must ensure that the group of study can easily understand the scale they create (or adapt from previous work). It is also important to ensure that the selected adjectives are applicable to the group of participants. For example, the adjective pair *religious—nonreligious* may be more appropriate in describing everyday life in Egypt than it is in Australia.

A related consideration is whether the adjectives are relevant to the topic of study. Asking people to rate the *religiosity* of the Democratic Party may be appropriate, whereas asking the same question about the field of psychology less appropriate.

Cultural Effects on Valence

Researchers also should consider whether people negatively or positively value the characteristic about which they are asking. Based on the culture or context of a study, people may imbue the same attitude object with positive or negative characteristics. For instance, research has found that people in Western cultures perceive words like *ambition* and *self-confidence* positively, whereas people in

Japan perceive them negatively.

Formatting

Semantic differential scales are most often presented in a manner similar to Figure 1. However, when creating one, researchers must make decisions about the specific format that they will use.

Polarity

In constructing a semantic differential scale, a core question is whether researchers should array the positive adjectives consistently on the same side (e.g., all on the right) or if they should randomize them (e.g., on the right, then left, then right). As noted earlier, one consideration is whether there is consensus regarding an adjective's positivity or negativity. If there is known debate over the polarity of a word (or words)—like in the *ambition* example—then it could be best to randomize the sides on which negative and positive words appear. Conversely, when adjective pairs have clear negative and positive words (e.g., *bad—good*, *unkind—kind*), research indicates that it is best to consistently array them in the scale. Always listing negative adjectives on the left and positive adjectives on the right helps respondents make easier judgments and is less mentally taxing.

Number of Scale Points

In Osgood's original conception, all semantic differential scales had seven blanks with which people could judge a person or object. In the years since, researchers have used 5-, 6-, and 9-point scales with varying degrees of success. Using a greater number of scale points allows people to make more fine-grained judgments, but the trade-off is that the differences between too many scale points may become meaningless. For instance, in a 9-point scale, participants may have difficulty choosing between a 7 and 8.

A related consideration is whether the scale should contain an even or odd number of blanks, the main consequence of which is the inclusion or exclusion of a "neutral" option in the middle of the scale. The key benefit of including a neutral option is that people are frequently neither negative nor positive toward a person or object; in this way, a neutral option can accurately reflect their evaluation. On the other hand, selecting neutral can also mean that people are undecided or do not have enough information to make a judgment. Without follow-up, researchers are unable to determine the truth. Thus, there could be circumstances in which *not* including a neutral option, and forcing a choice is preferable (although this may inject error into measurement procedures). In general, researchers most frequently use 7-point scales.

Osgood's semantic differential scaling technique, which uses a series of bipolar adjectives to measure people's judgments toward a range of stimuli, offers a simple and accurate means of data collection. Investigators interested in studying attitudes would be wise to consider employing semantic differential scales in their research.

Benjamin D. Rosenberg and Mario A. Navarro

See also Attitude Scaling; Instrumentation; Rating Scales; Self-Report Inventories; Survey Methods; Surveys

Further Readings

Banaji, M. R., & Heiphetz, L. (2010). Attitudes. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), Handbook of social psychology. Hoboken, NJ: Wiley.

Brinton, J. E. (1961). Deriving an attitude scale from semantic differential data. The Public Opinion Quarterly, 25, 289–295.

Crano, W. D., Brewer, M. B., & Lac, A. (2014). Principles and methods of social research (3rd ed.). New York, NY: Routledge.

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scoring of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. Personality and Individual Differences, 40, 873–884.

Garland, R. (1990). A comparison of three forms of the semantic differential. Marketing Bulletin, 1, 19–24.

Krosnick, J. A, Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), Handbook of attitudes and attitude change. Mahwah, NJ: Erlbaum.

Osgood, C. E. (1952). The nature and measurement of meaning. Psychological Bulletin, 49, 197–237.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

Snider, J. G., & Osgood, C. E. (Eds.). (1969). Semantic differential technique: A sourcebook. Chicago, IL: Aldine.

Semi-Partial Correlations

***See* Part Correlations**

Samantha B. Goldstein Samantha B. Goldstein Goldstein, Samantha B.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Sensitivity

Sensitivity

1507

1509

Sensitivity

Sensitivity, also called *true positive rate,* measures a diagnostic test’s ability to detect the correct number of positive elements in a binary classification test or to diagnose the correct number of students who have a given condition. Imagine, for example, a preliminary diagnostic test that determines whether or not a student has attention-deficit/hyperactivity disorder. The sensitivity of the test would measure the number of students who are correctly diagnosed with the condition. This entry further defines sensitivity, discusses the receiver operator characteristic (ROC) and positive predictive value (PPV), and looks at the practical application of sensitivity in testing.

Sensitivity assists in avoiding false negatives or classifying something as negative when it is in fact positive (e.g., identifying students as typically functioning when they have a disability). This is also known as Type II error. Sensitivity is calculated by the following formula:

$$\begin{aligned}\text{Sensitivity} &= \frac{(\text{true positives})}{(\text{true positives})+(\text{false negatives})}\\ &= \text{probability of a positive outcome given}\\ &\qquad \text{that the student has the condition.}\end{aligned}$$

Sensitivity is used alongside specificity when determining the efficacy of a binary classification test also called a binomial classification test. This kind of test divides elements of a group into two classes based on a certain characteristic. Diagnostic tests are a primary example of binary classification. They determine whether or not a child has a given condition, dividing students into either the typically functioning or condition group. Sensitivity measures how accurately the test predicts positive results, and specificity measures how accurately it predicts negative results. Specificity, also called *true negative rate*, measures a binary classification test's ability to detect the number of negative elements that are classified as negative or a diagnostic test's ability to identify the correct number of students who are normally functioning. In the example of the attention-deficit/hyperactivity disorder test, specificity would determine the proportion of students who do not have attention-deficit/hyperactivity disorder and are diagnosed as such by the test.

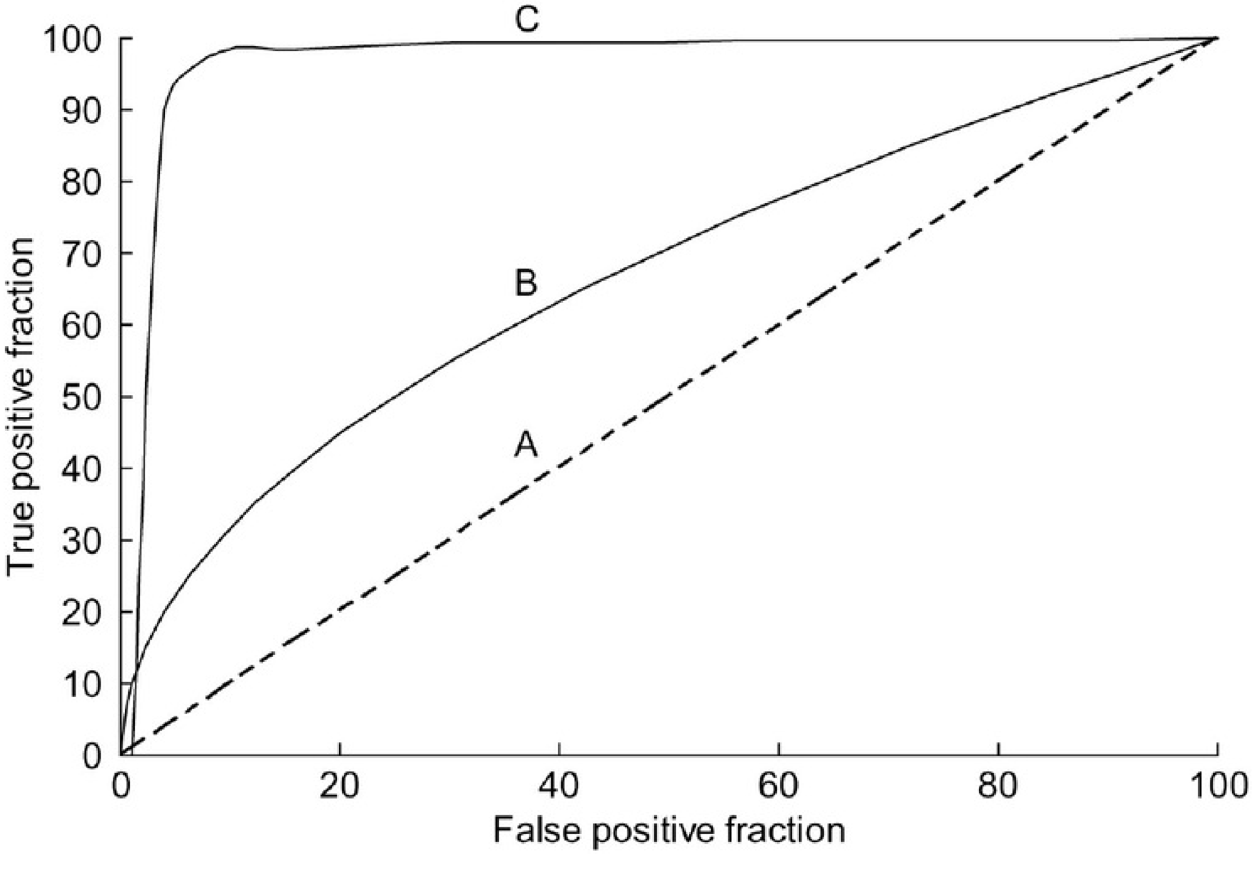
A perfect predictor would be 100% sensitive and 100% specific. A test with 100% sensitivity can correctly identify all students with a given condition, although that is unlikely. More commonly, a high, but not perfect, sensitivity score is useful in ruling out a condition when a student tests negative. For example, using a basic cognitive test to determine whether a student has a learning disorder would have high sensitivity because a high proportion of students who have learning disorders would test positive. However, imagine that this test is just the first of many tests in diagnosing a learning disorder. In that case, this test would not be very specific: A high proportion of students who do not have the learning disorder may also test positive and would be found to have no disorder with subsequent testing. A rule of thumb to go by is SnNOut (if the result of a highly sensitive [Sn] test is negative [N], it rules out the condition) and SpPIn (if the result of a highly specific [Sp] test is positive [P], it rules in the condition).

ROC Curve

An ROC curve, as shown in Figure 1, is a graphical illustration of sensitivity and specificity. A sensitivity ROC curve plots the false positive rate of a diagnostic test on the *x*-axis, against sensitivity, or true positive rate, on the *y*-axis to display sensitivity as a function of false positives. The area under the curve represents the overall accuracy of a diagnostic, or binary classification, test. A value of 1.0 indicates 100% sensitivity.

ROC curves take into account the cutoff point of a given diagnostic test. For example, a teacher uses the number of words a child can read to determine whether or not the child should be further tested for a learning disorder. The cutoff point of this test would be the number of words that require further testing. Raising the cutoff point would result in a lower false positive rate but a higher false negative rate or low sensitivity but high specificity. Few children would require further testing, but more of those children might now have undiagnosed learning disorders. Thus, sensitivity and specificity are inversely proportional. As sensitivity increases, specificity decreases and vice versa.

Figure 1 Receiver operator characteristic curve



Source: Lalkhen and McCluskey (2008, p. 222), by permission of the British Journal of Anaesthesia and The Royal College of Anaesthetists.

PPV

PPV measures the proportion of students who are correctly diagnosed with a condition, just like sensitivity, except that PPV depends on the population in question and the prevalence of the condition. Consider an example: Testing a whole school with that same basic cognitive test for learning disorders could have low PPV because of the high number of false positives. However, testing a population of only students who show symptoms of learning disorders (poor grades, attentional difficulty, etc.) would result in higher PPV due to the higher chance of students who present with symptoms having a learning disorder. Holding all other factors constant, the PPV of a test will increase with increasing prevalence of a condition, and its negative predictive value will decrease.

Practical Application

It is important to know which tests to use to most accurately diagnose a student with a given condition. A test with low sensitivity might not result in any definite answers. The *likelihood ratio* of a test is used to analyze the efficacy of a test: How much more likely it is that a student who tests positive has the condition than a student who tests negative.

$$\text{Likelihood ratio} = \frac{\text{sentivity}}{1 - \text{specificity}}.$$

A student who tests positive from a 50% sensitivity test is no more likely to have the condition than a student who tests negative. However, a test with high sensitivity will provide clearer answers for a teacher. A test with 98% sensitivity is more likely to correctly diagnose a student, thus giving a teacher more reason to rule out a condition when presented with a negative outcome.

Samantha B. Goldstein and Marc H. Bornstein

See also Bayes’s Theorem; Classification; Cognitive Diagnosis; Diagnostic Tests; Learning Disabilities; Specificity; Type II Error

Further Readings

Altman, D. G., & Bland, M. J. (1994). Diagnostic tests 1: Sensitivity and specificity. British Medical Journal, 308, 1552.

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain, 8(6),

221–223.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology, 56(1), 45–50.

Florence Wu Florence Wu Wu, Florence

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Service-Learning Service-Learning

1509

1511

Service-Learning

Preparing graduates to take on their responsibilities as citizens in their communities is often seen as one of the central goals of education. Yet even within the context of education, community service has been perceived as a reaction to social problems and societal needs, with less attention to the application of knowledge in the process of offering community service. In recent decades, the concept of service-learning has emerged as a way of offering students opportunities to serve and to learn simultaneously. This entry first defines service-learning and differentiates it from other forms of service. It then details the components of service-learning and research findings on outcomes for students who participate.

Service-learning is understood as responding to human and community needs while promoting educational growth. Service is seen as a part of a pedagogical method through which theories and facts are learned. The service experience enables students to put into practice the theories and concepts learned in their courses and prepares them to actively engage in social inquiry and problem solving, thereby promoting substantive learning. Students continuously reflect on the concrete situations, challenges encountered, and experiences gained throughout the process of offering services to the community.

Educational goals are achieved through students applying, integrating, and evaluating knowledge of related disciplines and developing perspectives and analytical skills to obtain first-hand understanding of social ecology. In the process of service-learning, both the service providers (i.e., the students) and service recipients mutually benefit, with equal consideration of all individuals in the service relationships.

the service relationships.

Service-learning has been widely adopted as a pedagogy across various disciplines, educational levels, and university settings. However, the appropriateness of university contexts to implement service-learning has been a heated debate in the past 2 decades. Traditionally, the purpose of higher education is to advance the professional competence of university students and give them better career prospects. However, there is also concern within higher education about helping students become responsible citizens. With valuable resources and a traditional mission of addressing the needs of the community, universities are particularly well suited for the development of service-learning. By emphasizing the importance of serving the community, service-learning can enrich students' learning and reconnect their academic learning to the authentic needs of the community. Service-learning can be seen as an important mechanism in universities to bridge the gap between universities and the community.

Service-Learning Versus Other Forms of Service

There are several concepts related to service-learning, namely, internships, field education, volunteerism, and community service that have some similarities to service-learning but also can be differentiated from service-learning. To represent the distinctions among these concepts, each can be defined by the intended beneficiary of the service activity and its degree of emphasis on service and/or learning.

Internships, also known as field education, involve opportunities for students to sharpen their skills in a particular field through devoting their time and effort to organizations in that field. Internship programs are primarily intended to benefit the participating students and their service in the programs mainly focuses on their learning. The service provided through an internship could be paid or unpaid, and those in internships may serve in profit-making or nonprofit organizations. Although there is an intention to benefit the recipients of the service, internships are primarily focused on students' learning about the recipients' needs and the professional skills required in the field.

Volunteerism and community service share more similarities with service-learning. In volunteerism, the students provide services based on a sense of altruism, and the benefit they receive is primarily the pleasure derived from serving others. The continuity of the services often depends more on the serving others. The continuity of the services often depends more on the volunteers’ will than concrete prior planning. Volunteers might receive benefits and learn something from the service delivery, but their learning experiences are unintentional and spontaneous.

In community service, more structured service delivery and volunteer commitment are involved. In the context of education, there is also an expectation that students providing services have opportunities to better understand the causes of social problems and ways to deal with these problems more effectively. Formal intellectual discourse is involved and students are able to integrate the service activities with their academic knowledge.

Service-learning moves beyond the context of charity and can be differentiated from volunteer service in that there is an expectation of reciprocity between the service providers and the service recipients. The needs of community members determine the nature of the service provided as part of service-learning. In addition, an academic context has to be present in service-learning. Service-learning is premised on experiential education as the platform for intellectual, moral, and civic growth. In service-learning, students take an active part in organizing service experiences and collaborating with members of the community. The services are closely tied to curricular objectives, involve students’ reflections, and contain an evaluative component.

Components of Service-Learning

The idea of introducing community service as a means of instruction can be traced to John Dewey’s concept of experiential learning. Dewey himself never explicitly delineated service-based learning. However, the pedagogical goals and methods of service-learning share elements of Deweyan theory.

Rebecca L. Carver has stated that service-learning addresses the three crucial purposes of experiential education, which are (1) encouraging students to become more effective change agents, (2) developing students’ sense of responsibility and belonging to their own communities, and (3) nurturing different students’ competencies. First, students acquire ideas and theories of their related discipline and learn the facets of their community. Then, through delivering organized and planned services, students are both physically and emotionally engaged in the subject of study. This engagement offers the students opportunities to integrate knowledge and enlarge their problem-solving capacities. Students can thus learn how to react more intelligently to changing capacities. Students can thus learn how to react more intelligently to changing situations in other service settings.

The process of reflection is also a core component of service-learning. As service-learning is seen as experiential learning and it rests upon the cyclic process of action and reflection on that action, students' understanding is continuously modified with more experiences, thoughts, and information gained from service delivery. In the process of reflection, students discuss their knowledge, skills, and attitudinal changes and accomplishments from the services in connection with their coursework.

Janet Eyler and her colleagues have proposed the four Cs to facilitate effective reflection, which are *continuous*, *connected*, *challenging*, and *contextualized*. The reflection process has to be carried on *continuously* over time throughout the course of service-learning. The reflection prior to the commencement of service delivery assists students in surfacing assumptions and sensitizing them to explore the possible complexities in the service settings. The reflection during the service helps students to derive meaning from their unique service experiences. The postservice reflection serves as the evaluation of the experience and facilitates students to consolidate the insights drawn from the experience.

In the four Cs of reflection model, reflection has to *connect* the service experiences to intellectual and academic pursuits in service-learning. The connected reflection helps bridge the theories learned in the classroom with the first-hand service experiences. *Challenging* reflection requires instructors' encouragement and stimulation to assist students in viewing traditional questions with new perspectives and putting new thoughts into the service design and delivery to better meet the needs of the community. Finally, *contextualized* refers to the contexts, where the services are delivered and how they provide a basis and orientation for reflection. For example, serving homeless community members prompts students to reflect on how poverty impacts one's living quality and orients the service plan to best serve the members. When designed with context in mind, reflection provides the linkage between thoughts and action.

Outcomes of Service-Learning for Students

Studies have found service-learning benefits students by cultivating their civic responsibility; helping them become more compassionate; enhancing their ability to solve social problems; and supporting their cognitive, attitudinal, moral, social, and personal development. In recent years, service-learning has moral, social, and personal development. In recent years, service-learning has been recognized as one of the crucial platforms in universities to increase students’ understanding of their roles in the community and encourage students’ reflection on community needs and universal virtues, and how these relate to what they are learning in their courses. Researchers have also found service-learning enhances students’ moral reasoning and judgment, psychological maturity, and self-esteem and lessens egocentric tendencies and prejudice.

Most of the research has shown significant improvements in students’ knowledge, critical thinking, and problem-solving skills. However, some scholars have cautioned that the findings on academic improvement in service-learning courses are still mixed. More appropriate test instruments and measures need to be developed to further investigate students’ intellectual gains from service-learning.

Florence Wu and Daniel Tan-lei Shek

See also Outcomes; Social Justice; Social Learning

Further Readings

Bringle, R. G., & Hatcher, J. A. (1999). Reflection in service learning: Making meaning or experience. Educational Horizons, 179.

Furco, A. (1996). Service-learning: A balanced approach to experiential education. Expanding Boundaries: Serving and Learning, 1, 1–6.

Kraft, R. J. (1996). Service learning: An introduction to its theory, practice, and effects. Education and Urban Society, 28(2), 131–159.

Speck, B. W., & Hoppe, S. L. (2004). Service-learning: History, theory, and issues. Greenwood Publishing Group.

Anita B. Delahay Anita B. Delahay Delahay, Anita B.

Lynne M. Reder Lynne M. Reder Reder, Lynne M.

Short-Term Memory Short-Term memory

1511

1513

Short-Term Memory

Short-term memory (STM) refers to what a person can remember from the immediate past. This is conceptually distinct from what a person can remember from all information stored during a lifetime, called long-term memory (LTM). There have been debates among memory researchers as to whether STM is a different memory store from LTM or whether STM is merely the information that is held in an active state within LTM. Regardless of how STM is conceptualized, in practical terms, it is one link in a much bigger chain of processes that begins with attention and perception and can lead to higher order cognition and learning.

This entry describes the *modal model of memory*, a model that helped define STM and was able to predict key features of STM such as its capacity and duration. It also describes challenges to the modal model, as well as a newer model that better explains the complex, real-world processes such as learning, multitasking, and intelligence.

The Modal Model of Memory

Building on the work of memory researchers in the 1950s and 1960s, Richard Atkinson and Richard Shiffrin, in 1968, proposed the *multi-store model*, which is the essence of the modal model. The theory was intended to account for how information from the outside world is encoded and gets into LTM. Visual or auditory information first goes into a sensory register (the first store), which acts as a perceptual buffer. Research by George Sperling in 1960 suggests that information can be held as physical features for about one quarter to one half of a second before it fades from sensory memory. Once stimulus information is categorized as known concepts by relying on LTM to identify its categorical features, the information is transferred to the second store, the short-term store. There the information is kept active in STM by rehearsal or repeating it over and over to oneself.

The longer an item is kept in STM and the more times it is rehearsed, the greater the chance it will be transferred to the third and final store, the long-term store. Some information will not be transferred to LTM due to longer times between opportunities to rehearse or more or lengthier items to rehearse. For example, in 1974, Alan Baddeley showed that the slower a person speaks, and therefore the longer the space between words, the greater the chance that the information will not transfer to LTM. Whether one can later retrieve the information depends on how much it is rehearsed or elaborated upon and the quality of the retrieval cues. In summary, according to the modal model, an individual must pay attention to stimuli in order to encode it; the individual must rehearse items in STM in order for them to transfer to LTM.

Measuring STM’s Capacity

Capacity of STM is often measured by the digit span test, in which an increasing span of digits (starting with four digits) is read quickly and then recalled from memory *in order of presentation*. Typical recall is 5 or 6 items, and recall above 7 is unusual. A useful psychological process to aid STM is called *chunking*, described by George Miller and Herbert Simon. They demonstrated that if information is chunked into high-level, meaningful units, then much more information might be recalled. For example, B I C I A F might be challenging to recall in order, but FBI CIA would not be. Yet those two letter strings are almost identical except for the transposition of one letter from the back to the front.

There is some debate as to how many chunks can be held in STM. Some suggest that it is seven plus or minus two, while others suggest that it is closer to four. However, the nature of the chunks is at least as important as the number. In 1980, researchers Anders Ericsson and Bill Chase reported on the undergraduate student Steve Faloon, who was able to repeat back very long strings of random digits, eventually reaching 82 digits, in order. He used many strategies but most critical was his knowledge of running times to recode the numbers into times to run various types of races, which were chunks of information that had a great deal of meaning to him as a runner. Therefore, chunking plus elaborative encoding, which helps a person organize information, can dramatically improve recall.

Encoding and Storing Information: Support for and Challenge to the Modal Model

Experiments involving free recall of word lists have provided empirical support for the modal model and a basic tenet of the model that, without rehearsal, STM has a very short retention interval of about 20 seconds. In these experiments, a distinct pattern of word recall, known as a serial position curve, emerges after 15 words are presented one by one (serially).

Reliably, the words that were presented last are recalled first. This is called the *recency effect*. Recent words benefit from still being in the rehearsal buffer of the short-term store. However, if the start of the recall task is delayed beyond the duration of STM, for example, to 30 seconds, the benefit of recency is no longer found.

The next words recalled are from the beginning of the list. This is called *primacy effect*. As the words are presented, the participant will begin to rehearse them (i.e., “apple,” “apple,” etc.). As more words are added to the list, more words need to be juggled in rehearsal, which becomes increasingly difficult. Because of this, the first items will benefit the most from having had the most rehearsal and are more likely transferred to LTM. If rehearsal is prevented, for example, by asking the participant to do a math task between the presentation of each word, then the primacy effect is disrupted.

However, other research has called the modal model into question by demonstrating that the number of rehearsals, per se, does not predict probability of recall, but rather how the information is processed. If information is elaborated on, or associated with information already stored in LTM, for example, by organizing it into a known category or forming visual or verbal images of it, then it is more likely to be encoded and stored long term, as was demonstrated in the Ericsson and Chase experiments.

Evolution of STM to Working Memory (WM)

The concept of STM has shifted in the past few decades as cognitive psychologists have learned more about how people use information in STM. A different theoretical construct called *WM* has replaced STM in that it has better predictive and explanatory power than the modal model when studying the complex, real-world processes such as learning, multitasking, and intelligence.

WM can be conceived of as a person's ability to focus attention, process information, and perform tasks. WM is required to perform STM tasks, and traditional STM measures such as the digit span or serial recall provide an estimate or a proxy for an individual's WM capacity. Indeed, such tasks are often used as a measure of intelligence. Although the modal model generated much research that advanced the field of memory science, WM surpassed the modal model due to its ability to more clearly explain what is happening in the present, including the role of attention and how STM is engaged in the real-world information processing tasks.

Anita B. Delahay and Lynne M. Reder

See also Long-Term Memory; Working Memory

Further Readings

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), The psychology of learning and motivation: Advances in research and theory (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. Science, 208, 1181–1182.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81–97.

Lisa L. Harlow Lisa L. Harlow Harlow, Lisa L.

Significance

Significance

1513

1516

Significance

Significance denotes that something is meaningful and of importance. Although significance tests can be used in different ways, *statistical significance* typically indicates the probability that a relationship among variables as large or larger as that found in a sample could have been drawn randomly from a population in which there is no relationship. In education and other fields, statistical significance has traditionally been the main focus, referring to a designation based on a dichotomous decision rule to reject or fail to reject a null hypothesis that there is not a relationship among variables in a population. Over the years, however, there has been increasing interest in encouraging a more detailed and informative approach to statistical inference to provide information on the practical significance or likely size of the relationship among variables in the population, and not just the statistical significance, of a research finding. In this entry, a brief history and some challenges of examining significance are provided, followed by practical significance methods, as well as limitations and future directions.

Brief History and Challenges of Significance Testing

In the history of statistical inference, Ronald Fisher, as well as Jerzy Neyman and Egon Pearson, provided early guidelines for researchers to follow when deciding whether a research study yielded noteworthy results or significance. The resulting procedure is referred to as null hypothesis significance testing, which is usually assessed by a statistical test such as a t-test, F test, or χ^2.

In its most simplistic practice, researchers historically would make a decision to either reject or retain a null hypothesis that usually stated that there were no meaningful effects. If the probability (i.e., *p* value) of the sample results was less than a designated value or α level (e.g., $p < .05$), the null hypothesis would be rejected and an alternative hypothesis, usually stating that there was some difference or association, would be retained showing a statistically significant effect (i.e., the result was significantly different from the null hypothesis value). Conversely, if the *p* value was at or greater than α (i.e., $p > .05$), a researcher could only claim that there was not enough evidence to find a significant result.

This practice of rejecting a null hypothesis when the *p* value was very small (e.g., < .05 or .01) helped researchers to have some degree of confidence in ruling out a chance finding. However, the overreliance on this dichotomous decision rule often resulted in statistical inferences that were not very informative. When used in isolation, a research result could be discussed as either significant or not, sometimes without the necessary specification of being statistically significant and many times without any reference to the size or meaningfulness of an effect. This is analogous to a physician saying that it is probable that you have an illness, with no additional input on what is the nature or seriousness of your condition. Staying with only this very limited approach is discouraged. Researchers are advised to provide more than just a statistical significance test result when presenting findings, offering a fuller basis on which to draw statistical inferences.

Statistical Inference Practices: More Than Statistical Significance

In recent years, researchers have increasingly enlarged their scope of what is involved in assessing significance. In addition to or instead of providing an indication of statistical significance, other inference procedures are encouraged to indicate the practical significance of a research study. This more meaningful approach, sometimes referred to as the *new statistics* entails the calculation of an effect size (ES) that provides a measure of the magnitude of a finding as well as a confidence interval (CI) that provides an indication of the degree of uncertainty around the estimate of a specific effect. Furthermore, replicating a study to verify findings, or conducting a meta-analysis that summarizes the effects over many studies, is preferred to trying to glean significance from the results of a single study. The following subsections present brief descriptions of some of these more informative inference practices, each of which provide supplemental information that can illuminate a finding beyond simple statistical significance.

ES

ESs are usually single numbers that provide an indication of the magnitude of a research finding, telling useful information about the practical significance of a research study. An ES can be as simple as a mean score. For example, an intervention study aimed at controlling weight may report an ES that is the average number of pounds lost in a treatment group compared to a control group. When presenting this information to the general public, this kind of ES can be very informative.

When comparing and reporting the averages across different samples or populations, particularly to a group of researchers, it may be more useful to choose an ES such as a standardized mean difference (i.e., the difference between two means, divided by the standard deviation, often averaged or pooled over the two groups being studied). This ES would be interpreted as the number of standard deviations difference there was between two groups, usually treatment and control groups. An often-used standardized mean difference ES, called Cohen's *d*, could be viewed as small, medium, or large, with values of 0.2, 0.5 and 0.8, respectively. Notice that these values correspond to almost a quarter of a standard deviation, half of a standard deviation, and close to a full standard deviation, although even higher values representing greater standard deviations of difference might be needed depending on the research area or if the cost of treatment is high. Consider that most people know that cutting back on heavy foods and increasing exercise could help in weight control. Thus, they might only be motivated to pay for an expensive weight control program if it showed very large differences or ESs. For example, if there were only two pounds' difference in weight loss between treatment and control groups, with a pooled standard deviation of 2.5 pounds, Cohen's *d* would be 2.0/2.5 = 0.8, a large ES, although the actual pound-loss sounds trivial. Doubling the weight loss to a 5-pound difference between groups, Cohen's *d* = 5/2.5 = 2.0, which could be viewed as a very large ES, but again, probably not enough to convince people to pay for such a program. It might take at least the hope of 10 or 20 pounds of weight loss, corresponding to Cohen's *d* ESs of 4.0 and 8.0, respectively, to show compelling and practical significance. Thus, it is important to take into account the context of a study and the pragmatic utility of an effect before claiming significance.

CIs

Whenever possible, researchers should provide a CI that provides an indication of the degree of uncertainty around a particular ES. CIs give lower and upper bounds on an effect to provide a range within which it could be expected that a large percentage (e.g., 95%) of numerous estimates of a CI would contain the true effect in the population. CIs that are fairly narrow indicate a more precise estimate of an effect, whereas very large CIs indicate a great deal of uncertainty and an imprecise effect estimate. For example, if a teacher said that the average score on an exam was 75, with scores ranging from 70 to 80, there would be more certainty about the value for a student’s particular score than if the mean was 75 and the interval ranged from 0 to 100.

Replication and Meta-Analysis

A significant finding should be replicated in independent samples to verify the results. When possible, it would be even better to conduct a meta-analysis that estimates an ES over a large number of studies, taking into account potential sources of variability within each of the studies. It may be that findings vary depending on specific characteristics of a group, called moderators. For example, there could be a higher degree of smoking for individuals who have a close friend or family member who smokes. In this case, we would say that the friend’s smoking status moderated the level of smoking for an individual. Examining the ESs from many studies, and identifying possible moderators that could have contributed to these effects, leads to stronger statistical inferences that would hopefully have both statistical and practical significance.

Limitations and Future Directions

Perhaps the biggest limitation when assessing significance occurs when researchers stop after assessing statistical significance. Thus, null hypothesis significance testing practices, although useful in ruling out a chance finding, do not guarantee that a finding is meaningful or practically important. A statistical test could have a large (e.g., *t*- or *F*-) value, and hence a small *p* value if there were a very large sample size, even if there were a small effect that is not practically meaningful. This would occur because there was a great deal of power to identify even minute effects, owing to the large sample size. For example, with a sample size of 10,000, a correlation of 0.03 would be evaluated as statistically significant (i.e., *p* = .00269), even though the effect is virtually 0. In this case, a behavior or characteristic with a minimal correlation could be believed to be a risk factor of an illness, which could be misleading if in fact the result was trivial and due to the particular sample. On the other hand, what is sometimes referred to as a medium correlation of 0.30 might emerge as being nonsignificant (i.e., *p* = .39969) with a small sample size of 10 in which researchers may have had difficulty recruiting participants for the treatment of a rare illness, for instance. In this latter case, there would not be enough power to notice a moderate ES. This could limit the possibility of a promising treatment being further explored if researchers based their conclusions solely on the results of a statistical significance test.

Another limitation is that a significant finding does not provide evidence in favor of a scientific or alternative hypothesis but rather evidence against the null hypothesis. That is, a significant finding *does not* indicate that there is a high probability for the alternate hypothesis prediction, only that there is little chance that you would find the specific result that emerged if the null hypothesis was true. Thus, it is important not to overstate, or even inadvertently misrepresent, the meaning of a significant finding.

Future research could continue to explore other options for assessing significance, such as Bayesian methods that take into account prior information, and could help provide evidence in favor of one hypothesis over another. A challenge with using Bayesian methods, however, is that they are computationally demanding and the procedures that are involved are not always widely known.

Researchers could also investigate how significance varies depending on the presence of moderators (e.g., age, gender, education, geographic area, income) or mediators (e.g., self-efficacy, social support, discrimination, powerlessness). Another area with growing interest is big data or data science, which might offer additional ways to discern the significance of patterns in large data sets as well as the challenges that are entailed. Finally, the field could benefit from more tutorials and clear expository articles that help make understanding and conveying the significance of statistical inferences more accessible and widely applied.

Lisa L. Harlow

See also Alpha Level; Confidence Interval; Effect Size; Hypothesis Testing; Power Analysis; *p* Value; Replication; Results Section

Further Readings

Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods, 21, 1–12. Retrieved from http://dx.doi.org/10.1037/met0000051

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. Theory & Psychology, 23, 251–270. doi:10.1177/0959354312465483

Harlow, L., Mulaik, S. A., & Steiger, J. H. (Eds.). (2016). What if there were no significance tests? (classic edition). New York, NY: Routledge.

Kelley, K., & Preacher, K. (2012). On effect size. Psychological Methods, 17, 137–152. Retrieved from http://dx.doi.org/10.1037/a0028086

Kline, R. B., (2013). Beyond significance testing: Statistics reform in the behavioral sciences (2nd ed.). Washington, DC: American Psychological Association.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling & Development, 80, 64–71. doi:10.1002/j.1556–6678.2002.tb00167.x

Van de Schoot, R., Mulder, J., Hoijtink, H., Van Aken, M. A. G., Dubas, J. S., de Castro, B. O.,… Romeijn, J.-W. (2011). An introduction to Bayesian model selection for evaluating informative hypotheses. European Journal of Developmental Psychology, 8, 713–729. doi:10.1080/17405629.2011.621799

Julius Sim Julius Sim Sim, Julius

Simple Linear Regression Simple linear regression

1516

1519

Simple Linear Regression

Linear regression is a form of statistical analysis whereby values on one variable (the outcome variable, denoted by *Y*) are predicted from values on another variable (the predictor variable, denoted by *X*) with which they are correlated. Here, "predict" does not necessarily have a temporal meaning but merely indicates that values on the outcome variable are estimated using values on the predictor variable. The analysis normally has one or both of two objectives: first, to obtain specific predicted values on *Y* that correspond to specific observed values on *X*; and second, to estimate the strength of this predictive relationship—that is, how well does *X* perform as a predictor of *Y*? The simplest case of linear regression, to be considered here, is where, in addition to the outcome variable, there is just one predictor variable; this is accordingly referred to as bivariate, or simple, regression. The case in which there are multiple predictors—multiple linear regression—is dealt with elsewhere.

Form of the Regression Model

The nature of the predictive relationship between the predictor variable and the outcome variable is expressed by two coefficients: the intercept (α) and the slope coefficient (β). These can be understood through a simple example. Imagine that a researcher wishes to predict students' exam scores (*Y*), measured on a 0–100 scale in a sample of 491 students, from a scale that measures their attitudes to schooling (*X*), with scores ranging from 0 to 30 (higher scores indicate a more positive attitude). The slope coefficient is the change in *Y* that is associated with a one-unit increase in *X*. A coefficient of .26 would indicate that for an increase of one point on the attitude scale, the predicted exam score increases by .26 marks. This relationship is constant across the scale of values—so that for a change in *X* from 12 to 13, or from 22 to 23, the change in *Y* is of the same magnitude. This is the basis of the term *linear* regression—the predicted values lie on a straight line.

The intercept is the predicted value of *Y* when *X* is 0 and is a constant. In some cases, the intercept has no real meaning—for example, if age were the predictor, no individual in this sample could have an age of 0—and it may also take a value that is not possible on the scale (such as a negative age). Nonetheless, the intercept is required to calculate the predicted scores. This will be clear if we look at the predictive equation:

$$\hat{Y} = \alpha + \beta X.$$

The symbol indicates the predicted value of the outcome variable. If we suppose that the intercept is 74.07, the predicted exam score for a student whose attitude score is 14 would be 74.07 + (0.26 × 14) = 77.71. Similarly, for a student with an attitude score of 21, it would be 74.07 + (0.26 × 21) = 79.53. Just as the intercept can be positive or negative, so can the slope. This will occur if the relationship between *X* and *Y* is negative. So, if we were seeking to predict exam performance from a measure of stress, we might find that a one-unit increase in stress is associated with a *decrease* in predicted exam score of, say, 1.4 marks and hence a negative slope coefficient of –1.4.

The aforementioned formula allows us to calculate the predicted scores on *Y* from scores on *X*. It does not tell us how strong this predictive relationship is. Let us consider a different formulation of the regression equation:

$$\hat{Y} = \alpha + \beta X + \varepsilon.$$

Here, *Y* indicates the *observed* value of the outcome variable, and ε indicates the residual, or the error of prediction—that is, the difference for a particular case between the predicted and the observed values of the outcome variable (*Y*–). If *Y* is underpredicted by , the residual will be positive, whereas if *Y* is overpredicted by , the residual will be negative. The smaller the residuals are on average, the better the goodness-of-fit of the regression model and the greater its predictive power. If we correlate the predicted and the observed values of *Y* and then square the resulting correlation coefficient, we obtain a statistic called the coefficient of determination, r^2. This can take values between 0 and 1 and can be interpreted as the proportion of the variance in *Y* that can be explained, or accounted for, by *X*. Higher values of r^2 indicate closer fit of the regression model to the data and are therefore better.

The method of ordinary least squares is normally used to fit the regression line to the data. This method finds the straight line, of all possible such lines, that minimizes the sum of the squared deviations of the observed values from this line—that is, it minimizes the sum of the squared residuals and thereby has the optimum fit to the data.

Estimating Population Parameters

The values of the intercept and the slope, α and β, are calculated from the data at hand and are therefore sample statistics. However, our interest normally lies not in the sample but in the population from which it was drawn. We are not directly concerned with the relationship between attitude scores and exam marks in the particular sample of students that we have measured; rather, we are interested in generalizing our findings to the population of such students. We therefore use the sample values of α and β as estimates of the corresponding population parameters.

Two issues are important here. The first is that we need to know how precise the sample values of α and β are as estimates of the corresponding population values of α and β. This can be accomplished by calculating a confidence interval. A 95% confidence interval around the sample estimate of .26 for β might run from .14 to .37. Although our best estimate of β is .57, there is range of plausible alternative values between .14 and .37. So, if the true value of β is not .57, we can be 95% confident that it is no smaller than .14 and no larger than .37. The wider this range, the lower the precision of the estimate and hence the greater the uncertainty about the true population value. The second issue is that we would normally want to determine whether our sample estimate of β is statistically significant (we can also see if the estimate of α is statistically significant, but we are not normally interested in this). The output from a regression analysis will provide a *p* value for a regression coefficient, and this can be used to test the null hypothesis that the population value (the parameter) is 0. If the *p* value lies below our predetermined cutoff for significance (e.g., $p \leq .05$), we can reject this null hypothesis and conclude that the coefficient in the population has a nonzero value. The same decision can be reached by examining a 95% confidence interval for the coefficient—if it excludes 0, the coefficient is statistically significant at $p \leq .05$.

As with all forms of inferential statistics, sample size is an important consideration. If the sample is too small, the sample regression coefficient will overestimate the corresponding population value and the r^2 will accordingly be inflated. In addition, the standard errors of the regression coefficient, and hence the associated *p* values, will be large; statistical significance will thereby be hard to achieve. A sample size calculation should therefore be performed, wherever possible, prior to collecting data.

Assumptions of the Analysis

Linear regression is a parametric statistical procedure and thereby makes certain assumptions about the data. First, the predictive relationship is assumed to be linear, and this should be tested by examining a scatterplot of *X* against *Y* and judging whether it is reasonable to fit a straight line to the plotted data. If the plot suggests that a curvilinear relationship is more plausible, linear regression will not be appropriate, unless adaptations are made to the basic model, such as an appropriate data transformation or the use of a polynomial (e.g., squared) term for *X* in the regression equation. The linearity of the relationship implies a second assumption to do with the level of measurement of *X* and *Y*. As the relationship between *X* and *Y* is constant across the scale of values of *X*, both must be on an interval or ratio scale (or, at least, a scale that can justifiably be treated as interval or ratio). A binary predictor can also, in fact be used, though this situation is probably more common in multiple linear regression. If the predictor variable is a nominal or ordinal variable with more than two levels, it must first be converted into a set of binary dummy variables (where there is one less dummy variable than there are levels on the original variable). If the outcome variable is ordinal rather than interval or ratio, an ordinal regression model should be used instead of linear regression.

Further assumptions concern the residuals. These are assumed to be independent (i.e., the value of one residual does not influence, and is not influenced by, the value of any other residual) and to have homogeneity of variance (also referred to as homoscedasticity). The latter assumption can be tested by plotting the residuals on the vertical axis of a scatterplot against the predicted values of *Y* on the horizontal axis. Homogeneity of variance implies that the degree of scatter of the residuals around their mean value will be constant from left to right across the range of the predicted values of *Y*. If confidence intervals are constructed or hypothesis tests are performed, the residuals are also assumed to be (approximately) normally distributed, with a mean of 0. Importantly, no assumptions are made about the distribution of *X* or *Y*, though if one or the other is markedly skewed this may give rise to nonlinearity. A final assumption concerning the residuals is that they are not correlated with *X*; this can be assessed from a simple scatterplot.

Strictly, *X* is assumed to be a fixed, rather than a random, variable and to be measured without error. Error in the measurement of *X* will cause the estimate of β to be biased (underestimated in the case of simple linear regression). These two assumptions are rarely met in practice, as there is frequently a need to utilize a random variable as a predictor and measurement error is present to varying degrees with most interval or ratio variables. However, even if *X* is a random variable, linear regression is generally considered to function well provided that the *X* variable is not correlated with the residuals. In addition, researchers can try to ensure that measurement error is minimized.

Other Considerations

It was noted earlier that the intercept, as the predicted value of *Y* when *X* is 0, may not have a meaningful interpretation. However, if the values of *X* are centered, a more useful interpretation is often possible. Centering is a transformation that involves subtracting every value from the mean, so that the resulting values are deviations from a mean of 0. For a centered predictor, the intercept becomes the predicted value of *Y* when *X* is at its mean value, rather than when *X* is 0, and this may be more useful information.

We saw that we can calculate a confidence interval for the regression coefficient, and we can do the same for the predicted values of *Y*. For an attitude score of 21, the predicted value of *Y* was 79.53. A 95% confidence interval around this value provides a range of plausible alternative predicted values of *Y* when *X* takes the value 21—in this case, the confidence interval runs from 78.58 to 80.27. Again, the narrower the confidence interval, the greater the precision. In addition, we can construct what is known as a prediction interval for predicted values of *Y*. This, however, has a different interpretation from that of a confidence interval. A 95% prediction interval for the predicted value of *Y* indicates a range of scores within which 95% of individual values of *Y* in the population are expected to lie, for a given value of *X*. In the current example, when *X* is 21, the 95% prediction interval runs from 61.47 to 97.38, telling us that 95% of students with an attitude score of 21 would be expected to have exam scores between these limits.

Two final caveats are worth noting. First, one should be aware of the effect of extreme values of *Y*, as these may exert a large influence on the coefficients and on the fit of the model. Cases with large residuals give an indication of such extreme values and more specific statistics (such as the leverage and the Cook's distance) can be used to determine the influence of such values on the regression model. Second, one should not try to predict beyond the range of values of *X* in the sample. If *X* were age and the observed values in the data were from 116 to 145 months, we have no information on the nature of the predictive relationship outside this range. It is, for example, possible that the relationship between *X* and *Y* becomes nonlinear at values of *X* greater than 145 months.

Julius Sim

See also Dummy Variables; Goodness-of-Fit Tests; Multiple Linear Regression; Residuals; Scatterplots

Further Readings

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Draper, N. R., & Smith, H. (1998). Applied regression analysis (3rd ed.). New York, NY: Wiley.

Kahane, L. H. (2008). Regression basics (2nd ed.). Thousand Oaks, CA: Sage.

Lewis-Beck, M. S. (Ed.). (1993). Regression analysis. London, England: Sage.

Miles, J., & Shevlin, M. (2001). Applying regression and correlation: A guide for students and researchers. London, England: Sage.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (4th ed.). New York, NY: Wiley.

David Westfall David Westfall Westfall, David

Simple Random Sampling Simple random sampling

1519

1522

Simple Random Sampling

Researchers are often faced with the task of making statements about entire populations. However, including every member of a population into a study is often not possible and simply not feasible. Thus, subsets of the population (samples) must be chosen to represent the population. If samples are collected properly, precise statements can be made about a population, with a fairly high degree of confidence, from relatively small samples. Numerous techniques have been developed to ensure that the subset, or sample, is representative of the overall population so generalizations can be made. Simple random sampling is a probability method of selecting a subset, or sample, from a larger population in such a manner that every element (individual member of the population whose characteristics are to be measured) has the same probability of being chosen into the sample during each stage of the sampling process.

Simple random sampling is one of the most basic and simplest forms of sampling used. The basic principle behind the method is that every element in a population retains the exact same probability of being selected into a sample. For example, a university campus has 2,000 parking spots and 6,000 students who applied for a parking permit. The 2,000 available permits are distributed so that each of the 6,000 students has the same probability of receiving a permit. This can be achieved by numerous means, such as putting names in a hat or using student identification numbers and a random number generator to choose the first 2,000 students. Typically, sampling in this manner is done without replacement. That is to say, once a student’s name is drawn it is removed from the pool of remaining students. Although this technically does impact the odds of the remaining students, during the initial setup each student has the exact same probability of selection. Sometimes, sampling with replacement is used to ensure exact probability of selection remains for each element.

Requirements

Simple random sampling requires a population, sampling frame, and elements. A population is the entire set of entities from which a sample will be drawn. In the previous example, the population is the 5,000 students. Elements are the individual members of the population. This could be people, families, nations, schools, classes or whatever the researcher is examining. In the example provided, the elements are the students. The sampling frame is similar but has an additional important feature. The sampling frame is a list of all of the elements or other units containing the elements in a population. In the example provided, a sampling frame would be a list of all of the student applicants. Developing a sampling frame and identifying all of the elements in a population can prove challenging for researchers. Some research environments are simply not conducive to this type of probability sampling. For example, conducting research on the homeless or research conducted in conflict-impacted areas may prove challenging to identify, or locate, members of the population. Developing the sampling frame can also prove challenging in less volatile environments. Something as simple as creating a sampling frame of faculty at a university can prove challenging. The conceptualization of "faculty," or whatever the researcher is studying, is vitally important to the generation of a sampling frame. A major challenge faced by people who conduct polls is that the phonebook or voter registration rolls are not complete. The increase in cell phone use and reduction in landlines mean that attempting to use the phonebook as a sampling frame does not allow for the inclusion of those who have moved away from landlines, as such not everyone in the population has the equal probability of selection. Although it may seem insignificant, and the researcher could decide to create a sampling frame from the phonebook, there are systematic differences in age, sex, race, and voting tendencies that have been known to significantly impact findings. In simple random sampling, the researcher must be sure that any differences in characteristics between the sample and characteristics of the population are due purely to chance and not introduced by selection bias.

Sampling Techniques

Simple random sampling is one of several probability sampling techniques (systematic random sampling, stratified random sampling, and cluster sampling are some others). Probability sampling is any method of sampling that uses random selection from a population for inclusion into the sample. Some processes must be used to ensure that all elements in a sampling frame have equal probability of inclusion in the sample and that the probability of inclusion can be calculated for each element. After creation of the sampling frame, there are numerous ways to choose the sample. Some of the more popular methods include random number generators, random number tables, placing names/numbers into a hat, or rolling dice.

Systematic random sampling is a very similar method to simple random sampling. It is another method of probability sample. The identification of elements and creation of a sampling frame remains the same as in simple random sampling. The major difference is in how elements are chosen. After the sampling frame is developed, the sample size (*n*) is determined. The number of elements in the population is divided by the desired sample size. The resultant number is the sampling interval. Finally, a random number is chosen between 1 and whatever number represents the sampling interval. The element associated with that number is the first element selected in the sample. Then, every *n*th element is selected as determined by the sampling interval. In the example provided earlier, the 6,000 students in the population would be assigned numbers in order between 1 and 6,000. Because there are 2,000 parking passes (*n*), 2,000 elements need to be selected. The population (*N*) 6,000 divided by the needed sample size (*n*) 2,000 results in a sampling interval of 3. Using a random number generator to select a number between 1 and 3, results in the number 2. The second element is selected into the sample. Then, every subsequent third element (the sampling interval) is selected into the sample until all 2,000 permits are filled (2, 5, 8, 11, 14, …).

Descriptive Versus Inferential Statistics

Probability sampling provides the most valid and credible results because they reflect the characteristics of the population they are chosen from. Because of this, probability sampling methods move beyond descriptive statistics and are the basis of inferential statistics. Descriptive statistics describe or summarize data. Although they allow for emergent patterns in the data to be viewed, they do not allow for conclusions beyond the data that has been collected from a particular sample. Descriptive statistics are useful for describing data and allowing for visualization. They involve measures of central tendency, such as the mean, median, and mode, to visualize the center of a group of data. Additionally, descriptive statistics summarize measures of dispersion or spread, such as the standard deviation, variance, and inclusive and exclusive ranges.

such as the standard deviation, variance, and inclusive and exclusive ranges.

Simple random sampling, as a probability sampling method, allows for the use of inferential statistics. Inferential statistics allow the researcher to make inferences or “infer” the value of a population parameter and thus make larger statements about a population (generalizability) from a relatively small sample. The researcher is able to draw conclusions beyond the sample alone. Political pollsters do this every election cycle, often predicting elections with high degrees of certainty from very small samples. Sampling involves a degree of sampling error. In any population, there is an infinite number of possible samples. The statistics, measures of central tendency and dispersion or spread, of each possible sample becomes an estimate of the parameters, measures of central tendency and dispersion or spread, of the population. The statistics of each sample can be viewed as an estimate of the population parameters.

The central limit theorem states that a sufficiently large number of independent random variables will be approximately normally distributed. Most cases will cluster around the mean of the population and the further away from the mean of the population, the fewer the number of cases, taking the shape of a normal (bell) curve. If a sample contains a large number of observations and the sample is collected randomly, the mean of the sample will become an estimate of the parameters of the population. Each sample of an infinite number of possible samples in a population acts the same way. As more samples are drawn, the estimates will take the shape of a normal curve. The central limit theorem states that the averages will be distributed according to the normal distribution.

Advantages and Disadvantages

Simple random samples reduce the potential for human bias when selecting elements. As such, samples are assumed to be representative of the population. The major advantage of simple random sampling is that the samples are chosen using probability sampling methods and thus allow for generalization or inferences to the larger population.

The disadvantages of simple random sampling are that a sampling frame, or a list of the entire population being studied, is identifiable and is not missing any of the elements. Complete lists of the population can be difficult to attain, may not be made public, may be expensive to gain access to due to privacy policies, the different means of contacting a sample spread out geographically can be challenging, or in the case of volatile environments, may be simply impossible to challenging, or in the case of volatile environments, may be simply impossible to know.

David Westfall

See also Descriptive Statistics; Inferential Statistics; Normal Distribution

Further Readings

Carlson, K. A., & Winquist, J. (2014). An introduction to statistics: An active learning approach. Thousand Oaks, CA: Sage.

Schutt, R. (2017). Understanding the social world: Research methods for the 21st century. Thousand Oaks, CA: Sage.

Szafran, R. (2012). Answering questions with statistics. Thousand Oaks, CA: Sage.

Inbal Yahav Inbal Yahav Yahav, Inbal

Galit Shmueli Galit Shmueli Shmueli, Galit

Simpson’s Paradox

Simpson’s paradox

1522

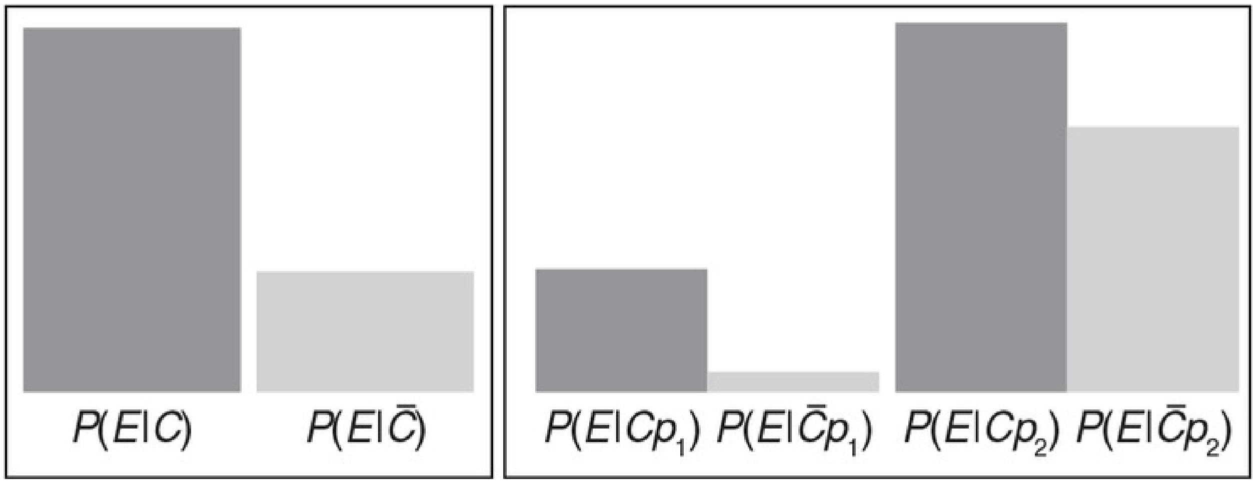
1524

Simpson’s Paradox

Simpson’s paradox, first defined by Edward H. Simpson in 1951, is a statistical phenomenon in which the association between two variables reverses or disappears when examining aggregate versus disaggregate data of a population via a third variable. Alternative known names of Simpson’s paradox are *Yule effect*, *reversal paradox*, or *amalgamation paradox*.

The practical implication to decision making that Simpson’s paradox raises is the question of which level of data aggregation presents the results of interest. This question further raises the challenge of identifying potential variables and then establishing a criterion for deciding if and which of the potential variables should influence the decision making.

Figure 1 Simpson’s paradox illustration for categorical cause and outcome variables



Simpson’s paradox is commonly defined for a categorical cause variable (*C*) and a categorical outcome variable (*E*) as the phenomenon whereby an event *C* increases the probability of *E* in a given population *p*, at the same time, decreases the probability of *E* in every subpopulation of *p* (see Figure 1).

Mathematically, the paradox is defined for two events and their complements: $Y = \{E, E^c\}$ and $X = \{C, C^c\}$, and a population *Z* with subpopulations $\{p_1, p_2, ..., p_n\}$, for which the following relationship holds:

$$P(E \mid C) > P(E \mid C^c), \text{ and}$$

$$P(E \mid Cp_i) < P(E \mid C^c p_i), \text{ for each } p_i \subseteq Z.$$

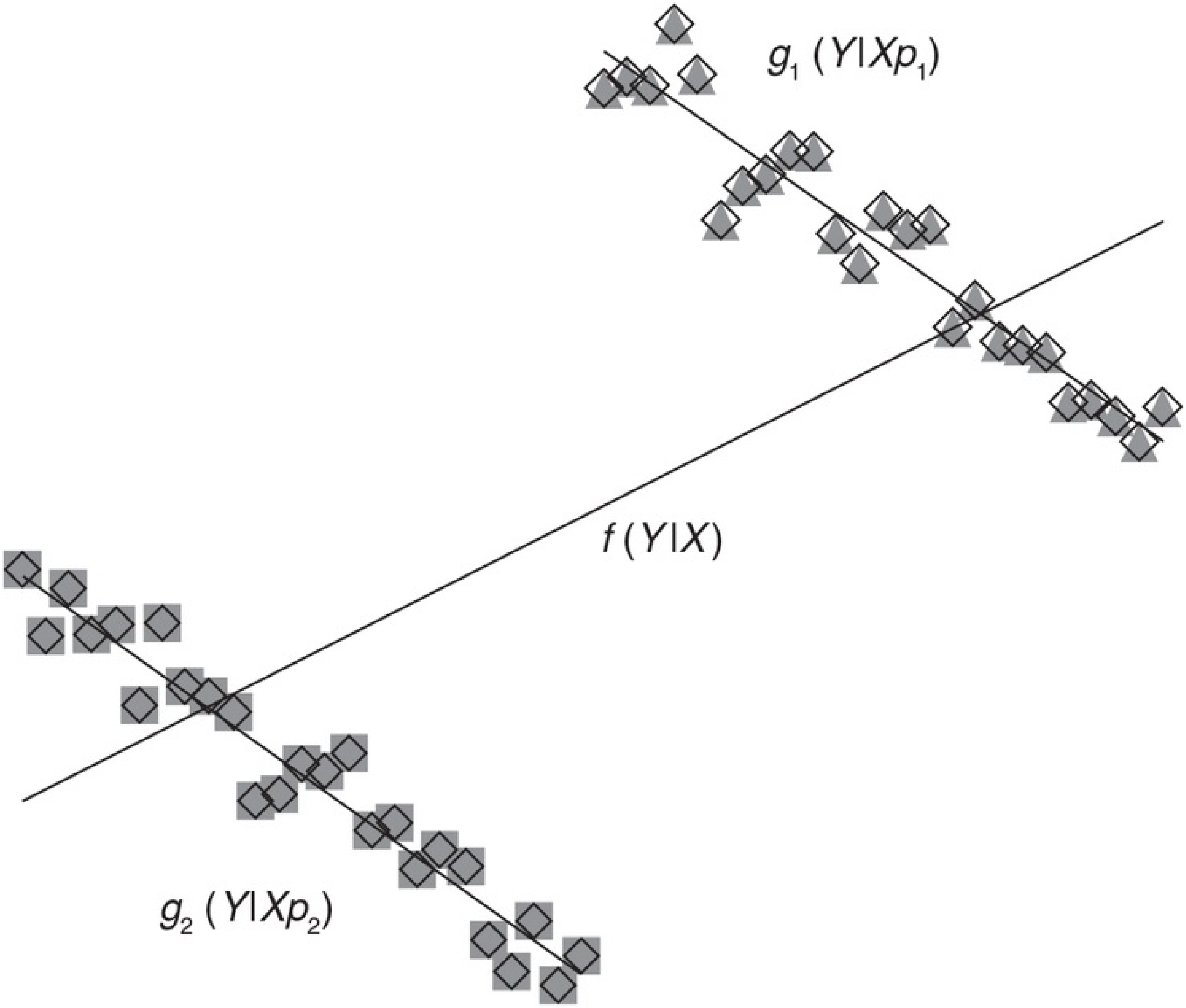
These inequalities can also be encountered in the form where the symbols < and > are reversed.

For continuous cause (*C*) *and* effect (*E*) *variables*, the association between *Y* and *X* are defined by two functions: the monotonic function *f*: $Y = f(X)$ for the overall data, and a monotonic function *g* for each of the data subpopulations: $Y = g_i(X|p_i)$, which has the opposite sign (see Figure 2).

Simpson’s paradox commonly arises when the underlying causal structure of the data, which is unidentifiable only from the data, is not considered by the researcher. Therefore, combining observational data with causal theory can resolve the paradox and determine the correct level of data aggregation.

We use the notation *X* to denote the cause variable, *Y* is the outcome variable, and *Z* is the potential reversal variable (or a vector of multiple potential reversal variables).

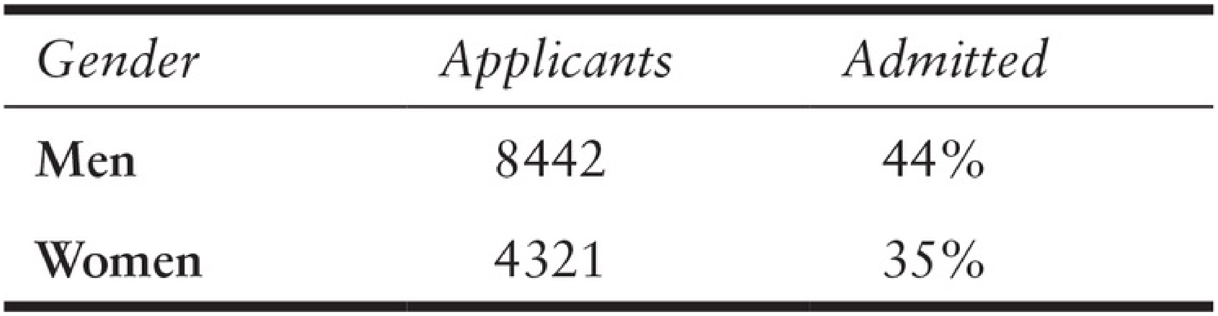
Figure 2 Simpson's paradox illustration for continuous cause and outcome variables



Examples

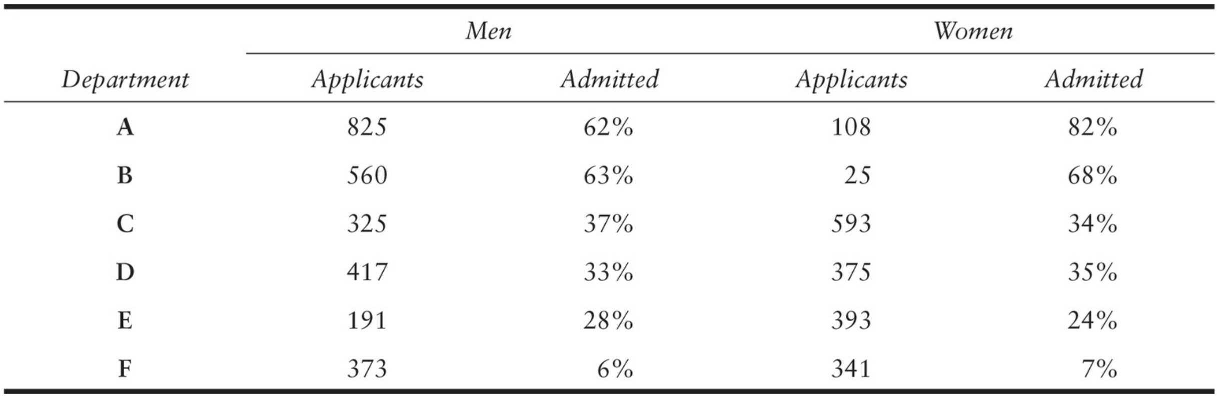
Example 1: Berkeley Admissions

Probably the most famous example of Simpson's paradox is the Berkeley admissions case. Given data on admissions to the different departments at UC Berkeley in 1973 (*Y*), and given the gender of each applicant (*X*), the aggregate data indicated a lower rate of admissions for women (see Table 1). The question that arose was therefore the existence of a gender bias against women in admissions. However, when broken down by department (*Z*), admission rates were found to be higher for women in almost every department (Table 2).



Gender	*Applicants*	*Admitted*
Men	8442	44%
Women	4321	35%

Source: Data from Bickel, P. J., Hammel, E. A., & O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404.



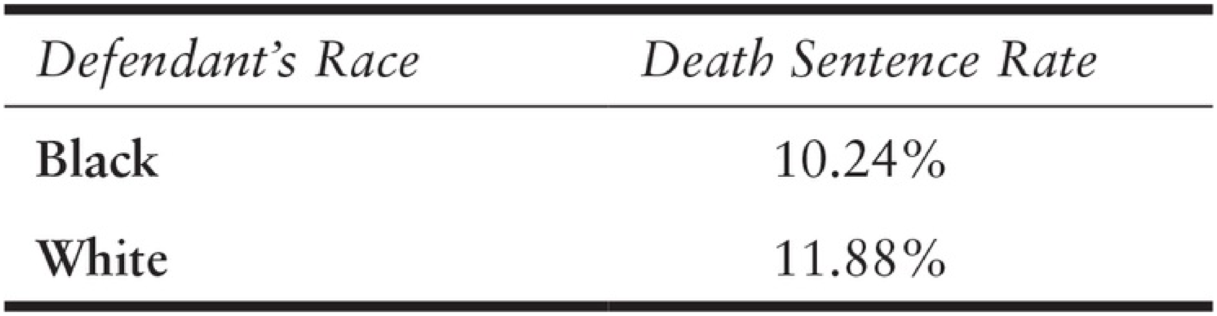
	Men		*Women*	
Department	*Applicants*	*Admitted*	*Applicants*	*Admitted*
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Source: Data from Bickel, P. J., Hammel, E. A., & O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*(4175), 398–404.

Example 2: Death Sentence Rates

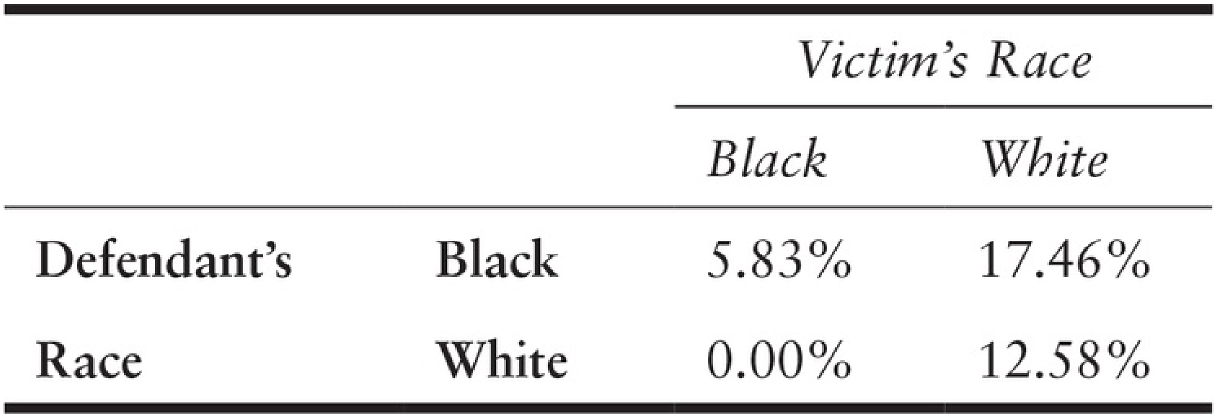
The death sentences example described by Alan Agresti contains information on 326 murder cases in Florida. In each case, data are available on the race of the defendant (*X*), whether the outcome was a death sentence (*Y*), and the race of the victim (*Z*). The question of interest is whether the defendant’s race affects the probability for a death sentence, thereby indicating racial bias. The potential reversing variable is the race of the victim.

Examining the contingency tables of the aggregate and disaggregate data indicates Simpson’s paradox. The aggregate data (Table 3) indicates that White defendants are more likely to get the death sentence than Black defendants. In contrast, the contingency table disaggregated by victim's race (Table 4) indicates that White defendants are less likely to get the death sentence when the victim is Black as well as when the victim is White.



Defendant's Race	*Death Sentence Rate*
Black	10.24%
White	11.88%

Source: Data from Agresti (2012).



		Victim's Race	
		Black	*White*
Defendant's	Black	5.83%	17.46%
Race	White	0.00%	12.58%

Source: Data from Agresti (2012).

Inferring a Paradox From a Sample to a Population

In cases in which the paradox is detected in a sample, in order to infer whether the paradox generalizes to the population, statistical inference is required regarding the effect of the interaction between *X* and *Z* on *Y*. For example, in the death sentence case, if the 326 murder cases are a sample from a larger population for which inference is required, then the presence of an apparent paradox in the sample does not necessarily mean the paradox exists in the population. We can use a logistic regression model of outcome *Y* on independent variables *X*, *Z*, and their interaction and assess the statistical significance of the interaction term.

Detecting Simpson's Paradox

Detecting Simpson’s Paradox

Detecting Simpson’s paradoxes in observational data is a challenging problem. Research shows that despite the commonality of the paradox, people are often poor at recognizing it. When it goes unnoticed, incorrect inferences may be drawn, and as a result, decisions may be misguided, sometimes leading to adverse consequences.

A statistical approach to detecting the paradox is the use of a regression model with interaction terms between the cause (X) and each of the potential reversal variables. The drawbacks of this approach are 2-fold. First, in high dimension data, the model becomes very complex and computationally inefficient. Second, the regression model treats the cause (X) and the potential reversal variable (Z) symmetrically, regardless of the strength of their effect on the outcome variable (Y). That might lead to misinterpretation of the causal structure of the data, as discussed earlier.

A recent alternative approach is the use of classification and regression trees for detecting the paradox. The trees consider the effects of variables sequentially, and therefore the tree structure can be used for determining whether a paradox might exist. This approach is also efficient when applied to big data.

Simpson’s Paradox and Big Data

Simpson’s paradox poses two challenges in big data. The first stems from considering a very large sample that can be broken down into many subpopulations. It is therefore more likely to find niche subpopulations, such as isolated communities, that behave differently than the mainstream. For these niche groups, data might exhibit partial paradoxes. The second challenge is the high dimensionality of big data, which, in the context of Simpson’s paradox, results in a large number of potential reversal variables (“needle in a haystack” problem).

Inbal Yahav and Galit Shmueli

See also Categorical Data Analysis; Chi-Square Test

Further Readings

Agresti, A. (2012). Categorical data analysis (3rd ed.). Hoboken, NJ: Wiley.

Blyth, C. R. (1972). On Simpson’s paradox and the sure-thing principle. Journal of the American Statistical Association, 67, 364–366.

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). New York, NY: Cambridge University Press.

Shmueli, G., & Yahav, I. (2016). The forest or the trees? Tackling Simpson’s paradox with classification and regression trees. Retrieved from https://ssrn.com/abstract=2392953

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society B, 13, 238–241.

Kathleen Lynne Lane Kathleen Lynne Lane Lane, Kathleen Lynne

David J. Royer David J. Royer Royer, David J.

Single-Case Research Single-Case research

1525

1528

Single-Case Research

Single-case research involves an experimental approach to answering the question: Is there an effect? Questions of causal inquiry can be answered using randomized control trials (RCTs), regression discontinuity designs, as well as single-case research designs (SCRDs). This entry focuses on SCRDs, which enables researchers and practitioners to determine whether a strategy, practice, or program (independent variable—*x*) impacts a given behavior (dependent variable—*y*). For example, one may want to determine whether incorporating instructional choices during writing instruction (*x*) increases student engagement (*y*) for four middle school students. This question can be answered using an SCRD. A common misconception is that single-case research is the same as case studies (which are not experimental designs). This is simply not correct.

SCRD can be a highly effective approach for examining effects when RCTs are not feasible. There are often times when there are too few participants involved to acquire the statistical power needed to detect an effect, when ethical considerations do not support the use of a control group (which is required in RCTs), or the cost of RCTs is prohibitive. Furthermore, there are instance in which SCRD is simply the preferred methodology due to the flexibility of the design process that enables phase changes to be conducted in response to student performance. SCRD can be particularly useful in early stages of inquiry when interventions are developed and tested, or when exploring solutions for participants requires more intensive intervention efforts than those initially planned. In fact, SCRD can be the methodology of choice in programmatic lines of inquiry to determine “what works.” Researchers use SCRD to develop, test, and refine a given strategy, practice, or program before moving to RCTs to examine efficacy and effectiveness of established interventions examine efficacy and effectiveness of established interventions.

This entry introduces the logic of SCRD followed by an overview of common SCRD designs: A-B-A-B withdrawal designs and multiple baseline designs (MBDs). The entry concludes with a brief discussion of strengths and considerations when employing SCRD.

Single-Case Design Logic

As mentioned, SCRDs are highly flexible experimental designs enabling one to determine “what works” when supporting an individual student, a group of students, or a school. SCRDs grew out of the field of applied behavior analysis, which involves the application of key behavioral principles (e.g., positive reinforcement) to applied contexts. The methodological logic includes four key components.

First, SCRD involves within-subject design whereby each participant serves as their own control. This is very different from group design methodology, which involves between-groups comparisons. For example, in traditional group designs, students with similar concerns (e.g., high levels of internalizing behaviors) are detected using consistent procedures (e.g., data from systematic screeners) and randomly assigned (such that each person as an equal and independent opportunity to be assigned to established conditions) to either an experimental (e.g., cognitive restructuring) or control (e.g., business as usual) condition. Using this methodology, students are often pretested on measures of interest (e.g., participation in class discussions) before the intervention begins (pretest) and again after the intervention concludes (posttest and later maintenance). In group designs, the intervention condition is fixed—the intervention does not change, participants in the control group do not access the intervention, and integrity data are collected to make sure there is no contamination between conditions. Then, analyses are conducted to determine how performance shifts over time for students who did (experimental condition) and did not (control condition) receive the intervention being tested. In SCRD, each student receives the intervention following a baseline condition. Using frequent, repeated assessment of student performance during baseline and then intervention conditions (rather than pre-/post intervention assessments), comparisons are made to see how each students’ performance shifts following systematic introduction of the intervention (e.g., cognitive restructuring) on the performance measures of interest (e.g., participation). These are within-subject comparisons.

comparisons.

Second, SCRD involves systematically introducing the independent variable (e.g., cognitive restructuring) that leads to systematic changes in the dependent variable of interest (e.g., participation). In this contingent relationship, a participant’s performance on the dependent variables *depends* on whether the intervention is in place (e.g., does participation in class activities increase for students as they learn and apply cognitive restructuring techniques?). In this design, a functional relation between *x* and *y* is established when this change in behavior is demonstrated one time and replicated at least two more times. In other words, a functional relation exists when students’ participation increases from baseline conditions only after the intervention is introduced (and does not increase before the intervention is put in place).

Third, SCRD involves frequent, repeated assessment of dependent variables rather than pre-, post-, and maintenance data collected in group design studies. For example, if a teacher wanted to test the usefulness of increasing students’ opportunities to respond using response cards for students with high levels of disruptive behavior, the teacher could collect baseline data on disruption and engagement using direct observation techniques for five class sessions before the intervention was introduced and again during the 10 class sessions when the intervention was introduced (as well as when the intervention was withdrawn and subsequently reintroduced—a point that will be discussed more fully). These data points within each phase (baseline, intervention, withdrawal, and reintroduction) are connected with lines to form data paths. By conducting repeated assessment, intervention effects can be determined by analyzing changes in stability, level, and trend over time, leading to the fourth feature of SCRD.

Fourth, SCRD involves analyzing data using these visual inspection techniques as well as statistical analyses recently developed to examine the magnitude of the effect. It is important to note that application of statistical tests (e.g., effect sizes) with data gleaned from SCRD is rather controversial; however, such tests are important to ensure treatment outcome studies conducted using SCRD can be taken into account when synthesizing bodies of evidence to determine whether a given practice (e.g., instructional choice, cognitive restructuring, and increasing opportunities to respond using response cards) is an evidence-based practice.

These four components of SCRD are the foundation of this experimental approach. There are several designs that can be used to answer questions such as: Does *x* lead to changes in *y*? (demonstration questions), does more or less of *x* lead to changes in *y*? (parametric questions), does *x* lead to more pronounced changes in *y* with some or all of the intervention elements in place? (component analysis questions), and does x_1 or x_2 lead to greater improvement in *y*? (comparative questions). The next section provides descriptions of two common SCRDs that can be used in isolation and combination to address these various questions.

Common SCRDs

The process does not begin with selecting a design. The process begins with selecting the outcome of interest: the dependent variable (*y*)—*y* drives the design. Researchers begin by defining what they want to change (e.g., increased engagement; decreased disruption), moving to the underlying logic model that establishes the theoretical rationale for this change, the questions of interest, and finally design selection. This section discusses two core designs: A-B-A-B and MBDs. These designs provide a pattern of responding where change is desired (e.g., increasing engagement or decreasing disruption), determined by examining baseline performance (which should contain a minimum of five data points). These designs also offer the *possibility* of establishing an experimental effect by showing changes in dependent variables that occur only when the intervention is implemented, with data from adjacent phases compared to examine stability, level, and trend. Finally, a functional relation (experimental control) is established by three instances of basic effects over time (one demonstration and two replications).

A-B-A-B Designs

In an A_1-B_1-A_2-B_2, the A (baseline A_1 or withdrawal A_2) and B (intervention) phases are described with precision such that anyone could read the description and replicate each phase. Each phase must contain a sufficient number of data points to be certain the pattern of responding is clear. Recently, the minimum number of data points required has shifted from three to five required per phase. With an A_1-B_1-A_2-B_2 design, the first basic effect (A_1-B_1) can be shown when a change in the main dependent variable of interest (e.g., engagement) occurs when the intervention is introduced in Phase B_1 (e.g., instructional choice) and all other elements of the Phase A_1 remain. The second basic effect can be shown when the intervention is withdrawn and elements of the Phase A remain in place after removing the intervention (B_1-A_2). The third basic effect can be shown when the intervention is reintroduced (A_2-B_2), allowing a chance to replicate the first basic effect when the intervention was first introduced. Although some people may be opposed to a withdrawal phase, in this design, the withdrawal is essential to establishing a functional relation between the introduction of the intervention and changes in the dependent variable. However, there are instances in which it would be unethical to use an A_1-B_1-A_2-B_2 design such as when addressing extremely self-injurious behaviors. Also, there are times when an A_1-B_1-A_2-B_2 design is not possible. For example, one cannot “undo” or withdraw the effects of learning. If the intervention is addressing an acquisition (skill) deficit such as teaching a student how to decode or a cognitive restructuring intervention (which involves teaching the student a new method of self-talk that is constructive rather than self-defeating), this newly learned skill cannot with withdrawn. In these instances, an MBD is needed.

MBDs

The MBD provides an opportunity to answer questions of effect when it is not feasible or reasonable to withdraw the intervention of interest (reverse the initial effect). The MBD enables one to show a change when the intervention is introduced and provide evidence that is likely the intervention (and not maturation or other variables) is likely responsible for the change on the dependent variable.

A key feature of the MBD is staggered introduction of the intervention. First, it is important to determine a minimum of three data series (but ideally four to protect against attrition). A data series refers to a set of repeated measurements such as the percentage of intervals in which a student is engaged academically, the number of times the students raise their hand during a 15-min discussion held each day, or the rate of positive interactions occurring during recess. The data series can be defined across three or more participants (e.g., Nathan, Katie, and Gabe), across three or more settings (e.g., classroom, playground, and after-school-care), or across three behaviors (e.g., walking, bike riding, and swimming). These refer to as MBD across participants, settings, and behaviors, respectively. In each, the person collecting data begins collecting baseline data on the defined dependent variables at the same time for all data series (e.g., all participants) to ensure three or more baselines suggest a steady pattern of responding. The desired direction for change is specified (e.g., increase engagement, decrease disruptive behavior, or reduced variability in contributing to group discussions). Then, under optimal baseline conditions, the series would contain five data points and if the data path was stable, the intervention would be introduced for the first series (e.g., the first student, Nathan). Data collection would continue in all series until the dependent variables begin to demonstrate the desired change (e.g., increase in engagement). Once this effect was evident, the intervention would be introduced in the second series (e.g., with the second student, Katie), whereas data collection continues in all series. Once the effect was evident in the second series, the intervention would be introduced in the third series (e.g., Gabe) and so on. In MBD, the intervention (e.g., cognitive restructuring and reading intervention) is not withdrawn. The demonstration of effect in the first series and two replications or more in additional series (minimum three effects) establish experimental control.

There are but two designs, each of which shows how SCRD can demonstrate a functional relation between the intervention and changes in the dependent variable via one demonstration and two replications of effects. As discussed, SCRD is a flexible methodology used to answer questions surrounding: Is there an effect? These experimental designs (e.g., A-B-A-B, MBD) have been criticized historically, noting issues such as the small number of participants involved, heavy reliance on visual inspection techniques rather than statistical analyses, and the inability to quantify the magnitude of effects (reported as effect sizes). Yet, recent advances have been made enabling data collected from SCRD to be analyzed quantitatively, including the use of effect sizes. SCRDs are feasible, practical, and affordable, holding particular benefit for examining treatment outcomes when working with low-incidence populations and developing (and testing) new strategies, practices, and programs. Furthermore, SCRDs are beneficial for examining effects of more intensive interventions for individuals requiring modification or supplements to interventions initially designed.

Kathleen Lynne Lane and David J. Royer

See also ABA Designs; Experimental Designs; Nonexperimental Designs; Regression Discontinuity Analysis

Further Readings

Gast, D. L., & Ledford, J. R. (Eds.). (2014). Single case research methodology: Applications in special education and behavioral sciences (2nd ed.). New York, NY: Routledge.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design: The what works clearinghouse standards. Remedial and Special Education, 34, 26–38.

Odom, S. L., & Lane, K. L. (2014). The applied science of special education: Quantitative approaches, the questions they address, and how they inform practice. In L. Florian (Ed.), The SAGE handbook of special education (2nd ed., Vol. 1, pp. 369–388). Los Angeles, CA: Sage.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). The role of between-case effect size in conducting, interpreting, and summarizing single-case research. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Hae-Young Kim Hae-Young Kim Kim, Hae-Young

Skewness

Skewness

1528

1530

Skewness

Skewness is a measure of degrees of asymmetry in a distribution. A skewed unimodal distribution can have a longer tail either to the right side or the left side. If the mass of the distribution is concentrated on the left side and the longer tail is on the right side, then the distribution is referred to as skewed to the right or positively skewed; reversely, if the mass of the distribution is concentrated on the right side and the longer tail is on the left side, then it is referred to as skewed to the left or negatively skewed. Generally, the relative positions of central tendency measures appear different according to the direction of skewness. A positively skewed distribution tends to have a relatively larger mean followed by median and mode (order of size: mode < median < mean); a negatively skewed distribution tends to have a relatively larger mode followed by median and mean (order of size: mean < median < mode). However, for multimodal distributions, the meaning of skewness becomes obscure and the relative position of central tendency measures do not follow the aforementioned order.

An example of right skewness is the distribution of income where a small number of persons possess a much higher income compared to the rest of the people. Many variables have distributions skewed to the right when the characteristics are bounded on the left side; counts (>0), durations (>0), or weight/height (lower bound). Left skewness is less common than right skewness. An example of left skewness may be a distribution of scores from a very easy exam where most examinees receive a perfect score. Among formal probability distributions, many distributions are right skewed such as the Poisson distribution, gamma distribution, chi-square distribution, Weibull distribution, and lognormal distribution. The ß distribution and binomial distribution can have and lognormal distribution. The β distribution and binomial distribution can have distributions skewed to either the right side or left side, depending on the values of parameters.

Types of Skewness Measures

Over the years, numerous types of skewness statistics have been suggested. Some were derived using descriptive summary statistics such as central tendency measures or quantiles, whereas others were calculated by using all observed values. When skewness statistics are obtained by summary statistics, the main advantages are simplicity in calculation procedure and an intuitiveness in interpretation. Two popular skewness statistics in this category are the Pearson skewness using relative positions of mean and median and Bowley’s quartile skewness.

Karl Pearson suggested a skewness statistic by standardizing the difference between mean and mode . Later, the sk_1 measure was replaced by three times the standardized difference between mean and median because a more accurate estimation of a population median was possible compared to that of a population mode. The measure was subsequently modified into the standardized difference between mean and median after removing the The meaning of positive or negative Pearson skewness can be easily understood by simple comparison of relative positions of mean and median in right-skewed or left-skewed distributions, respectively.

The Bowley’s quartile skewness is defined as , or more simply , where Q_1, Q_2, and Q_3 are the first quartile, the median, and the third quartile, respectively. The quartile skewness is also easy to interpret. Positive quartile skewness reflects that the right quartile is longer than the left quartile , and negative quartile skewness means the reverse. Bowley’s quartile skewness can be generalized by choosing any two symmetrical points of percentiles, the 100*p*th and 100(1 − *p*th), instead of first and third quartiles. Generally, both the Pearson skewness and Bowley’s quartile skewness are expected to range from −1 to 1 and have a value of zero when the distribution is symmetric, such as the normal distribution.

The Fisher–Pearson skewness is calculated by using all observed values instead of summary statistics. Although the calculation procedure is more complex and the resulting statistic may be strongly affected by extreme values, the Fisher–Pearson skewness produces more precise estimates and is also more powerful in statistical testing compared to the skewness by summary statistics. The Fisher–Pearson skewness is a formal mathematical skewness statistic that is adopted by most statistical packages. The Fisher–Pearson skewness based on moments is expressed as the expectation of the third standardized moment . Generally, an adjusted form of the Fisher–Pearson skewness estimate , which corrects bias related to small sample size, is currently adopted as a sample skewness estimator by most statistical software packages such as Excel, SPSS, STATA, and SAS. The standard error formula for the adjusted Fisher–Pearson skewness estimate is given as . The value of standard error approaches for large sample sizes (e.g., n > 50). Using the estimate of standard error, we can construct interval estimations such as the 90% or 95% confidence interval for population skewness.

Statistical Testing and Other Issues

Symmetric unimodal distributions such as normal distributions have approximately zero skewness. Null hypothesis of zero population skewness is testable using the *z* test statistic under assumption that the sampling distribution is approximately normal for large sample sizes. The *z* test statistic is expressed as skewness statistic divided by the standard error of the skewness statistic () and can be used to test the null hypothesis of zero population skewness. As the standard error decreases, the sample size increases, and calculated *z* values for larger samples tend to be smaller than those for smaller samples. Therefore, the *z* test may be more liberal in rejecting null hypothesis in a large sample compared to that in a small sample. A large absolute skewness value larger than two is considered the indication of nonsymmetric distributions. Various transformation methods such as log or square root transformations can be considered to ameliorate the asymmetry.

Still, most types of skewness have been criticized for their imperfectness. When sample size is small, most skewness statistics do not provide an adequate power in detecting mildly asymmetric distribution. Moreover, sometimes, they may produce inconsistent results. Interested scholars are trying to explore alternative superior skewness statistics or making efforts on the quality improvement of existing ones.

Hae-Young Kim

See also Distributions; Moments of a Distribution; Normal Distribution

Further Readings

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? Journal of Statistics Education, 19(2).

Tabor, J. (2010). Investigating the investigative task: Testing for skewness. An investigation of different test statistics and their power to detect skewness. Journal of Statistics Education, 18(2).

West, S. G., Finch, J. F., Curran, P. J. (1995). Structural equation models with nonnormal variables; Problems and remedies. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues and applications (pp. 56–75). Newbery Park, CA: Sage.

Nancy N. Boyles Nancy N. Boyles Boyles, Nancy N.

Smarter Balanced Assessment Consortium

Smarter balanced assessment consortium

1530

1531

Smarter Balanced Assessment Consortium

Smarter Balanced Assessment Consortium is an assessment consortium supported by 15 states, the U.S. Virgin Islands, and the Bureau of Indian Education that created an online assessment system aligned to the Common Core State Standards. The assessment system measures performance on these standards in both English language arts/literacy and mathematics in Grades 3–8 and high school. The summative end-of-year test determines progress toward college and career readiness by evaluating students in two ways: proficiency (how much the student knows at the end of the year) and growth (how much the student has improved since the previous year). The summative assessment consists of two parts: a computer adaptive test and a performance task.

The computer adaptive feature is a key component of the Smarter Balanced assessment system. Questions get harder when students answer correctly and easier when they answer incorrectly. This provides greater accuracy in measuring students' strengths and needs and gives teachers and parents more meaningful information to guide instruction. The ELA/literacy summative test measures reading, writing, listening/speaking, and research. The math summative assessment addresses concepts and procedures, problem solving, modeling and data analysis, and communicating reasoning. Items measure competence in these areas across four depths of knowledge: recall, skills and concepts, strategic reasoning, and extended thinking.

Stimuli for both math and ELA/literacy include a variety of information forms such as readings, video clips, and data, as well as an assignment or problem situation. Item types include selected-response items that prompt students to choose one or more responses from a set of options; technology-enhanced items choose one or more responses from a set of options; technology-enhanced items that could include drag and drop, editing, and drawing; constructed-response questions that ask for a short written answer or numerical response; and performance tasks that measure a student’s ability to think critically and creatively and solve the real-world problems. The performance task is a collection of questions and activities that are connected to a single theme and set of stimuli. Although the performance task is taken on the computer, it is not computer adaptive.

Beyond the end-of-year summative assessment, the Smarter Balanced Assessment Consortium supports standards-based learning in two additional ways. The Smarter Balanced Digital Library is a collection of instructional and professional learning resources that help educators apply principles of formative assessment to improve teaching and learning. There are resources to clarify learning goals, elicit evidence, interpret evidence, and act on evidence. The intent is to use these tools during the year to adjust instruction based on feedback from curriculum-based tasks.

The consortium also offers interim assessments so teachers can monitor students’ progress periodically. These assessments are optional, used at the discretion of districts, and include item types and formats that are the same as those found on summative assessments. Interim assessments are delivered online and provide results that can be examined in relation to grade-level benchmarks. Beyond their benefit in determining proficiency and growth, these interim tests familiarize students with item types as well as the rigor expected by new standards-based assessments.

Nancy N. Boyles

See also Achievement Tests; Common Core State Standards; Formative Assessment; Partnership for Assessment of Readiness for College and Careers; Summative Assessment

Further Readings

Smarter Balanced Assessment Consortium. (2016, November 10). ELA/literacy summative assessment blueprint. Retrieved from https://portal.smarterbalanced.org/library/en/elaliteracy-summative-assessment-blueprint.pdf

Smarter Balanced Assessment Consortium. (n.d.). The formative assessment process. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/09/Formative-Assessment-Process.pdf

Smarter Balanced Assessment Consortium. (n.d.). Sample questions. Retrieved from https://www.smarterbalanced.org/assessments/sample-questions/

Smarter Balanced Assessment Consortium. (n.d.). Smarter assessments. Retrieved from http://www.smarterbalanced.org/assessments/

Websites

Smarter Balanced Assessment Consortium: http://www.smarterbalanced.org/

Toni Crouse Toni Crouse Crouse, Toni

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Snowball Sampling Snowball sampling

1531

1532

Snowball Sampling

Snowball sampling is a sampling method used by researchers to generate a pool of participants for a research study through referrals made by individuals who share a particular characteristic of research interest with the target population. It is also referred to as chain sampling or chain referral sampling.

In snowball sampling, a subject from an initial sample group is asked by researchers to recommend individuals to act as future participants. The prompting for recommendations may take the form of an informal question, such as “Who are your best friends?” The subjects who are recommended by these individuals and agree to participate in the research are then considered to be the first wave of participants. The subjects in the first wave will be asked to make their own referrals of future participants. This second group of referrals will make up the second wave of research participants. This method may be repeated over and over again, thereby continuing the cycle, and just like a snowball rolling down a hill, the sample gets bigger and bigger.

Although the snowball sampling technique is applicable to a variety of study designs, it has been utilized most frequently in qualitative sociological research. In particular, this method has been employed in cases where the research focused on a sensitive issue, such as individuals who are HIV-positive, or when target subjects were difficult to locate in the general population. For example, snowball sampling has been particularly useful in research concerning deviant behavior, such as with participants who may be drug users or prostitutes.

Researchers’ use of the snowball sampling method has several unique advantages. First, due to the established familiarity between participants and those they refer, valuable social and interactional knowledge may be generated. Participants are observed within the context of their naturally formed relationships and social networks. Consequently, it may be easier to build rapport with referred participants, as researchers have already spoken with a friend, relative, or colleague at an earlier time.

The snowball sampling method also allows for the collection of both group and individual qualitative data simultaneously. For example, information may be gathered on group movements and routes of travel, in addition to individual backgrounds and histories. Utilization of the snowball technique allows researchers to overcome cultural boundaries such as lower literacy levels and language barriers, which may traditionally affect a participant’s likelihood of volunteering for a study.

Despite these advantages, there are also distinct limitations to snowball sampling. Due to the lack of randomization across study phases, data collected from participants cannot be considered generalizable to the target population as a whole. Definitive conclusions regarding the population may be inherently biased. For example, individuals from the target group that are isolated from others may be less likely to be referred to researchers, thus excluding a subset of the population. This technique also introduces the potential for a lack of confidentiality across participants. Researchers may be asking participants to disclose personal or sensitive information about others related to target group membership. Referred participants are then faced with the decision of whether to disclose their eligibility status to the researcher.

Toni Crouse and Patricia A. Lowe

See also Qualitative Research Methods; Selection Bias; Simple Random Sampling; Stratified Random Sampling; Systematic Sampling

Further Readings

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. Sociological Methods & Research, 10(2), 141–163.

Goodman, L. A. (1961). Snowball sampling. The Annals of Mathematical Statistics, 148–170.

Kyrsten M. Costlow Kyrsten M. Costlow Costlow, Kyrsten M.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Social Cognitive Theory Social cognitive theory

1532

1535

Social Cognitive Theory

Social cognitive theory (SCT) is a psychological model of behavior that asserts that learning occurs through observation within a social context. According to SCT, people observe the behaviors of others and the resulting consequences and use those observations to inform their own behaviors. The theory emerged largely from the work of Albert Bandura. According to Bandura, SCT is founded on reciprocal triadic relations among personal, behavioral, and environmental factors. The major theoretical components of SCT include modeling, outcome expectations, self-efficacy, goal setting, and self-regulation. SCT has been applied to a variety of disciplines such as psychology, education, business, and health communication. Within education, SCT has been used to understand classroom learning, student motivation, and academic achievement. This entry describes the history of SCT, its core components, and its applications within education.

Albert Bandura and the History of SCT

In 1941, Neal Miller and John Dollard introduced social learning and imitation theory, which proposed that humans learn behavior through observation and modeling. To test this theory, Bandura and his colleagues exposed children to models either playing aggressively with a Bobo doll or nonaggressively with tinker toys. The children were then placed in a room with a variety of toys, including a Bobo doll, which could be used for aggressive or nonaggressive play. Children who were exposed to aggressive models showed more imitative aggression than children in the nonaggressive and control groups. In a follow-up study, Bandura found that these results held when aggressive models were presented on film. These studies provided early evidence of the importance of imitation and modeling in learning.

In 1977, Bandura used the Bobo doll studies to expand on Miller and Dollard’s original theory and create his own social learning theory. To emphasize the role of cognition in encoding behavior, Bandura renamed the theory from social learning to social cognitive in his 1986 book *Social Foundations of Thought and Action: A Social Cognitive Theory*. There, Bandura proposed reciprocal triadic relations among personal, behavioral, and environmental factors as the basis of SCT. According to this triadic reciprocality model, an individual’s personal attributes, overt behaviors, and external environment affect one another bidirectionally. By providing the individual with agency, SCT opposes earlier behaviorist views of the learner as a passive responder to the environment.

Core Components of SCT

As a model of behavior, SCT incorporates multiple theoretical components into its larger framework. These core components include modeling, outcome expectations, self-efficacy, goal setting, and self-regulation.

Modeling

SCT is built on the premise that people learn by observing and imitating the behaviors of others in the environment—a process known as modeling. As Bandura demonstrated in his Bobo doll experiments, the models used for learning can be real people or media images. Modeling may also occur indirectly through verbal and written material. Modeling can be used both to promote and to inhibit behaviors. For example, students may imitate actions they see praised in the classroom, such as raising hands to answer questions, and inhibit actions they see reprimanded, such as talking out of turn.

According to SCT, modeling relies on four processes: attention, retention, production, and motivation. Attention is the first step to modeling because an individual must focus on the model’s behavior and consequences to learn. Retention is then used to transform the observed behavior into a symbol to be stored and accessed in the future. This symbol is then accessed in the third step, production, when the individual would like to reproduce the observed behavior. Motivation makes up the final step where individuals decide whether they would like to repeat the action in the future based on the responses they received.

Outcome Expectations

Outcome expectations refer to the responses that individuals expect to get from their behaviors, based on their previous behaviors and the observed behaviors of others. For example, a student who is consistently given detention for fighting in the classroom will expect this same consequence in the future. Outcome expectations shape people’s decisions about which behaviors to continue and which to inhibit. While observing the behavior and resulting consequences of a model, an observer will usually expect similar but not identical outcomes. Students who observe their peer scolded by the teacher for shouting in class may therefore expect a similarly negative consequence. According to Victor Vroom’s expectancy theory, individuals weigh the probability and desirability of their outcome expectations to inform future behavior.

Self-Efficacy

In conjunction with outcome expectations, self-efficacy also motivates future behavior. Self-efficacy refers to the success with which individuals believe that they can perform a particular skill. Individuals with greater self-efficacy are more confident in their abilities to master a given task than individuals with lower self-efficacy. As such, observers are more likely to imitate a model’s behavior if they have a high level of self-efficacy. Therefore, self-efficacy can influence outcome expectations such that individuals with greater self-efficacy will expect more positive performance outcomes. Although based on perceived abilities, self-efficacy beliefs have been found to have weak correlations with objective measures of ability. According to Bandura, people’s self-efficacy beliefs are more important than “objective truth” in determining behavior and motivation.

Individuals base their self-efficacy beliefs on four factors: mastery experience, vicarious experience, social persuasions, and somatic and emotional states. Mastery experience refers to increases in self-efficacy after positive behavioral outcomes and decreases after negative outcomes. Although less influential than mastery experience, vicarious experiences of others can also influence an individual’s self-efficacy. When observers see models with similar characteristics to themselves successfully performing a task, these observers may gain confidence in their own abilities and increase their self-efficacy. Vicarious gain confidence in their own abilities and increase their self-efficacy. Vicarious experiences are particularly salient when people are unsure of their abilities or have no previous experience in a given task.

Self-efficacies are also influenced by social persuasion and the person’s somatic and emotional states. Social persuasion refers to external encouragements or discouragements of behavior. The influence of social persuasion may be positive or negative. For example, students who are praised for their reading ability will increase their self-efficacy in this domain, but students who are discouraged by poor grades in this subject will decrease their self-efficacy. Somatic and emotional states brought on by a given task may also influence self-efficacy. When individuals are highly anxious in the face of a difficult task, they are more likely to lose confidence and decrease their self-efficacy. Relaxing and promoting physical and emotional well-being before a novel task can therefore be used to increase self-efficacy.

Goal Setting

Through the process of goal setting, individuals determine the expected and desired outcomes of their behavior. Outcome expectations and self-efficacy shape goal setting because individuals set goals based on the outcomes they expect and desire as well as on the confidence they have in achieving set goals. Goal setting reflects a central tenet of SCT that individuals are active agents in their environments and can determine their own behaviors. Goal setting enables individuals to determine standards of self-performance and to set a plan for self-regulation.

Self-Regulation

Closely related to goal setting, self-regulation refers to the ability to manage and change one’s own behaviors to reach particular outcomes. Bandura splits self-regulation into three subprocesses: self-observation, self-judgment, and self-reaction. Individuals must first monitor their behaviors through self-observation and evaluate the consequences of their behaviors through self-judgment. After forming this evaluation, individuals can respond by modifying their behaviors through the process of self-reaction. Like goal setting, self-regulation reflects SCT’s focus on human agency and is connected to other core concepts of SCT, such as outcome expectations and self-efficacy.

Applications in Education

SCT can be applied to education to improve student learning in the classroom. The triadic reciprocality tenet of SCT allows teachers to intervene in student learning through personal, behavioral, or environmental factors. Teachers can target personal factors, such as students’ emotional states and self-efficacies; behavioral factors, such as academic skills and student self-regulation; or environmental factors, such as the classroom setting. Teachers can act as models in the classroom to promote observational learning of targeted skills.

To successfully apply SCT to the classroom, teachers must model the positive behaviors they would like to bring out in their students. They should also employ other types of positive models, such as peers, parents, and the media. Because students’ classmates act as models for learning, teachers must administer consistent rewards and consequences to teach students responses to expect from good and bad behavior. Because Bandura’s Bobo doll studies showed media models to be influential, teachers may use educational programs in the classroom to encourage observational learning as well. Teachers should also support each of the four processes within modeling: attention, retention, production, and motivation. They can try to increase student attention by making class material interesting, engaging, and personally relevant. Student retention can be improved through mnemonics, learning visuals, repetition of key concepts, and other such strategies. Finally, teachers can encourage students to practice the behaviors and skills being taught and should use rewards and punishments to motivate students in the classroom.

Using modeling in this way allows teachers to create appropriate outcome expectations for their students. Consistent rewards and punishments will teach students what to expect from certain types of behavior in the classroom. Moreover, teachers should work to assure their students that rewards and punishments are meaningful and important. For example, consistent discouragement may teach students that acting out in the classroom results in disciplinary action, such as detention or grade reductions. However, students will only be motivated to inhibit this type of behavior, if they recognize the value of these consequences. If students do not care about their grades, reducing their grade as punishment may not influence their behavior. Teachers must therefore teach students the relevance of their educations to their futures. Applying lessons to the real-world contexts can help students to realize the curriculum’s importance in their own lives.

Teachers can also focus on increasing students’ self-efficacy to enhance students’ motivation in the classroom. Lessons that start with basic foundations and build stepwise from there can avoid discouraging students with overly difficult curricula. If the class masters the basic foundations of a topic first, they can then move on to more advanced lessons with increased self-efficacies. Students’ self-efficacies can improve when they master skills on their own or when they watch peers complete tasks successfully. Students should be given individual attention whenever possible and encouraged verbally when they succeed or show effort. Teachers can also promote relaxation and positive emotional states in their students when students become overwhelmed in the face of a difficult task.

Finally, teachers can use goal setting and self-regulation to improve student learning. Teachers can use the curriculum to set challenging yet attainable goals for their students. Initiatives such as the Common Core State Standards can assist in disseminating these specific, age-appropriate student goals in K–12 education. Teachers can also support self-regulation by encouraging self-observation, self-judgment, and self-reaction. Teachers can use self-assessment methods to teach students to monitor and evaluate their academic behavior. Teachers may also guide students through self-reaction by rewarding successes and modifying strategies in the face of failure. SCT strategies may be used to modify student behavior in regard to academic performance as well as socioemotional functioning. SCT and its core components therefore have important implications for education and improve student learning and academic achievement.

Kyrsten M. Costlow and Marc H. Bornstein

See also Attention; Goals and Objectives; Learning Theories; Motivation; Operant Conditioning; Self-Efficacy; Self-Regulation

Further Readings

Bandura, A. (1976). Social learning theory. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1998). Self-efficacy: The exercise of control. New York, NY: Freeman.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. Journal of Abnormal and Social Psychology, 63(3), 575–582. doi:10.1037/h0045925

Bandura, A., Ross, D., & Ross, S. A. (1963). Imitation of film-mediated aggressive models. The Journal of Abnormal and Social Psychology, 66(1), 3–11. doi:10.1037/h0048687

Denler, H., Wolters, C., & Benzon, M. (2014, January 28). Social cognitive theory. Retrieved May 6, 2016, from http://www.education.com/reference/article/social-cognitive-theory/

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. Journal of Vocational Behavior, 45(1), 79–122.

Pajares, F. (2002). Overview of social cognitive theory and of self-efficacy. Retrieved May 5, 2016, from http://www.emory.edu/EDUCATION/mfp/eff.html

Ivar Krumpal Ivar Krumpal Krumpal, Ivar

Social Desirability

Social desirability

1535

1536

Social Desirability

Social desirability can be conceptualized as an individual's constant need for social approval and impression management in social interactions. Social desirability has its origin in shared social norms. To conform to social norms, research subjects often present themselves in a positive light and mask their true behaviors or intentions in an empirical study. This entry describes the problem of social desirability and some remedies in the research practice.

In empirical social and educational research, especially in surveys using question-and-answer data collection methodology, it is a challenge to accurately measure private or norm-violating issues (e.g., sexual behavior, income, health-related issues, illicit drug use, delinquency or unsocial opinions such as racism) because respondents may choose to deliberately misreport on such sensitive topics and adjust their answers in accordance with perceived social norms. More specifically, social desirability refers to the respondents' tendency to self-report socially desirable characteristics (systematic overreporting) and to deny undesirable ones (systematic underreporting). Research subjects give socially desirable answers due to self-presentation concerns to avoid negative emotions of embarrassment in social interactions or because of fear of legal sanctions in consequence of self-reporting illegal behavior. Such response behavior introduces serious bias to the measurement of the sensitive characteristics and lowers the overall data quality. Another severe problem is nonresponse. Some respondents may refuse to answer the sensitive questions at all. If nonresponse is systematically related to the behaviors or attitudes of interest, estimates will be distorted.

Cumulative evidence indicates that the use of appropriate design features and data collection methods could reduce the respondents' data protection concerns, improve the validity of measurements of the sensitive characteristics, and achieve better data quality in empirical studies in which social desirability bias is a potential problem. In this context, social scientists, survey designers, and educational researchers could use confidentiality and data protection assurances to decrease the respondents' concerns in admitting to some sensitive behavior. Furthermore, they could try to reduce interviewer and bystander effects or use mixed-mode designs or nonreactive methods to increase the anonymity of the data collection situation. If respondents trust their privacy protection, they will be more likely to reveal embarrassing or self-discrediting information. Finally, special questioning techniques have been developed to anonymize answers at the individual level via a random mechanism. Among these techniques, the randomized response technique, the crosswise model, and the item count technique are the most prominent ones. Advanced statistical methods have been developed to analyze multivariate relationships between response variables generated via these methods and background covariates.

Ivar Krumpal

See also Confidentiality; Data; Hawthorne Effect; Mixed Methods Research; Survey Methods; Surveys; True Score; Validity

Further Readings

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. Public Opinion Quarterly, 76, 32–49. Retrieved from https://doi.org/10.1093/poq/nfr036

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. Quality & Quantity, 47, 2025–2047. doi:10.1007/s11135–011-9640-9

Preisendörfer, P., & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. Public Opinion Quarterly, 78, 126–146.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. Psychological Bulletin, 133, 859–883.

Annette Woods Annette Woods Woods, Annette

Social Justice

Social justice

1536

1539

Social Justice

The term *social justice* refers to moves to encourage and achieve equity, in a framework of human rights and recognition of diversity. However, there is no consensus about a definition for the term, and it is often put to use without a serious engagement with underpinning philosophies or standpoints. As a social practice, social justice relates to attempts to provide access to quality services such as education and health to all, regardless of their gender, race, religion, social standing, class, or social, cultural, and language practices. Discrepancies that arise from the advantage and access available to some groups of society and not others cannot be understood to result only from force—a situation that Michel Foucault would discuss as violence. It is as likely that inequities result from the structural and institutional mechanisms of society that are entrenched in the political and public routines of systems. It is these struggles for equity that equate to seeking a socially just society. This push for parity has often seen social justice advocates positioned in opposition to the established institutions of society—whereby when one group asks for more, the assumption is that others must somehow have less. However, while the foundation of social justice is that everyone deserves a fair and equitable portion of what is available, social justice is not just about the shifting of resources toward a more fair distribution. This entry discusses social justice as a multidimensional concept.

The Difference Between Equality and Equity

Social justice is about fairness; however, this should not lead to thinking that everyone should be treated the same or that social justice will have been achieved if everybody is given the same resources. There is a difference between equality and equity. In an equal society, everyone would be able to access the same services and approaches as everyone else. In its simplest terms, this may provide the same inputs for all, but it will never produce the same outputs for all. So social justice is about equity rather than equality, in that all persons should receive what they need to participate and benefit equally in society. The emancipatory dimension of social justice comes about in the drive to somehow readjust society so that the privilege that certain dominant groups experience over others is recalibrated toward a more fair and just approach to access and quality.

Distribution and Recognition

Early understandings of social justice were based in the struggles of social groups or classes of people and the basic connection of workers to the economy. The writings of Karl Marx from the mid-1800s have formed the basis of understanding social justice as a fight against social systems such as capitalism and the unequal distribution of financial resources across a classed society. It is within these ideals that much of the various struggles for equality, including those focused on gender and class battles, have been based on until the latter decades of the 1900s. As an example, if oversimplified at best, women's positioning as second to men can be explained as a distributive issue. Such an argument would claim that the fact that women in many societies have traditionally taken up the responsibility for unpaid labor in the home while men have engaged in paid work has set up a situation of structural division based on gender. This has led to a situation whereby women are disadvantaged through inequitable divisions of labor and lack of economic control.

Commentators often claim the 1990s as the period of the cultural turn—when the focus of political debates on such issues as gender or race equality shifted from foregrounding labor and its connection to the economy, to issues of identity and representation. In terms of social justice and the fight for equality, this led to a shift from focusing on redistribution to redress social disadvantage, to calls for recognition of difference and diversity. By this way of thinking, denial of full citizen rights to marginalized groups in society results from institutional practices that center certain groups because of gender, race, and class. This centering relegates those who represent the binary opposite of such categories to "other." So from a recognitive way of thinking, social injustice results from dominant groups marginalizing the interests and cultures of others from social dominant groups marginalizing the interests and cultures of others from social systems and discourses.

Through the first 2 decades of the 21st century, Nancy Fraser has consistently raised problematics with understanding social justice as *either* a matter of redistribution *or* recognition. She has led a theoretical move to understand social justice as a multidimensional concept, and her work is at the foundation of recent understandings of social justice. Her arguments not only bring to the fore the complexities of creating a socially just society but also help to explain the difficulties of actually achieving a more just society regardless of either economic or cultural will. Taking a socially just approach to our social institutions requires *both* redistribution of resources to create a more equitable playing ground, *and* recognition of difference and diversity. This allows for social categories to be understood as both political and cultural, about economies and social interaction. So solutions to disadvantage can no longer be about redistribution *or* recognition but must relate to both. As Fraser has pointed out, maldistribution and misrecognition are at the foundation of injustice and inequality.

To develop these ideas further, it is interesting to consider what these dimensions of socially just practice might look like within a social institution such as education. At a conceptual level, redistributive justice relates to initiatives to redistribute funds, initiatives, and policies toward addressing the financial disadvantage of some groups as compared to others. However, once this concept is taken to a practical field, such as education, it becomes increasingly clear that redistributive moves are about much more than economic or material resources. As important in this challenge is the need to (re)distribute access to the valued knowledge, languages, ways of producing texts and communication, ideas, and understandings more equitably. Children being schooled in communities of high poverty need access to the basic skills of literacy as well as the means to think critically about how they and their communities are positioned in text. They require access to conventional ways of using the valued languages of the dominant group if they are to have access to jobs and full citizenship. The distribution of appropriate economic resources to schools in communities of high poverty—while important—is only part of the redistributive solution. Recognitive moves in education relate to ensuring that all students have the opportunity to be recognizable as a member of society in the curriculum, and classroom spaces of the schools they attend; in the images and texts used to represent school and used in school; and in the policy, material, and human resources to which they are exposed and interact. It is about questioning the institutional systems that prevent some members of society from participating in the same ways and to the same extent as others.

Fraser develops this concept of participation. A society can be judged on the extent to which all members are able to participate in its actions, routines, and benefits. She draws on the term *parity of participation* to represent at the most general level what the aim of social justice should be, and this highlights her focus on social interaction as key to our political and social lives.

Participation and Representation

Many commentators have positioned the marketization of social institutions as implicated in social injustice and inequity for marginalized groups in society. In fast capitalism, the relationships between distributive and recognitive approaches to social justice are complicated. Once social injustice is conceptualized as resulting from cultural and social processes and systems, it risks being decoupled from the practical concerns of equitable distribution of power and wealth. On the other hand, once political approaches to redress disadvantage are narrowed to financial and labor concerns, a critical understanding of institutionalized domination and oppression can be removed from key social debates.

Shifts toward a more socially just society require shifts in the taken for granted rules and procedures of social institutions, and this is likely to occur only if a more diverse range of perspectives are included in decision making and political representation. These ideas were put forward in the work of Iris Marion Young in the early 1990s and have led to a more recent shift by Fraser and many others to think about social justice as including dimensions of distribution, recognition, and representation. This shift is intended to ensure that the political dimension of social justice is highlighted. So a notion of representation takes up the call to ensure participation and representation of a diverse range of citizens in the governance of institutions and political systems. Representation requires that those groups who may be the target of social justice moves are shifted from a positioning as “problems” to having some voice in the decision-making processes of citizenship. With these rights comes civic responsibility.

In order for social institutions such as education and health to come to a position where representative justice is a dimension of socially just practice, there is a need for critical, analytic reflection on why the institutions are currently unjust and how this situation is perpetuated through routines, policies, and practices that are taken for granted. This moves conceptualizations of social justice into a transformative agenda, led by scholars such as Sonia Nieto, Marilyn Cochran-Smith, and many others. When social justice is framed as a multidimensional concept, simple solutions to injustice are not plausible. A commitment to social transformation and recalibrating privilege will be required.

Annette Woods

See also Gender and Testing; Outcomes; Race to the Top; School Leadership

Further Readings

Cochran-Smith, M. (2003). The multiple meanings of multicultural teacher education: A conceptual framework. Teacher Education Quarterly, 30(2), 7–26.

Connell, R. (2002). Making the difference, then and now. Discourse: Studies in the Cultural Politics of Education, 23(3), 319–327. Retrieved from http://dx.doi.org/10.1080/0159630022000029812

Fraser, N. (1997). Justice interrupts: Critical reflections on the “postsocial” condition. New York, NY: Routledge.

Fraser, N. (2014). Transnationalizing the public sphere. Cambridge, UK: Polity.

Fraser, N., & Honnneth, A. (2003). Redistribution or recognition?: A political-philosophical exchange. New York, NY: Verso.

Luke, A. (2010). Documenting reproduction and inequality: Revisiting Jean Anyon’s “Social class and school knowledge.” Curriculum Inquiry, 40(1), 167–182. doi:10.1111/j.1467–873X.2009.00474.x

Nieto, S. (2000). Placing equity front and center: Some thoughts on transforming teacher education for a new century. Journal of Teacher Education, 51(3), 180–187. Retrieved from https://doi.org/10.1177/0022487100051003004

Young, I. M. (1990). Justice and the politics of difference. NJ: Princeton University Press.

Lina Goldenberg Lina Goldenberg Goldenberg, Lina

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Social Learning

Social learning

1539

1541

Social Learning

Social learning is learning through the experiences or observations of others. Behaviors are influenced by environmental antecedents and consequences, which either increase or decrease the chance of the behaviors occurring again. Through social environments, which may include peers and teachers in school, family, coworkers, and supervisors, humans are able to learn various behaviors and in turn shape their own beliefs, morals, and ideas about the world. Influenced by B. F. Skinner's theory of behaviorism, Albert Bandura coined the term *social learning theory*, which was utilized by many researchers such as Richard Walters, Robert Sears, and others, to study how humans learn through their social environment. This entry covers social learning theory, its components, and briefly reviews how the theory has been utilized in research.

Learning Through Experience

A punisher or reinforcer immediately following a certain behavior unconsciously shapes that behavior. According to Skinner and later confirmed by Bandura, when an individual is continuously and constantly punished for a behavior, the likelihood of engaging in that behavior decreases over time. In contrast, when the behavior is rewarded, the behavior is more likely to continue. According to Bandura, due to this mechanism, people develop hypotheses about what types of behaviors are more likely to be successful, and these hypotheses guide future behaviors. Successful hypotheses are unconsciously strengthened. Although behavior is controlled by external stimuli, it is not always controlled by immediate outcomes. For example, individuals are able to anticipate based on prior knowledge or experiences and, as a result, are unlikely to wait until they are in a car accident to buy car insurance; thus, the potential outcome serves as a motivator for current behaviors. Learning is difficult when individuals are unaware that they are being reinforced for a behavior. However, when made aware of the wanted behavior, individuals are able to discern what is wanted of them, which increases the probability of learning.

Learning Through Observation or Modeling

Although learning can occur through reinforcement or punishment, the world is full of complex concepts (e.g., religion, morals, language, cultural norms) that are difficult to teach through punishment or reinforcement. For example, if an individual has not been exposed to music, it would be difficult for the individual to associate a certain note with the corresponding sound without another individual modeling this association. Therefore, according to social learning, behavior can be learned through viewing examples of other's behavior, and reinforcement is considered an unnecessary facilitator for learning. Furthermore, learning through modeling or observation prevents individuals from making mistakes and learning through trial and error.

Because learning occurs through the observation of others, the most likely behaviors to be learned are those of individuals whom we are frequently around. However, learning from models is dependent on certain variables. Specifically, various individuals within a social group possess different modeling power. For example, if an individual is perceived as important or interesting, the individual will be observed more closely. The model's status can be judged by various factors, such as individual's appeal, speech, or knowledge. Learning through observation can also occur when people watch television. Although models on television are not physically present, if they are considered to be important or interesting, learning from their behaviors is likely to occur. This theory has been used to examine aggressive behavior through modeling. When individuals are frequently around aggressive people, they are more likely to learn that type of behavior. However, it is not enough to be exposed to modeled activities. Various factors, such as anticipation, influence whether modeled activities will be learned.

To function in the world, humans need to be able to anticipate the outcomes of various situations, and information about these outcomes is acquired from environmental cues. The anticipatory response is very important for safety and protection against threats. Knowing that the modeled behavior will result in positive or desired outcomes will increase the attentiveness to the model and will, therefore, increase the retention of the learned behavior. As a result, the more attention paid to the model, the more likely learning of the behavior will occur. Once the behavior has been anticipated and attended to, the individual needs to be able to retain what was learned in order to utilize it in the future. Those who achieved positive and better outcomes in the past from observing others are more likely to be responsive to modeling influences.

Types of Modeling

There are various types of modeling according to social learning theory, one of which is *verbal modeling*. This is the least difficult type of modeling to do, as it is easier to convey more information of varying complexity through words than through behaviors. An additional type of modeling is *symbolic modeling*, which can be acquired through television, social media, films, or other pictorial displays (e.g., magazines). Both children and adults can acquire attitudes, various behaviors, and emotional responses through symbolic modeling.

Although the means of the modeled behavior is not contingent on the behavior that is being learned, different types of modeling may not be equally effective for various types of behaviors and people. For example, a demonstration of how to assemble a desk may be more effective than verbally describing it. Furthermore, various forms of modeling may be more interesting than others. For instance, watching something on television may be more interesting for a child than hearing a lecture about the same concept. The more relevant and interesting the means of modeling are, the more likely the information will be encoded and used in the future.

Vicarious Learning

Although learning can occur through reinforcement and observing others, behavior is not always controlled by direct reinforcements. In social situations, individuals frequently have the opportunity to observe others being rewarded or punished for their actions. Thus, behavior of the observer can be changed based on the consequences of the observed behavior. Negative consequences experienced by the observed individual will reduce the likelihood that the observer will behave in a similar manner. Learning from the consequences of others is termed v*icarious punishment* and has been most commonly studied with aggression. When aggressive behavior is observed being punished, it is less likely to be imitated by the observer.

Reinforcement Types

There are various frequencies of reinforcements that shape behavior, one being *intermittent reinforcement*. In this type of reinforcement, an individual is reinforced after a certain time has elapsed. In contrast, if the reinforcement occurs regardless of the individual's behavior, the motivation to produce the desired behavior will be low. For example, if an individual is getting paid regardless of the amount of work produced, then the individual is likely to produce less work. This particular type of reinforcement is called a *fixed-interval schedule of reinforcement*. On the other hand, if the individual is paid only after completion of a certain amount of work, the individual is likely to complete more work to receive the reinforcement. This type of reinforcement is called *fixed-ratio reinforcement*. However, employers typically do not check on their employee's work at the exact time each day; therefore, individuals are generally reinforced for their work on a *variable-ratio schedule of reinforcement*. In this type of reinforcement, the time of reinforcement varies, but the behavioral outcome expectation remains the same. This type of reinforcement produces the highest and most consistent performance.

In everyday life, individuals behave in varied manners, and their behavior is not always reinforced. Behavior may be applauded by one individual, whereas it may be ignored or frowned upon by another. Therefore, variable-ratio schedule of reinforcement is the most prevalent type of reinforcement in social situations. External reinforcement is most powerful when it matches personal beliefs. Therefore, people usually engage and associate with those individuals who fit that criteria and provide social support for their self-evaluation. Reinforcement varies across age-groups and behaviors. For example, children often require frequent reinforcement during initial stages of skill development; however, as they progress further in their skill development, they are more likely to be naturally reinforced (e.g., reading). Although external reinforcement is important in some situations, the highest level of autonomy is achieved when individuals are able to regulate their own behavior, and the consequences are intrinsically rewarding.

The Cycle of Self and Environment

Although external reinforcements are important, according to social learning theory, control eventually shifts from external to internal sources. Much of our behavior is self-controlled and self-reinforced, which is done when we evaluate ourselves (e.g., editing while writing a research paper) and produce an external outcome (e.g., final version of the research paper). Self-evaluation and self-reinforcement are highly related to environmental or societal standards. For example, if an individual's society does not value higher education, it is unlikely that the person will pursue such a path. Therefore, while self-reinforcement is dependent on the individual's values and beliefs, the individual's values are socially dependent.

Moreover, the reaction of the social environment to the individual's behavior shapes the person's self-esteem, self-concept, and self-efficacy. Based on social learning theory, self-esteem is defined as the difference between an individual's behavior and what standards the person selected to indicate merit. Self-concept is defined as the individuals' evaluation of their own behavior. When the evaluation is overly negative, the individual is deemed to have low self-concept. Self-efficacy is defined as the individuals' beliefs about their own abilities. The individual's beliefs will determine how much effort will be put into a certain behavior. When individuals believe that they will not perform well, their emotional state may distract them from performing the activity at the desired level. Therefore, individuals do have their own ideas and beliefs; however, their social environment shapes those ideas and beliefs.

Contributions of Social Learning

Human behavior and learning are complex concepts; however, social learning theory has provided the foundation to study various important behaviors and experiences that humans encounter. Social learning has been used to study the development of aggressive and criminal behavior, given one's social environment. Research on aggressive behavior and familial influences on child behavior has resulted in the development of various interventions for children with behavioral problems using the principles of social learning. Development of social skills, social relationships, and the internalization of culture have also been investigated using social learning theory. Furthermore, because our social environment shapes our beliefs and values, social learning has been used to evaluate how the social environment shapes personality.

Lina Goldenberg and Patricia A. Lowe

See also Behaviorism; Learning Theories; Reinforcement; Self-Efficacy; Social Cognitive Theory

Further Readings

Bandura, A. (1973). Aggression: A social learning analysis. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice Hall.

Grusec, J. E. (1992). Social learning theory and developmental psychology: The legacies of Robert Sears and Albert Bandura. Developmental Psychology, 28, 776–786. doi:10.1037/0012–1649.28.5.776

Jennings, W., & Akers, R. L. (2011). Social learning theory. In B. Clifton (Ed.), The Routledge handbook of deviant behavior (pp. 106–113). New York, NY: Routledge.

Nangle, D. W., Erdley, C. A., Adrian, M., & Fales, J. (2010). A conceptual basis in social learning theory. In D. Nangle, D. Hansen, C. Erdley, & P. Norton (Eds.), Practitioner’s guide to empirically based measures of social skills (pp. 37–48). New York, NY: Springer.

Patricia M. Noonan Patricia M. Noonan Noonan, Patricia M.

Amy S. Gaumer Erickson Amy S. Gaumer Erickson Erickson, Amy S. Gaumer

Chunmei Zheng Chunmei Zheng Zheng, Chunmei

Social Network Analysis Social network analysis

1541

1543

Social Network Analysis

David Knoke and Song Yang defines social network as a structure depicting interconnections among a set of members or actors. Social network analysis is an approach to better understand the exchange of information or other resources within these interconnections. The goal of this relational approach is to conceptualize and quantify how actors (e.g., people, groups, or organizations) interconnect and influence other actors. Social network analysts use qualitative and quantitative methods to (a) conceptualize social network ties with visual tools, such as graphs, tables, and figures and (b) characterize the nature of those ties, such as the strength of the relationships.

The application of social network analysis in research and evaluation is fairly recent, emerging in the 1930s, appearing in limited research articles in the 1970s, and then gradually increasing to date, with a recent sharp incline. Fields such as social sciences, computer science, and organizational management have recognized the benefit of using this approach to study patterns of relationships in a structure. This set of techniques can be used to capture complex patterns of interaction among actors as well as depict the structural change of interactor relationships over time.

There are three assumptions underlying social network analysis. First, structural relations are critical for understanding and predicting behavior more than attributes such as age, gender, and background. Second, social relations or networks affect perceptions, beliefs, and actions through a variety of structural mechanisms that are socially constructed among entities. Third, structural relations should be viewed as dynamic processes. This indicates that the relations among entities are not fixed; on the contrary, relationships change all the time. Better understanding the interconnectedness among actors further informs our understanding of context, patterns, and systems of groups of actors.

Social network analysis accommodates six types or measurement levels of variables: binary, multiple-category nominal, grouped ordinal, full-rank ordinal, interval, and ratio. A binary measure of relations refers to 1 representing the presence of a relation and 0 representing an absence. Multiple-category nominal measures of relations refer to the nominal measure with multiple groupings (e.g., participant selects among a series of options: friend, business relationships, or no relationship). Grouped ordinal measures of relations refer to ordinal data such as dislike, neutral, and like options. Full-rank ordinal measures of relations refer to rank data in which participants rate the relations from the strongest to the weakest. Interval and ratio measures of relations refer to data in which the measure is continuous.

Social network analysis results are mathematically calculated and then commonly visually represented via tables, graphs, and figures that illustrate characteristics such as density, degree centrality, closeness, and betweenness. Graphs efficiently highlight key features of a social network structure and consist of nodes that represent actors and lines that represent ties.

Concepts such as the density of the network, or the degree of connectedness of groups of actors, which is calculated to be a value between 0 and 1, can be illustrated via a graph of nodes and lines. Degree centrality concerns the extent to which a person or group is connected to other actors and is used to identify prominent actors within the network. Closeness and betweenness are centrality measures that help determine an actor’s proximity to others and can illustrate the depth of a relationship.

Strength and direction of the ties are also important to quantify and illustrate. Tie strength is the intensity, frequency, or strength of interaction between pairs of actors. Direction refers to the direction of relations between dyad members. The intensity or the strength of interaction between pairs of actors can be quantified with a scale. For example, Bruce Frey’s 5-point collaboration questionnaire uses the following gradations: no interaction at all, networking, cooperation, coordination, coalition, and collaboration. In this example, point scale indicates the strength of interaction with 0 representing the *lowest strength*, while 5 represents the *highest strength* of interaction; lines are graphed of varying thicknesses to illustrate the level of connectedness. Nondirected relations occur when relations mutually occur (e.g., conversing) and directed relations occur when one actor initiates and the second actor receives (e.g., advising or e-mailing resources). Therefore, nondirected relations are symmetric and the strength of interactions between dyads is interchangeable, while directed relations are asymmetric and not interchangeable.

The sum of relations (ties) for the receiver and sender illustrates the nodal *indegree* and nodal *outdegree*, respectively. Nodal degree is the total number of relations for a certain entity, where degree refers to number of ties. Because the data are directional, nodal degree can be split to provide nodal indegree and nodal outdegree values. Nodal indegree is the number of ties received by one entity from other entities, while nodal outdegree is the number of ties sent by an entity to others. The indegree and outdegree of one entity may differ from each other. If an entity has a greater outdegree value than indegree value, the entity is expansive; if an entity has a large nodal indegree value, the entity is popular.

Social network analysis can provide summary or descriptive statistics such as mean (density) for each entity and for the entire network, standard deviation, sum of relations (ties), and minimum and maximum value in the data. Entities can be senders and receivers (e.g., if one agency rates level of relationship with nine other agencies, the one agency is the sender and the nine agencies are receivers) and the summary statistics reported from both senders and receivers, respectively, can disclose more comprehensive information. The mean (density) is the average link, describing the average strength of collaboration across all relations. The sum is the total number of ties, suggesting which entities are more influential. The standard deviation indicates the variability of the distribution, with larger values representing greater variability. Together these data characterize the social network.

Patricia M. Noonan, Amy S. Gaumer Erickson, and Chunmei Zheng

See also Dyadic Data Analysis; Program Evaluation; Social Network Analysis Using R; Sociometric Assessment

Further Readings

Carolan, B. V. (2014). Social network analysis and education: Theory, methods and applications. Sage.

Durland, M. M., & Fredericks, K. A. (2005). An introduction to social network analysis. New Directions for Evaluation, 2005(107), 5–13.

Knoke, D., & Yang, S. (2008). Social network analysis. Sage.

Paul E. Johnson Paul E. Johnson Johnson, Paul E.

Social Network Analysis Using R

Social network analysis using R

1543

1549

Social Network Analysis Using R

Social network analysis (SNA) involves analysis of the formation of relationships and the transmission of information (or possibly products, diseases, and so forth). Research in social networks has grown rapidly since 1990, a reflection of the improvement in statistical computing (faster computers accommodate more complicated models) and the growth of the Internet, which provides both data through participation in social media, including blogs, and an open environment in which researchers can exchange software. This entry looks at the growth of interest in researching social networks using the statistical software framework called R. It first discusses the development of R, then looks at how it is used to represent social networks and analyze the formation of new networks.

R and Its Community

R is one of the outstanding successes of the free software movement. The code is open for inspection, editing, and redistribution without restriction. *A New York Times* article published January 7, 2009, speculated that R was becoming a lingua franca of statistics and data science.

The R framework is modular; most of the work is done by functions contained in packages. It is easy for new R users to overlook the difference between the base of R (the software distributed by the R Core Team, which includes 30 packages) and packages provided by the R community, which now number more than 10,000 (taking together the Comprehensive R Archive Network as well as smaller repositories such as Bioconductor, R-Forge, and GitHub). The openness to addition of packages is a significant part of the explanation for R’s growth.

Another reason that R is becoming the lingua franca of statistics is that R can absorb functions written in fast, low-level programming languages such as C and Fortran. Experts may prefer to write in C++, for example. R packages that incorporate those functions often appear.

The base R distribution does not include tools for SNA. The general purpose social network frameworks considered here are found in packages igraph, statnet, and graph. Depending on the researcher’s taste and needs, any or all of these may be useful. In combination with tools in R base, each one of these is able to handle the following:

> importation of data,
> description of networks (summarize connections among individuals),
> visualization (plotting and interaction with graphic displays), and
> Simulation of artificial networks.

Users can expect differing degrees of difficulty when using these packages. One package will offer nicer plots, but at the expense of more difficult data preparation, for example. Researchers will have to pick and choose among the functions offered by different packages. There are significant stylistic differences among packages and it is not always easy to navigate among them. The packagers are aware of these concerns and make frequent revisions. As a result, many blogs and tutorials about SNA are outdated. Books published as recently as 2014 describe functions in packages such as igraph, which no longer exist or are being phased out.

There is no single R package that can handle all of the more advanced needs of social network researchers. The statnet suite is the closest to that objective. It links together 14 separate R packages; especially noteworthy are the data importers in network, network connectivity analysis in sna, and exponential random graph models (ERGM) in ergm.

The igraph package for R is a “wrapper” around a general purpose C library. The library can be accessed from programs written in R, Python, or C. igraph has areas of strength in calculations for huge network data sets, especially in community detection.

The graph package is used in conjunction with others in the Bioconductor

Repository. It offers especially good plotting routines based on GraphViz, a publicly available graph layout library that was prepared at AT&T Labs. Because R cofounder Robert Gentleman is a team leader in graph, it should not be surprising that the style of coding (function names and data storage structures) is more consistent with R itself than the other packages (graph uses S4 formal classes, rather than the less formal S3 style).

Users are likely to be confused by some terminology when they explore these packages. First, the word *attribute* has dual meanings. In SNA, *attribute* is a characteristic (e.g., a person’s age). On the other hand, R design uses *attribute* to refer to a marker that can be inserted on any R object. (See the R functions attributes and attr.) Documentation for the SNA packages is frequently confusing because of this dual usage.

Second, the style of function names is idiosyncratic and somewhat confusing. The R run-time system, and most packages affiliated with R Core, tends to use a style for function names that helps users differentiate purpose from the nature of objects on which action depends. However, the R “namespace” system segregates functions so that R packagers are allowed to name functions in any style that they choose. Moving from one package to the next, one is struck by the differences in style. In addition, to discourage the proliferation of new names for common chores, R has generic functions. Generic functions, such as plot or summary, exist as abstract labels. They don’t do work, they send work to “method functions,” known as “methods” for short.

Methods have names such as summary.igraph (in the igraph package) or plot.network (in the network package). The aim of this design is simplicity: Users need to know only the generic name (plot or summary) not the full name. However, R packagers are not required to participate in that scheme. For example, in the statnet suite, there are two virtually identical functions: plot.network is a method function in network, while gplot in sna is a standard function that appears to be almost identical.

Social Network Representations

In SNA software, individuals are referred to as vertices or nodes (sometimes also “agents” or “actors”). An edge is a connection between two nodes (also referred to as a “link,” “connection,” or “tie”). The vertices and edges, taken together, are known as a graph (note that “graph” does not mean “plot”).

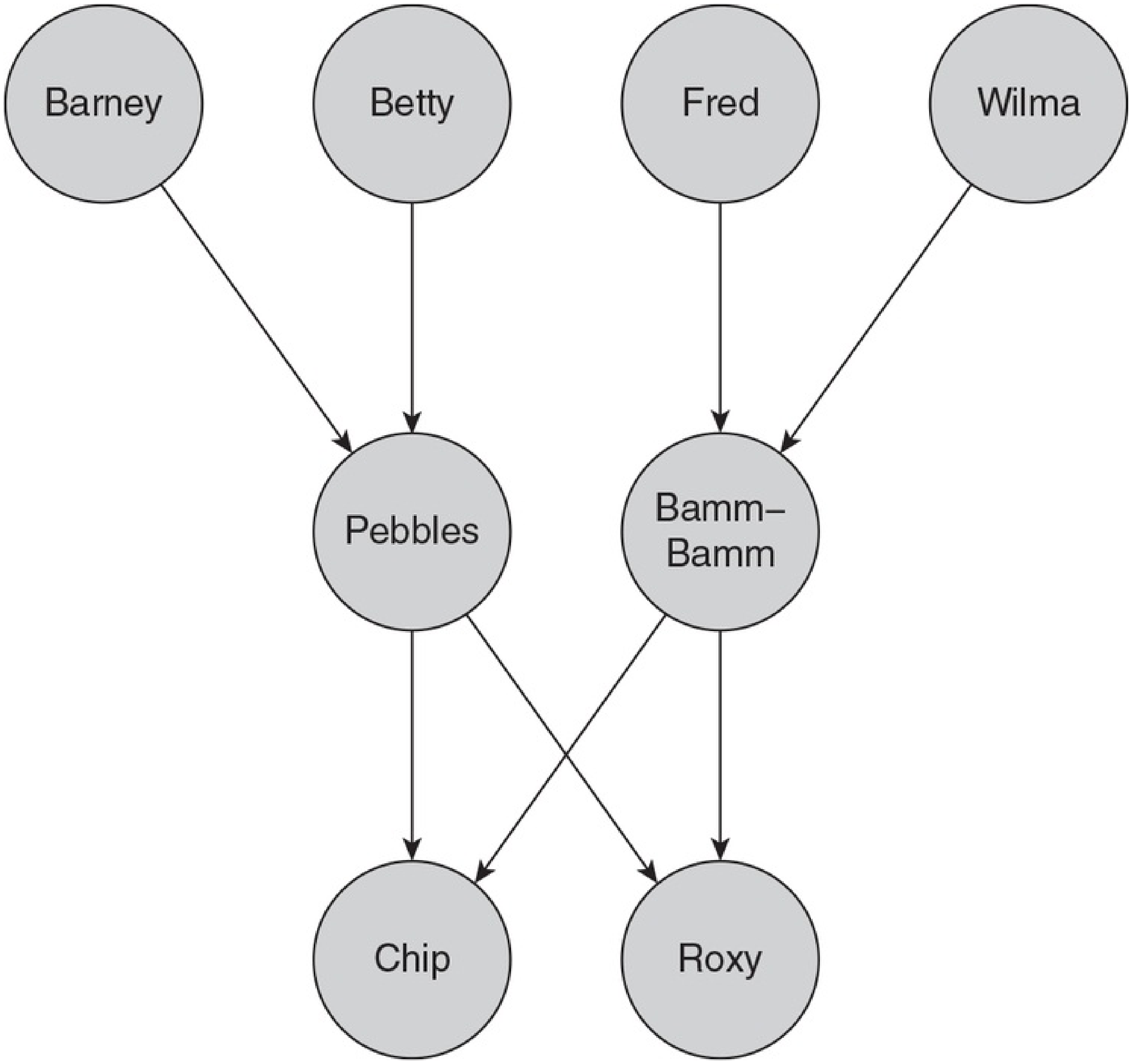
An adjacency matrix is an array of 1s and 0s indicating the presence (or absence) of a relationship. Table 1a shows marriage ties in the fictional town of Bedrock. This matrix is symmetric (identical above and below the main diagonal) because it represents an undirected network: Fred is married to Wilma implies that Wilma is also married to Fred. Some adjacency matrices have weighted edges, representing the idea that connections might vary in strength.

The adjacency matrix is often used to exchange data among packages. However, it is not generally used for storage within packages. Packages use more compact formats, such as the edge matrix (Table 1b), or edge list (Table 2b). The columns are labeled "ego" and "alter" (sometimes they are named "from" and "to"). The edge matrix requires much less storage, and yet it contains all of the same information.

Many interesting social networks are directed networks, where there is a significance in the "fromto" direction of the relationship. The edges might represent social dominance, affection, kinship, or the like. In Table 2a, we have the adjacency matrix for parent–child relationships in Bedrock. In Table 2b, we have the corresponding network edge list. The igraph package uses an edge matrix as the default storage format, whereas graph uses a list in graphNEL (NEL means "network edge list").

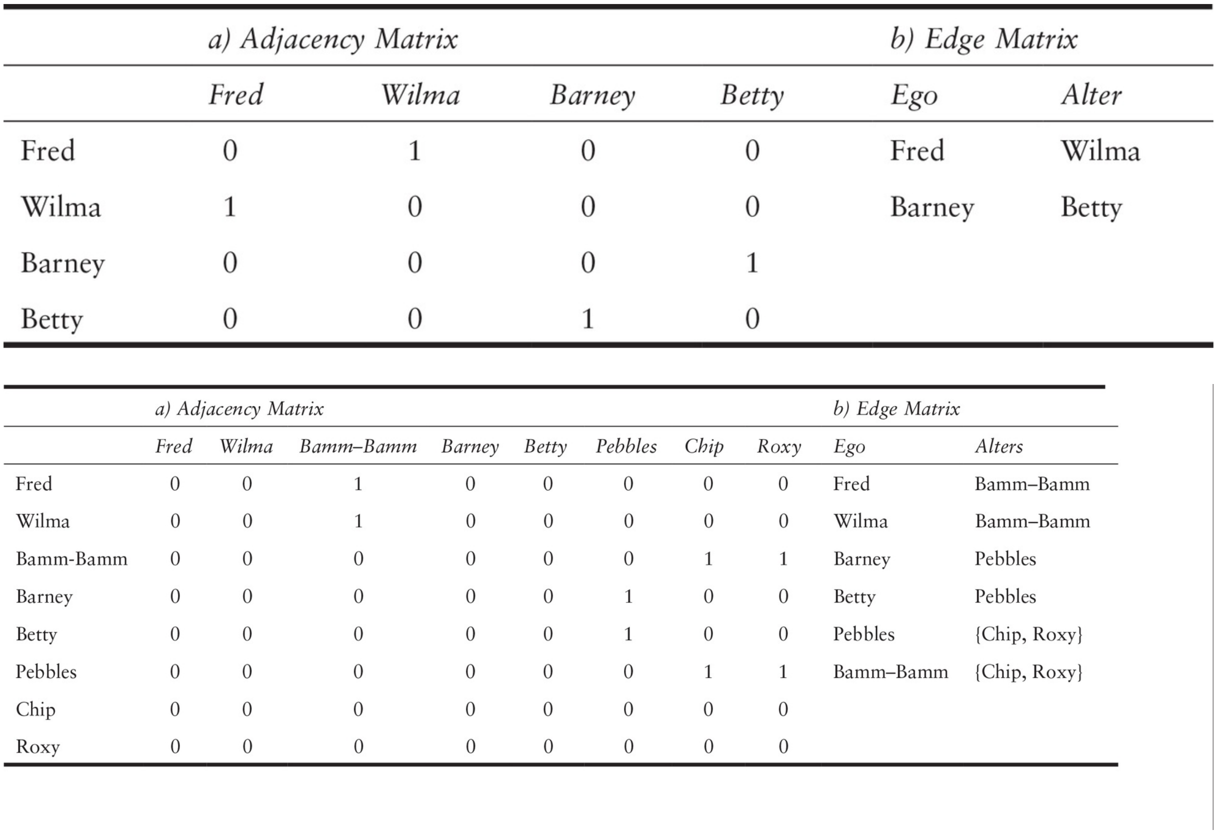
Visualization is a key part of the social network research process. Networks have no natural (*x*,*y*) coordinates. Instead, the nodes are placed algorithmically to convey information. There are many competing layout algorithms. The "dot" layout is displayed in Figure 1, an illustration of parent–child relationships prepared with graph and Rgraphviz. Circles represent nodes; arrows represent edges. For small data sets, the graphs produced by those packages are usually the most visually appealing.

Figure 1 Plot of a directed graph



Rewiring Social Networks

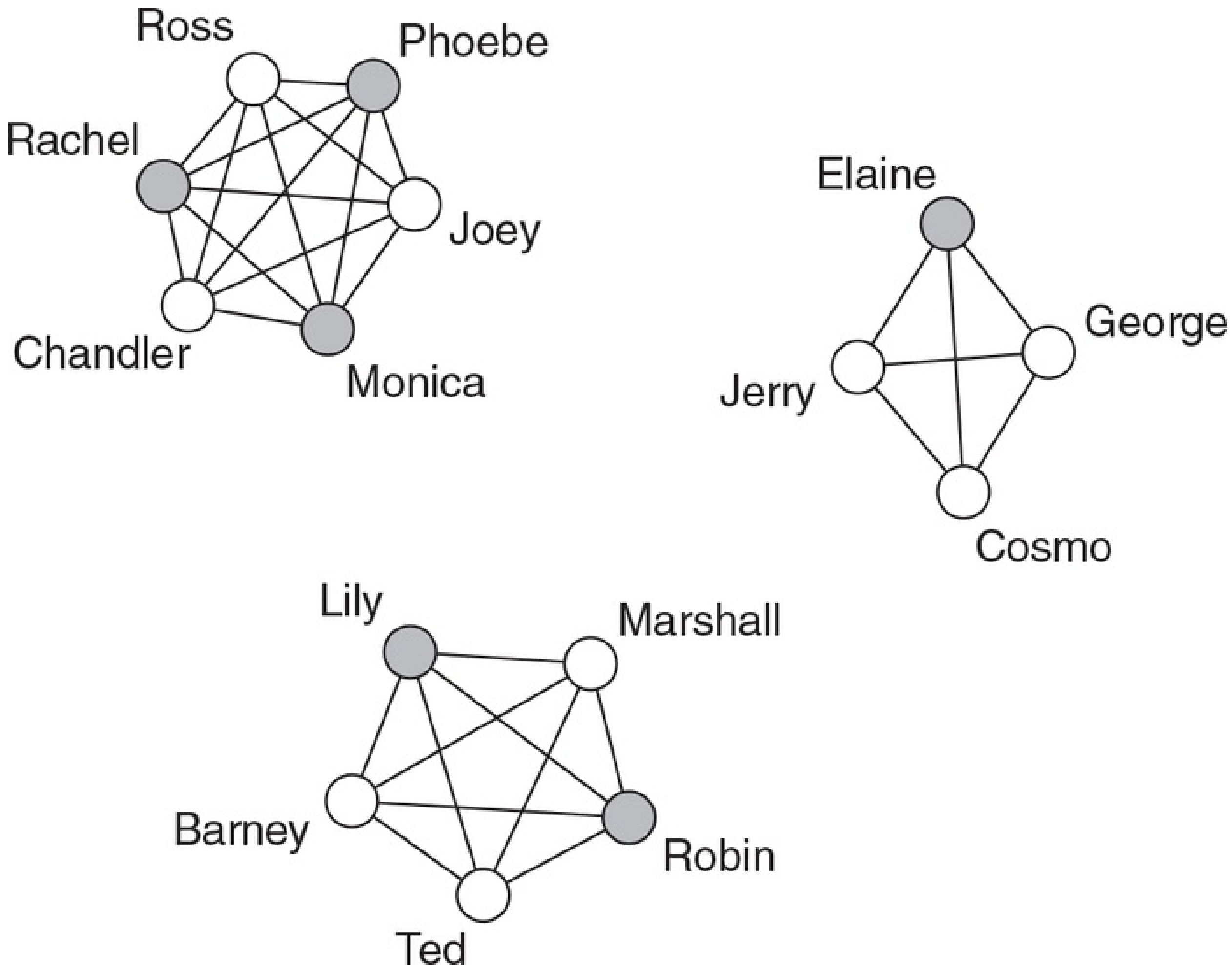
To explore the different styles of storing and editing network data, a small data set about fictional sets of friendship networks was created. The ties within three isolated subnetworks are illustrated in Figure 2. This plot was created by the sna package’s gplot function. To color code the vertices, gender and gender color attributes were assigned to the nodes. With gplot, node labels can be requested and are printed beside the nodes.



	a) Adjacency Matrix				*b) Edge Matrix*	
	Fred	*Wilma*	*Barney*	*Betty*	*Ego*	*Alter*
Fred	0	1	0	0	Fred	Wilma
Wilma	1	0	0	0	Barney	Betty
Barney	0	0	0	1		
Betty	0	0	1	0		

	a) Adjacency Matrix								*b) Edge Matrix*	
	Fred	*Wilma*	*Bamm–Bamm*	*Barney*	*Betty*	*Pebbles*	*Chip*	*Roxy*	*Ego*	*Alters*
Fred	0	0	1	0	0	0	0	0	Fred	Bamm–Bamm
Wilma	0	0	1	0	0	0	0	0	Wilma	Bamm–Bamm
Bamm-Bamm	0	0	0	0	0	0	1	1	Barney	Pebbles
Barney	0	0	0	0	0	1	0	0	Betty	Pebbles
Betty	0	0	0	0	0	1	0	0	Pebbles	{Chip, Roxy}
Pebbles	0	0	0	0	0	0	1	1	Bamm–Bamm	{Chip, Roxy}
Chip	0	0	0	0	0	0	0	0		
Roxy	0	0	0	0	0	0	0	0		

Figure 2 Friendship networks



This plot used the popular Fruchterman-Reingold layout algorithm which is widely available. The “spring tension” model pushes apart nodes that are not connected. Users may be surprised to find that running the gplot command over and over will generate a new arrangement each time (random numbers are used to anchor some nodes).

Researchers often want to test the impact of “rewiring” a graph by adding or deleting edges. It is possible to insert a new node, named Janice (connected to Chandler), and insert new ties (between Jerry and Monica as well as between Marshall and Chandler).

Figure 3 shows the drawing created by the igraph package. Before creating the plot, a community detection procedure named cluster_fast_greedy was used to generate the “blobs” that unite some nodes. This kind of analysis can be used to find out where separate communities become indistinguishable as new edges are inserted.

The interface to add nodes and edges is easier to manage in the graph package, but the chore is manageable with igraph and network. Both the latter packages offer several different avenues to manipulate vertices and edges, including some special purpose operators that create a distinctive code style.

The Fruchterman-Reingold layout of the new network has a peculiar property illustrated in Figure 3. Inserting new connections has an unexpected consequence of changing the depiction of other relationships that were not altered. The depiction of the linkage between Phoebe, Joey, and Rachel changes when Chandler gains outside links. Visualization algorithms are based on algorithms that make assumptions about the impact of additional edges that we might not expect.

All of the SNA packages include tools to summarize edge connectivity in graphs. Classical network analysis affords a wealth of summary statements (node centrality and so forth). The selection of summary statistics will be wider in igraph and sna, but it is also substantial in the package RBGL, which is associated with graph in Bioconductor.

Exploration of statistical and graphical tools benefits from the ability to simulate network data. The graph package provides a few basic simulators for elementary graphs, whereas igraph and the sna packages provide more simulators for network configurations.

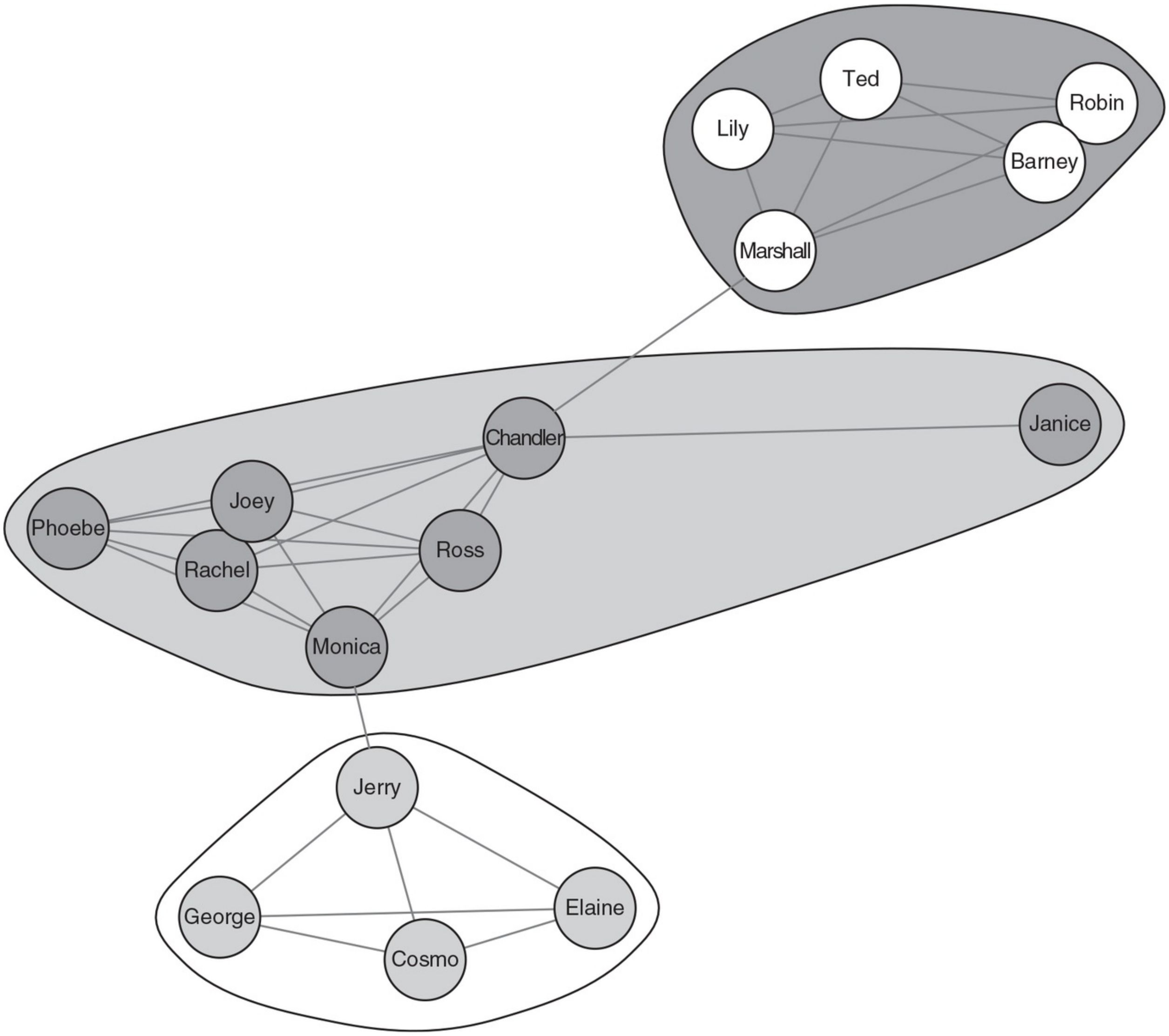
Statistical Modeling

In ERGM research, we think of network data differently. Suppose we gather data about sexual interactions among teenagers in a society in which homosexuality and interracial relationships are taboo. The information is collected into an edge list object with the tools in the network package. Next, consider two possible models. One model supposes that edges are formed randomly (equally likely among teens [boys or girls], without regard to race). The other supposes that boys are more likely to connect with girls and same-race teens are more likely to connect. In this case, it is much more likely the data came from the second model. That is ERGM in a nutshell. ERGM tools choose the most likely type of network and estimates coefficients to assess the relative impact of gender and race on edge formation.

Introduction of these models in the early 1990s revolutionized SNA. Two of the leading R packages are ergm and RSiena. The ergm is part of the statnet suite, while RSiena is a separate, but related, project.

The ERGM framework makes it possible to form very elaborate conjectures about what connections might form. Predictive terms might include network properties (general proclivity to form ties), individual characteristics (age, race, and gregariousness), dyads (match or mismatch of node attributes), triads (if one node is linked to two others, are those two others more likely to form a connection?), and so forth. It appears as though the sky is the limit, as the ergm package now includes more than 70 of these predictive terms for network relationships. Recent enhancements of ERGM to be found in statnet focus on longitudinal changes in networks to represent, for example, the spread of rumors or disease.

Figure 3 Communities isolated by igraph



The package RSiena is a more recent development. This began as a way to estimate ERGM but has now transitioned into a stochastic actor model. The terminology is similar to ERGM, but there are two differences. First, RSiena models can predict not only formation of ties but also the impact of changing ties on individual attitudes and behaviors. The model allows us to ask, for example, will a teenager become more likely to use drugs if new connections are formed with other teens who have done so in the past? Second, RSiena is intended for longitudinal network data. The logic of the model is inherently dynamic; repeated observations are required. Details, such as missing data due to withdrawal of study participants, have been taken into account and work is proceeding on alternative implementations of the parameter estimator.

Paul E. Johnson

See also R; Social Network Analysis; Sociometric Assessment

Further Readings

Handcock, M., Hunter, D., Butts, C., Goodreau, S., & Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. Journal of Statistical Software, 24(1),1–11.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008b). ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software, 24(3), 1–29.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3), 299–314.

Kolaczyk, E. D., & Csárdi, G. (2014). Statistical analysis of network data with R. New York, NY: Springer.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org.

Snijders, T. A., van de Bunt, G. G., & Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. Social Networks, 32(1), 44–60.

Vance, A. (2009, January 7). Data analysts are mesmerized by the power of program R. New York Times, p. B6.

Samantha B. Goldstein Samantha B. Goldstein Goldstein, Samantha B.

Marc H. Bornstein Marc H. Bornstein Marc H. Bornstein

Socio-Emotional Learning Socio-Emotional learning

1549

1550

Socio-Emotional Learning

Socio-emotional learning (*SEL*) aims to teach children to make responsible decisions, manage their emotions, and maintain a positive attitude toward school and social relationships. SEL uses classroom instruction, community service, extracurricular activities, and supportive home environments to help children develop into responsible and constructive members of society. SEL begins at home with caregivers and continues with school programs spanning preschool through high school, teaching students social-and self-awareness, impulse control, empathy, cooperation, and problem resolution. This entry describes the history, approaches, and effectiveness of SEL programs.

History of SEL

Social and emotional learning became popular in the early 1990s when Peter Salovey and John D. Mayer published research about emotional intelligence, defined as the ability to recognize and monitor one’s emotions and use them to direct cognition and behavior. In his State of the Union address in 1997, President Bill Clinton raised the issue of character education. Defined as a nation-wide focus on teaching children to develop a sense of core ethical values, character education brought national attention to the importance of SEL. SEL stems from character education, but it focuses more broadly on an active learning process and the ability to generalize socio-emotional skills across multiple settings.

Approaches

SEL programs utilize many outlets (classroom, community, and extracurriculars) to implement social and emotional learning. The most common forum is classroom instruction, involving a structured curriculum that addresses five key competencies: self-awareness, self-management, social awareness, relationship skills, and responsible decision making. Experts stress the importance of applying curriculum across contexts so students can practice these skills in different settings (e.g., school and home). Therefore, a fundamental aspect of SEL programs is to ensure that teachers and families work closely together to support students' socio-emotional growth. SEL programs teach educators to use positive discipline techniques and to be emotionally supportive of their students and families.

Why SEL Is Important

Emotions are important in cognitive learning; much of what we learn is linked to specific events or social and emotional situations. Therefore, SEL correlates with academic achievement. The quality of student–teacher interactions and classroom instructional practices predicts higher academic performance and social adjustment. Additionally, caregiver support is crucial in child SEL. Effective SEL programs can increase academic achievement, decrease problem behaviors, and improve the quality of students' relationships. By promoting communication and setting positive goals, these programs help students learn to be active and constructive members of a community.

Samantha B. Goldstein and Marc H. Bornstein

See also Curriculum; Emotional Intelligence; Learning Styles; Out-of-School Activities; Problem Solving; Self-Regulation; Social Learning

Further Readings

Durlak, J., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. Child Development, 82(1), 405–432.

Elias, M. J., Zins, J. E., Weissberg, R. P., Frey, K. S., Greenberg, M. T., Hayness, N. M.,… , Shriver, T. P. (1997). Promoting social and emotional learning. Association for Supervision and Curriculum Development (ASCD).

Yvonne H. M. van den Berg Yvonne H. M. van den Berg van den Berg, Yvonne H. M.

Sociometric Assessment Sociometric assessment

1550

1552

Sociometric Assessment

Sociometric assessments have a long history that dates back to the 1934 work *Who shall survive?* by Jacob Moreno. In his book, Moreno introduced sociometric assessment as a method to measure attraction and repulsion between individuals and to study individuals' functioning within groups. Although sociometric assessments were initially used among adult prison inmates and in psychiatric hospitals, the method was quickly used in other social settings as well. Currently, sociometric methods are most commonly used to measure children's social position at school but are also used in other group structures such as sports teams or professional work environments.

The underlying assumption of the method is that each group member is an expert observer of daily interactions in the group and can therefore evaluate the group and its members on a variety of social characteristics. The respondents are those who are actually part of the group and insiders in the group culture. Importantly, scores are based on judgments by multiple respondents rather than a single individual. As a result, research has repeatedly and consistently shown that sociometric assessments provide highly reliable and valid information about the structure of social groups.

Traditionally, the term s*ociometric assessment* or *sociometry* referred to a wide variety of methods assessing relationships in groups. Yet, a distinction should be made between sociometric assessments and peer assessments. Sociometric methods are used to measure (mutual) liking and disliking and individuals' social position in the group (e.g., whether a person is liked), whereas peer assessments are used to measure behavioral characteristics (e.g., what a person is like). Although both methods are often used together and share the same basic procedure, the purpose and type of information gathered are different.

This entry describes the basic procedure and several important considerations when using sociometric and peer assessment methods.

Basic Procedure

The most commonly used method to collect sociometric and peer assessment data is peer nominations. With peer nominations, a distinction is made between the persons who answer the questions (i.e., the voter population) and the persons who are evaluated (i.e., the reference group). Ideally, all members of the group are part of both the voter population and the reference group. However, it is possible that a member of the reference group is not part of the voter population or vice versa. For instance, when individuals are willing to participate but absent on the day of testing, they can be evaluated by the other group members while not filling in the questionnaire themselves.

Typically, respondents fill in a questionnaire consisting of several questions on social relationships (sociometric assessment) as well as social behaviors (peer assessment). For example, students at school are asked “In your classroom, who do you like most?” and “In your classroom, who is most helpful?” Then, each student is asked to evaluate the other classmates on these traits by nominating those who best fit each description. Students can usually nominate as many or as few members of the group for each question as they like but are generally discouraged from nominating themselves. In paper-and-pencil versions of sociometry, respondents can nominate others by writing down names or code numbers. Sometimes, a roster with the names of the members of the reference group is provided so that respondents can simply check off the ones they want to nominate. Increasingly, paper-and-pencil versions are being replaced by computerized testing, but the underlying procedures are the same.

By using this basic procedure, information is collected not only about each individual in a group but also about dyadic relationships and group structure. At the individual level, scores can be computed for individual social status or behaviors. This is done by counting the number of nominations received for each question. To compare results between groups (e.g., classrooms of different sizes), it is important to correct for group size. That is, being named 10 times in a classroom of 11 students has a different meaning than being named 10 times in a classroom of 30. Controlling for group size is done by different methods of sociometric standardization, such as standardizing nominations received to *Z* scores, computing proportion scores, or by using a regression-based technique. These scores indicate how liked or popular a person is in the group relative to the other group members or whether a person is seen as more or less aggressive or prosocial than other group members.

It is also possible to derive information about dyadic relationships, such as reciprocal friendships or mutual antipathies. To do this, one codes whether a nomination given is reciprocated. For example, if person A nominated person B as “best friend,” did B also nominate A?

At the group level, a sociometric and peer assessment can yield information about the structure of the social network and individuals’ position in the network. For instance, one can visualize friendship networks that show clusters of friendships but also who are more central versus isolated in the larger group.

Considerations

Although the basic procedure is often the same across different assessments, there are several considerations and choices to make when conducting sociometric and peer assessments.

One of the first decisions is whether to ask respondents about positive behaviors and relationships only (e.g., “Who do you like most,” “Who is helpful”) or also about less desirable behaviors and relationships (e.g., “Who do you like least,” “Who is aggressive”). Teachers and parents sometimes are concerned about the negative consequences of sociometric and peer assessments when children are asked to evaluate classmates on potentially negative characteristics such as aggression, bullying, or dominance. Despite the fact that research has found minimal to no evidence for malicious effects, some ethical review boards share these concerns and do not allow researchers to ask about negative behaviors and relationships. However, studies have found that one acquires the most reliable and valid information when using both positive and negative nomination questions. It is therefore recommended to ask respondents about positive as well as less desirable relationships or behaviors in the group. Yet, always try to start and end with a positive question. Moreover, many concerns can be eliminated by carefully following the ethical principles and code of conduct for psychological data collection. Careful verbal instructions must be given to the respondents to explain the purpose of the assessment and the meaning of anonymity and confidentiality. Respondents should also be told that they are free to participate (or not) and could stop at any given moment.

In addition to the choice of questions, one needs to choose between limited and unlimited nominations. With limited nominations, respondents are restricted in the number of group members they can choose for a question. Traditionally, respondents were often asked to nominate three group members they liked most. With unlimited nominations, respondents are free to nominate as many or as few group members as they like. There are advantages and disadvantages to both limited and unlimited nominations. For instance, limiting the number of nominations saves time to complete the questionnaire but yield less ecologically valid data than unlimited nominations. In general, limited and unlimited nominations lead to very comparable results. Yet, the use of unlimited nominations is recommended for sociometric questions regarding social status.

Third, when using a roster or list with names there is the chance of order effects; group members at the top of a list are nominated more often than those lower down the list. There are various ways to deal with this issue. Statistically, one can control for order effects. One can also try to avoid order effects by randomizing the order of the names per respondent. The downside is that it takes respondents more time to find the name of the group member they want to nominate, especially when the group is large and the list of names is long (e.g., when the school grade is the reference group). It is therefore advised to randomize the order of the names across respondents, yet to keep it constant across questions. In modern computerized applications of sociometric and peer assessments, this is relatively easy to implement.

Next, writing down multiple names or circling names from a roster for each question can be labor-intensive, time-consuming, and even frustrating for respondents. This is especially the case when there are many questions or the reference group is large. At the same time, data processing can also be time-consuming and further complicated when answers are illegible due to poor handwriting. Computerized assessments are a good solution and alternative to the traditional paper-and-pencil questionnaires, as respondents can simply click on the names of the group members and data are processed automatically. Still, one should try to keep the number of questions as small as possible.

Finally, it is important that a substantial proportion of the group participates in order to obtain reliable results. Ideally, all group members evaluate one another order to obtain reliable results. Ideally, all group members evaluate one another on all questions. However, one often ends up with a subsample of the larger group as some group members are not willing or able to be part of the assessment. Yet, low response rates may result in unreliable results. A recent study showed that for some criteria (i.e., overt aggression) participation rates of 40% could still provide reliable information, but for other traits (i.e., friendship), participation rates of at least 85% are needed. Regardless of the trait of interest, higher response rates lead to more reliable results. As a rule of thumb, a minimum participation rate of 60% to 70% is recommended.

Given these concerns and the time-consuming nature of sociometric and peer assessments, researchers have thought about alternatives. For example, some have assessed random subgroups of the larger group or only interviewed group members who were seen as the most knowledgeable and aware of the social processes in the group. Researchers also have wondered whether teachers could provide the same information. Although these alternatives are valuable, results are not as reliable and valid as when they are provided by the peers themselves. Therefore, sociometric and peer assessments remain a popular and useful method to obtain information about individuals, dyads, and networks in social groups.

Yvonne H. M. van den Berg

See also Classroom Assessment; Dyadic Data Analysis; Social Network Analysis; Socio-Emotional Learning; Survey Methods; Z Scores

Further Readings

Bukowski, W. M., Cillessen, A. H. N., & Velásquez, A. M. (2012). Peer ratings. In B. Laursen, T. D. Little, & N. A. Card (Eds.), Handbook of developmental research methods (pp. 211–228). New York, NY: Guilford Press.

Gommans, R., & Cillessen, A. H. N. (2015). Nominating under constraints: A systematic comparison of unlimited and limited peer nomination methodologies in elementary school. International Journal of Behavioral Development, 39, 77–86. doi:10.1177/0165025414551761

Marks, P. E. L., Babcock, B., Cillessen, A. H. N., & Crick, N. R. (2013). The effects of participation rate on internal reliability of peer nomination measures. Social Development, 22, 609–622. doi:10.1111/j.1467–9507.2012.00661.x

Mayeux, L., Underwood, M. K., & Risser, S. D. (2007). Perspectives on the ethics of sociometric research with children: How children, peers, and teachers help to inform the debate. Merrill-Palmer Quarterly, 53, 53–78. doi:10.1353/mpq.2007.0002

Poulin, F., & Dishion, T. J. (2008). Methodological issues in the use of peer sociometric nominations with middle school youth. Social Development, 17, 908–921. doi:10.1111/j.1467-9507.2008.00473.x

van den Berg, Y. H. M., & Cillessen, A. H. N. (2014). Computerized sociometric and peer assessment: An empirical and practical evaluation. International Journal of Behavioral Development, 37, 68–76. doi:10.1177/0165025412463508

Mario A. Navarro Mario A. Navarro Navarro, Mario A.

Jason T. Siegel Jason T. Siegel Siegel, Jason T.

Solomon Four-Group Design Solomon four-group design

1552

1554

Solomon Four-Group Design

The Solomon four-group design, developed by Richard Solomon in 1949, was devised to overcome the problem of pretest sensitization. Pretest sensitization occurs when participants' scores on a posttest are influenced as a result of a pretest being administered. The central feature of the Solomon four-group design is that participants are randomly assigned to either receive or not to receive a pretest and then randomly assigned to either a treatment or a comparison group. All participants then receive a posttest. This approach enables researchers to acquire the benefits of using a pretest, while also allowing an assessment of pretest sensitization.

Benefits of Pretest Inclusion

Researchers implement pretests with the goal of obtaining information regarding baseline levels of specific variables of interest (e.g., self-esteem, knowledge) prior to the implementation of an experimental stimuli or intervention. The central benefit of collecting pretest data is that it provides a comparison point for posttest data. Illustrative of the benefits of a pretest, consider a study where two classes received educational interventions. If the average SAT score of Class A was 1,000 and Class B was 800, it is possible that the intervention provided to Class A was more useful than the one provided to Class B. However, if SAT scores of Class A were 1,200 at pretest and Class B's were 200 at pretest, a very different picture of the intervention impact is presented. In this instance, without knowledge of the pretest scores, researchers could make inaccurate conclusions about the effectiveness of the treatment. On the other hand, if pretest scores indicated that the two classes did not significantly differ from each other at pretest, the investigator can have increased confidence that posttest differences were caused by the treatment.

Costs of Pretest Inclusion

Even though the benefits of a pretest include greater control over an experimental or quasi-experimental design, several drawbacks accompany their use. These shortcomings include increased monetary cost, increased time consumption, and pretest sensitization. As noted, pretest sensitization occurs when the implementation of the pretest leads participants to respond to the stimuli or the posttest assessment differently than they would have otherwise.

The problems associated with pretest sensitization are manifold. Firstly, a pretest can alert participants to the questions that are likely to appear on the posttest. For instance, if students receive a difficult pretest examination prior to a math-based intervention, some of them might take it upon themselves to learn the answers to the pretest’s challenging questions—irrespective of the quality of the intervention that was received. If students in both conditions learn the answers to the posttest as result of the pretest, any possible differences between groups caused by the educational intervention will be obscured as, due to the pretest, both the treatment condition and the control condition will have high scores on the posttest. Another possibility is that the students will focus only on the aspects of the educational intervention that were covered on the pretest, thus resulting in scores that they would not have received if they did not have prior knowledge of questions on the posttest.

In other instances, the pretest could make the participants aware of outcomes that researchers are hoping to influence, regardless of whether they intended to do so. For instance, a health intervention may seek to reduce drug use through an intervention that intentionally never mentions drugs, thereby reducing rebellion to the message. The inclusion of a pretest measuring drug use could make the goal of the intervention obvious, thus increasing the likelihood of participant rebellion. A pretest can also impair data integrity when participants, in either the treatment or the comparison condition, attempt to answer questions in an identical manner as the pretest. If this occurs, the impact of the treatment on the outcome of interest could be obfuscated.

The Solomon Four-Group Solution

As noted, the Solomon four-group design was devised to overcome the problem of pretest sensitization while maintaining the benefits associated with conducting a pretest. The design achieves this aim by randomly assigning participants to either receive or not to receive the pretest and then to receive or not to receive the treatment. By randomly assigning these two factors of treatment and pretest, four conditions are created. Researchers can readily discern the influence of the pretest by contrasting differences in posttest scores between both groups that received the treatment (one of which received a pretest) and the two groups that did not receive the treatment (one of which received a pretest). For example, if participants who did not receive a pretest before an educational intervention scored an 80 on a math posttest, but those who received a pretest and an educational intervention scored a 100 on the math posttest, there may be cause for concern. These results would be particularly revealing if this pretest effect was not present for the control group. Another possibility could be that there was a main effect for the treatment, a main effect for the pretest, and an interaction between the two.

The central drawbacks to the Solomon four-group design are its cost and feasibility—this design requires twice the number of participants, materials, and resources to implement. For example, in a traditional pretest–posttest study with random assignment to just two conditions, roughly 100 participants would be needed; however, because implementing a Solomon four-group design doubles the number of conditions to four, 200 would be needed. Larry Howard, Thomas Tang, and M. Jill Austin have suggested that researchers can curtail this issue by randomly assigning a smaller percentage of participants to the control conditions than to the treatment conditions. By doing this, investigators can reduce the cost of their studies but still reap the benefits of a Solomon four-group design.

Recent Addition

Although the Solomon four-group has been used sporadically throughout its existence, a recent modification of the design, the *Solomon postgroup design*, could be used to obtain the benefits of including manipulation checks (i.e., assessments to determine whether the experimental stimuli impacted on the participants as expected) while avoiding the potential harms. Similar to the inclusion of a pretest, using a manipulation check also comes with costs and benefits. A key benefit of using a manipulation check is that researchers can be confident that the experimental treatment worked as expected; a drawback is that placing measures between the manipulation or treatment and the outcome measure could lead to a reduction of the treatment's influence on the key outcome measure. The Solomon postgroup design is similar to the Solomon four-group design in that there are four conditions—two conditions receive a treatment, while the other two conditions do not receive the treatment. However, rather than randomly assigning participants to receive or not receive a pretest, participants are randomly assigned to receive or not receive a manipulation check measure between the treatment and the outcome measures. This, in turn, tests the effects of the manipulation check on the outcome measure.

Mario A. Navarro and Jason T. Siegel

See also Experimental Designs; Survey Methods; Surveys

Further Readings

Crano, W. D., Brewer, M. B., & Lac, A. (2014). Principles and methods of social research. New York, NY: Routledge.

Howard, L. W., Tang, T. L. P., & Austin, M. J. (2015). Teaching critical thinking skills: Ability, motivation, intervention, and the Pygmalion effect. Journal of Business Ethics, 128, 133–147. doi:10.1007/s10551–014-2084-0

Sawilowsky, S., Kelley, D. L., Blair, R. C., & Markman, B. S. (1994). Meta-analysis and the Solomon four-group design. The Journal of Experimental Education, 62, 361–376. Retrieved from http://dx.doi.org/10.1080/00220973.1994.9944140

Solomon, R. L. (1949). An extension of control group design. Psychological Bulletin, 46, 137–150.

Mary L. McHugh Mary L. McHugh McHugh, Mary L.

Spearman Correlation Coefficient Spearman correlation coefficient

1554

1558

Spearman Correlation Coefficient

The Spearman correlation coefficient is a nonparametric, correlation statistic that measures the strength of association between two rank-ordered variables. The Spearman rho is symbolized by the Greek letter, rho (ρ). The ρ was developed to measure the strength of association between two ordinal variables, although it can also be used with interval and ratio variables. The ρ is a robust statistic and works well with ordinal variables that have either a small number or a large number of levels and is often used with interval/ratio variables that do not meet the normal distribution assumption of parametric statistics.

Significance statistics for which ρ is often used include the Mann-Whitney *U* test, and the Kruskal-Wallis *H* statistic. It is not used with variables measured at the nominal level or when both variables are dichotomous, even if ordinal dichotomous. The Spearman is also not the preferred statistic when there are many ties in the data.

Background

ρ was developed by Charles Edward Spearman, a professor of psychology who was known for his application of statistical concepts to the study of psychology. He is most famous for his developmental work in factor analysis and for his development of the Spearman ρ.

Because the Spearman ρ is a correlation statistic, it measures the *strength* of an association between two variables. Correlation statistics provide 4 items of information: First, they answer the question, “Do these two variables covary?” That is, does one variable change when the other changes? Second, when two variables do covary, these statistics describe the direction of the association, which can be positive or negative. A positive correlation means as one variable increases the other also increases. A negative correlation means that as one variable increases the other decreases. Third, correlations describe the strength of the association. Strength in this context means how closely do the two variables change together? In a perfect correlation, for every one level of rise in one variable, the other variable would change exactly one level; it would either rise (positive correlation) or fall (negative correlation) that one level. The ρ value can range from −1.0 to + 1.0. Fourth, the significance of the obtained value can be determined using a significance table (if the statistic is hand calculated), and the statistical programs that produce the ρ provide a significance level as part of the output.

Assumptions

The ρ, like virtually all inferential statistics not specifically designed to test matched pairs or other related measures, assumes that the sample was randomly selected from a defined population. It assumes subjects were independently sampled from the population. That is, selection of one subject is unrelated to selection of any other subject. It is not appropriate for use with paired or otherwise related samples.

The relationship between the two variables must be generally linear. That is, for the ρ to be useful, there must be a single direction of the correlation (Figure 1). Specifically, as one variable increases, the other variable either increases (positive correlation) or decreases (negative correlation). If the relationship has one or several distinct curves (Figure 2), the ρ is not an appropriate statistic and may find little or no association because it cannot test curvilinear associations.

Figure 1 A linear relationship: Relationship appropriate for ρ

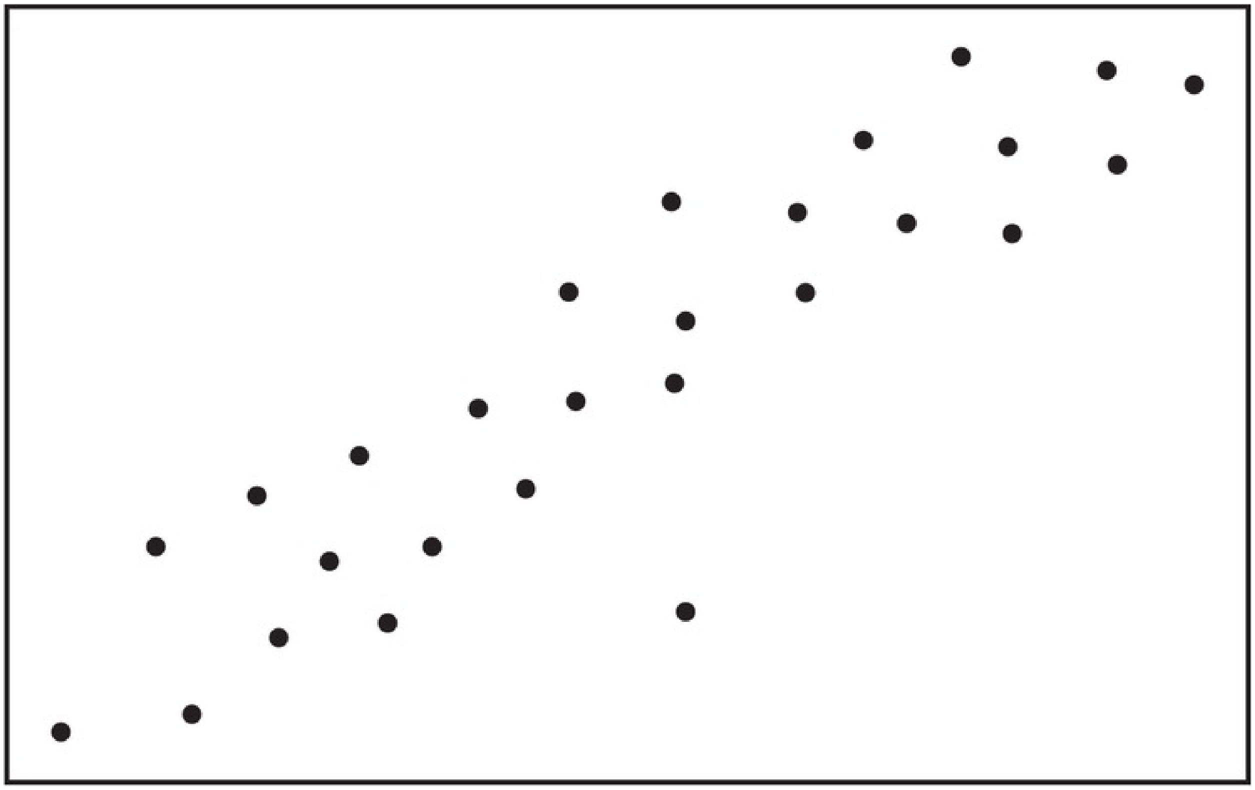
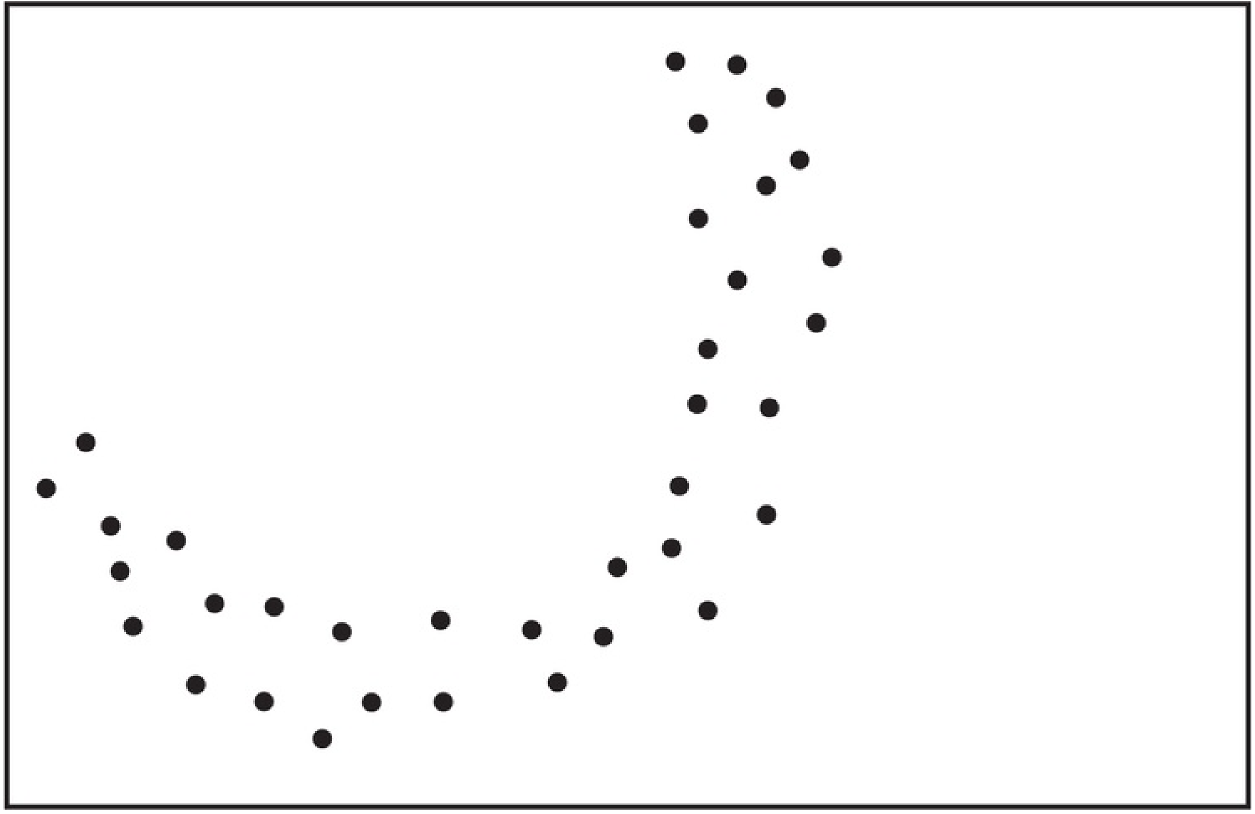


Figure 2 A curvilinear relationship: Relationship not appropriate for ρ



Calculation

The calculation is not a simple task unless the data set is small and there are few or no ties in the data. The general formula for the Spearman ρ is as follows:

$$\rho = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}.$$

In this formula, the values for each variable are ordered from low to high, ranked, and for each case. Then, the rank on Variable 2 is subtracted from the rank of Variable 1. Then the obtained difference is squared, all the squares are summed and the result is multiplied by the constant, 6. Those processes are represented in the following part of the formula: . The “*n*” in the formula is the sample size.

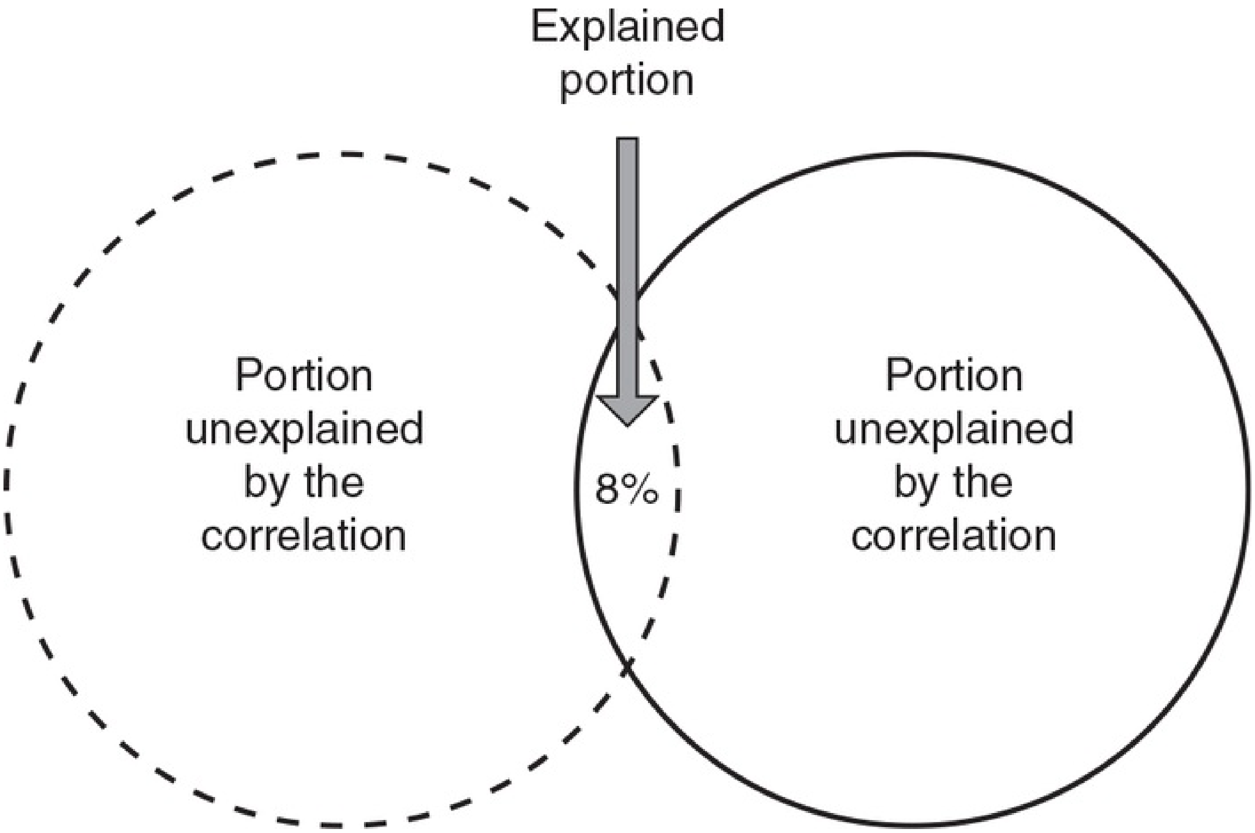
This formula is used only when there are few or no ties in the data, which happens if both variables for a case have the same rank. Then the subtraction results in a value of zero, and that case drops out of the formula. Many ties in the data will seriously underestimate the strength of the association, so a different formula must be used.

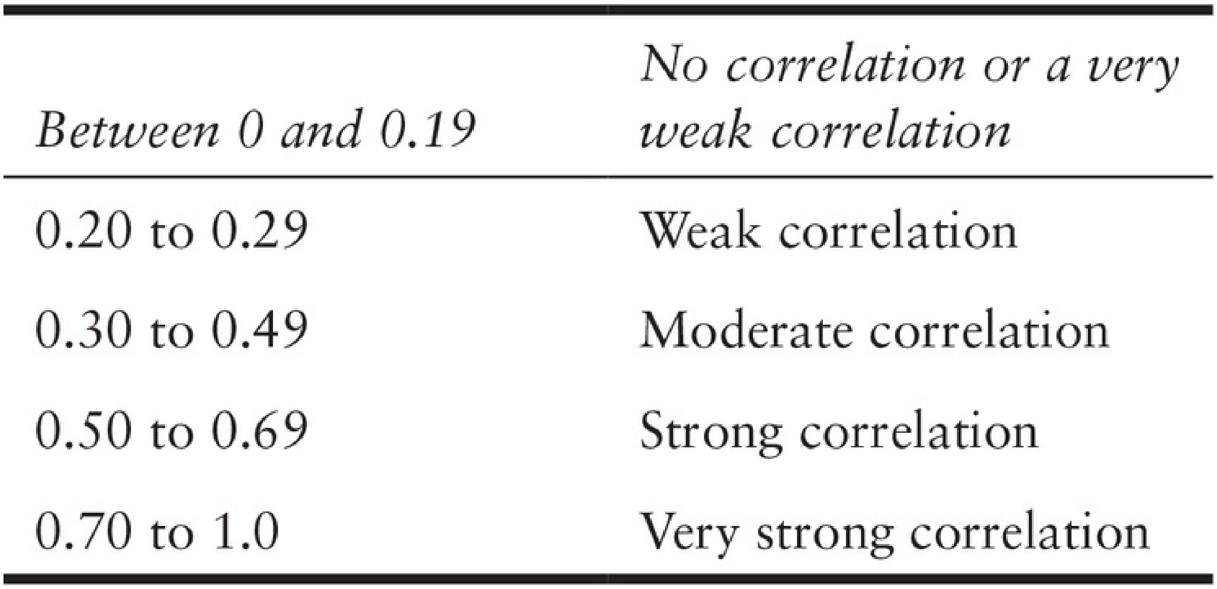
Interpretation

Values for the Spearman ρ can range from −1.0 to +1.0. A value of 1.0 means there is a perfect one-to-one correlation between the two variables.

Although different authors may use different values for weak, moderate, and strong correlation measures, the following table can be used as a general guide to interpretation of the strength of effect size represented by various values of the φ correlation coefficient.

Figure 3 Amount of variance explained by the independent variable


Explained portion
Portion unexplained by the correlation
8%
Portion unexplained by the correlation



Between 0 and 0.19	*No correlation or a very weak correlation*
0.20 to 0.29	Weak correlation
0.30 to 0.49	Moderate correlation
0.50 to 0.69	Strong correlation
0.70 to 1.0	Very strong correlation

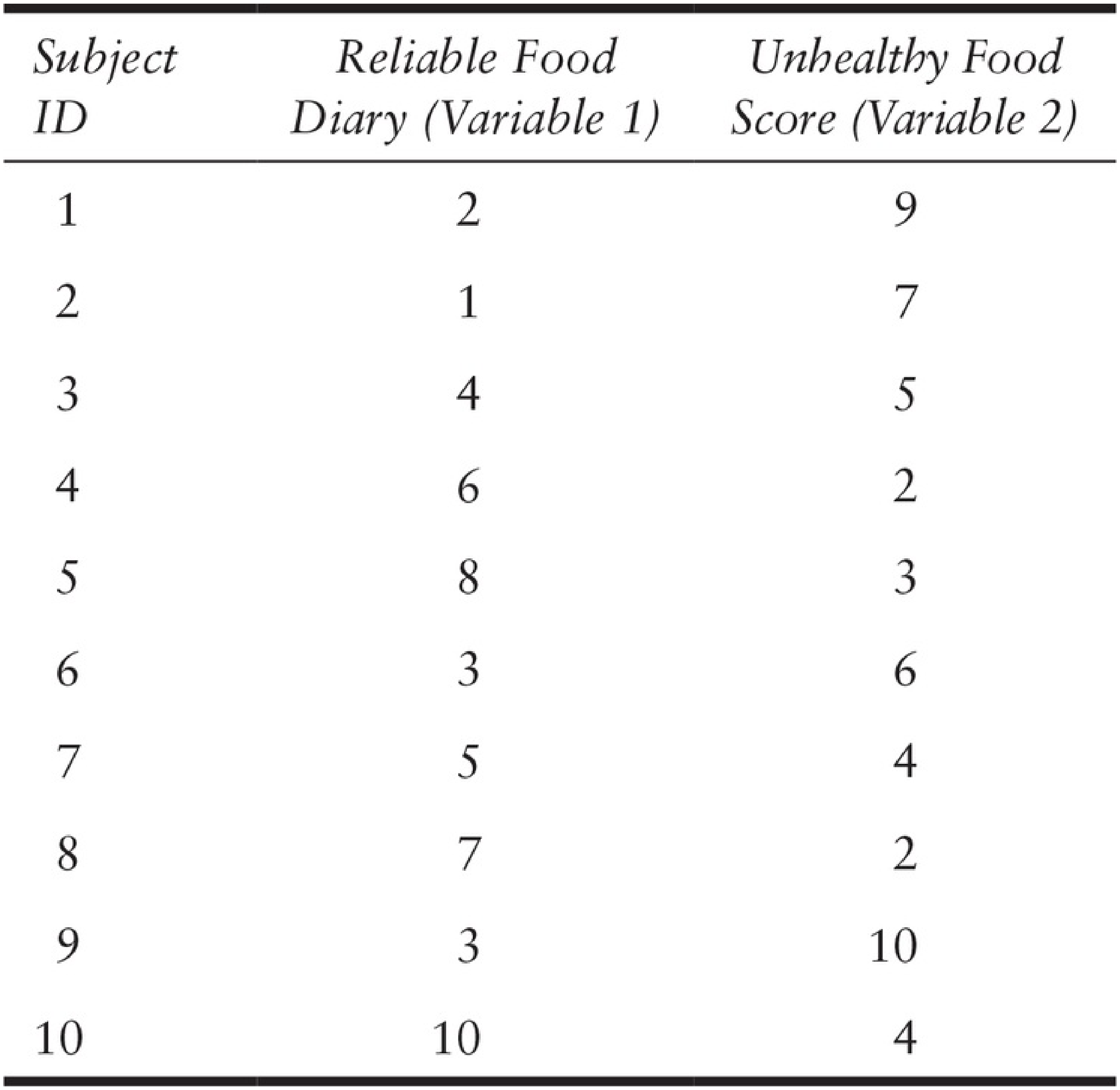
These interpretations are based on the amount of variance in the dependent variable explained by the independent variable. A correlation of +0.29 means that even if statistically significant, only about 8% of the variance in the dependent variable is explained by the independent variable. This is a small amount of overlap, and Figure 3 above demonstrates the concept.

Example of Spearman Rho Use

Assume that a teacher in secondary school responsible for teaching nutrition in a health class was interested in the teens changing from an unhealthy diet of snack foods and sugary soft drinks to a healthy diet. The teacher rated each student's performance on keeping a food diary. At the end of the term, the teacher had each student complete a food diary for 2 days. The variables were coded as follows:

1. Students received 1 point for each day they filled in the food diary, and an additional 3 points were awarded for completeness of the diary. As a result, students could achieve a score of from 1 (*almost nothing submitted*) to 10 (*very complete diary*).
2. Students received points for each sugary drink or unhealthy snack they had eaten in the end-of-term diary. The scores ranged from 1 (*no unhealthy food/drinks*) to 10 (*most of the diet consisted of unhealthy food/drinks*).

The teacher wanted to know whether the discipline needed to change dietary habits is related to the discipline needed to keep a food diary for a whole week. Because higher scores on the diary variable (V-1) represent a good diary, and low scores on dietary change (V-2) represent few unhealthy foods, she expected to obtain a negative correlation. Her hypothesis was: The students who kept a full food diary would consume the least amount of unhealthy food/drink. The Spearman ρ is an excellent statistic to choose to find the answer to the question.

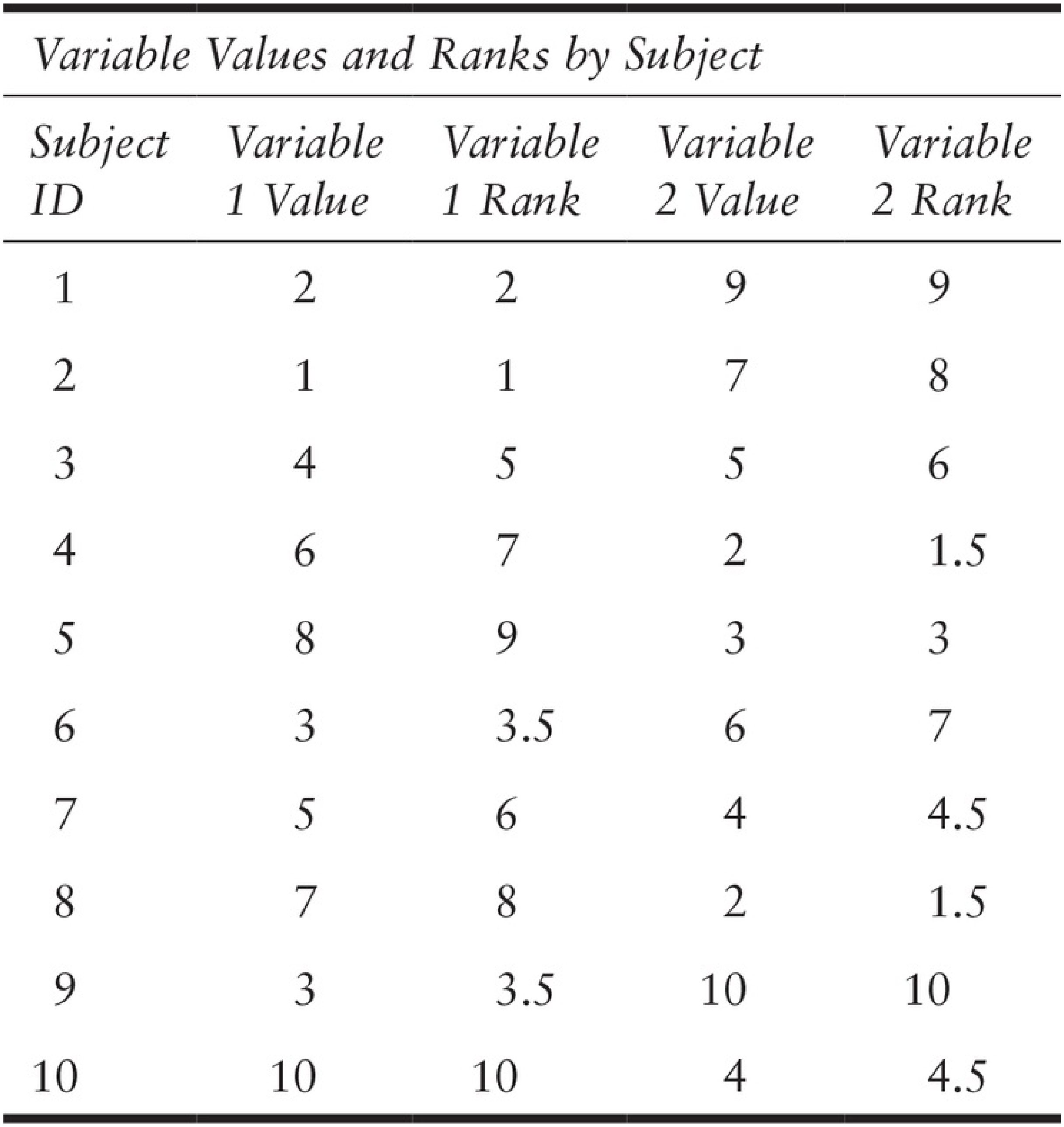


Subject ID	*Reliable Food Diary (Variable 1)*	*Unhealthy Food Score (Variable 2)*
1	2	9
2	1	7
3	4	5
4	6	2
5	8	3
6	3	6
7	5	4
8	7	2
9	3	10
10	10	4

The first step in calculating this statistic is to rank order the student's answers on each of the variables. In this hypothetical data set, the following table presents the students' scores on each of the questions.

The first step is to order the values of each variable from highest to lowest and then rank order the values for each of the variables:

Variable Values and Associated Ranks			
Variable 1 Values Ordered	*Variable 1 Ranks*	*Variable 2 Values Ordered*	*Variable 2 Ranks*
1	1	2	1.5
2	2	2	1.5
3	3.5	3	3
3	3.5	4	4.5
4	5	4	4.5
5	6	5	6
6	7	6	7
7	8	7	8
8	9	9	9
10	10	10	10



Variable Values and Ranks by Subject

Subject ID	*Variable 1 Value*	*Variable 1 Rank*	*Variable 2 Value*	*Variable 2 Rank*
1	2	2	9	9
2	1	1	7	8
3	4	5	5	6
4	6	7	2	1.5
5	8	9	3	3
6	3	3.5	6	7
7	5	6	4	4.5
8	7	8	2	1.5
9	3	3.5	10	10
10	10	10	4	4.5

Using the formula, the difference between ranks must be calculated, and each of those differences will be squared. The sample size is 10, and the formula can be completed.

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}.$$

Difference Scores and Difference Scores Squared

Subject	*V-1*	*V-2*		
ID	*Rank*	*Rank*	*Difference*	d^2
1	2	9	–7	49
2	1	8	–7	49
3	5	6	–1	1
4	7	1.5	5.5	30.25
5	9	3	6	36
6	3.5	7	–3.5	12.25
7	6	4.5	1.5	2.5
8	8	1.5	6.5	42.25
9	3.5	10	–6.5	42.25
10	10	4.5	5.5	30.25
Sum d^2 values				294.5

$$\rho = 1 - (6 \times 294.5) / (10 \times 99).$$

$$\rho = 1 - \left(1767 / 10\left(99\right)\right).$$

$$\rho = 1 - \left(1767 / 990\right) = 1 - 1.785$$

$$= -.785. \text{Rounding yields a } \rho \; of - .79.$$

The ρ value is −.79. This is a very strong negative correlation. Looking up a correlation of −.79 for a sample size of 10, the correlation is significant at the $p <$ .01 level. The result would be presented as follows: There was a strong, negative correlation between reliably keeping a food diary for a whole week and the amount of unhealthy food and drinks the student consumed at the end of the term ($\rho = -.79$, $p < .01$, $n = 10$).

Mary L. McHugh

See also Correlation; Phi Correlation Coefficient

Further Readings

Heavey, E. (2014). Statistics for nursing: A practical approach (2nd ed.). Jones & Bartlett Learning.

McClave, J., & Sincich, T. (2016). Statistics (13th ed.). Pearson.

Polit, D. (2013). Statistics for nursing research (2nd ed.). Prentice Hall.

Ravid, R. (2014). Practical statistics for educators (5th ed.). Rowman & Littlefield.

James Dean Brown James Dean Brown Brown, James Dean

Spearman-Brown Prophecy Formula Spearman-Brown prophecy formula

1558

1560

Spearman-Brown Prophecy Formula

The Spearman-Brown prophecy formula provides a *rough* estimate of how much the reliability of test scores would increase or decrease if the number of observations or items in a measurement instrument were increased or decreased. This formula is called the Spearman-Brown (S-B) formula because the idea was introduced by both C. Spearman and W. Brown in articles they wrote in 1910. This entry demonstrates two ways to calculate the S-B formula and show how the predictions in score reliability typically vary with increases or decreases in the numbers of items on a test.

The S-B formula is commonly used to estimate the full-test reliability from the half-test correlation when calculating split-half reliability. *Split-half reliability* is an internal-consistency strategy for estimating reliability that is similar to the parallel-forms strategy, except that the parallel forms in this case are created by scoring two equal halves of a test separately—usually by scoring the odd-numbered and even-numbered items separately. The tester then calculates a correlation coefficient for the odd-and even-numbered scores, and the result is an estimate of the reliability of the odd-numbered scores or of the even-numbered scores but not of both halves together. Because testers are typically concerned with the full-test reliability including all of the items and that a longer test can reasonably be expected to be more reliable than the short halves, an adjustment is made to the half-test correlation using the S-B formula for a test that is twice as long to estimate the full-test reliability. One S-B formula often applied in such cases is:

$$r_{xx'} = \frac{2 \times r}{1 + r},$$

where is full-test reliability and *r* is half-test reliability.

For example, consider a 40-item test that has an odd-even half-test (with each half having 20-items) correlation of .70. The full-test (40-item) reliability would be:

$$r_{xx'} = \frac{2 \times r}{1 + r} = \frac{2 \times .70}{1 + .70} = \frac{1.40}{1.70} = .8235 \approx .82.$$

Thus, adjusting the half-test correlation using this formula predicts that the full-test reliability is likely to be about .82.

A more general version of the S-B formula can be used for estimating the reliability of a test that is increased in length by any number of times (e.g., 2 times, 3 times, 4 times, 2.5 times):

$$r_{xx'} = \frac{n \times r}{(n-1)r + 1}.$$

Here, the symbols are the same except that *n* is the number of times the test length is increased. Applying this formula to the same example as earlier where the test length is doubled to estimate the full-test reliability from the half-test correlation, the result is the same:

$$r_{xx'} = \frac{n \times r}{(n-1)r + 1} = \frac{2 \times .70}{(2-1).70 + 1}$$

$$= \frac{1.40}{(1).70 + 1} = \frac{1.40}{1.70} = .8235 \approx .82.$$

Applying the same formula to estimate the reliability for a 60-item version of that same test would involve adjusting the reliability for a test that has 3 times (*n* = 3) as many items as the half-test correlation (for 20 items) as follows:

$$r_{xx'} = \frac{n \times r}{(n-1)r+1} = \frac{3 \times .70}{(3-1).70+1} = \frac{2.10}{(2).70+1}$$

$$= \frac{2.10}{1.40+1} = \frac{2.10}{2.40} = .8750 \approx .88.$$

It is even possible to adjust for a test that is shorter. For example, a tester who has a 100-item test with an estimated reliability of .9211 might want to know for practical reasons how reliable the scores would be at a more manageable 50-item length. To adjust in this direction involves estimating the reliability for a test that is half as long (*n* = .5):

$$r_{xx'} = \frac{.5 \times r}{(.5-1)r+1} = \frac{.5 \times .9211}{(.5-1).9211+1} = \frac{.46055}{(-.5).9211+1}$$

$$= \frac{.46055}{-.46055+1} = \frac{.46055}{.53945} = .8537 \approx .86.$$

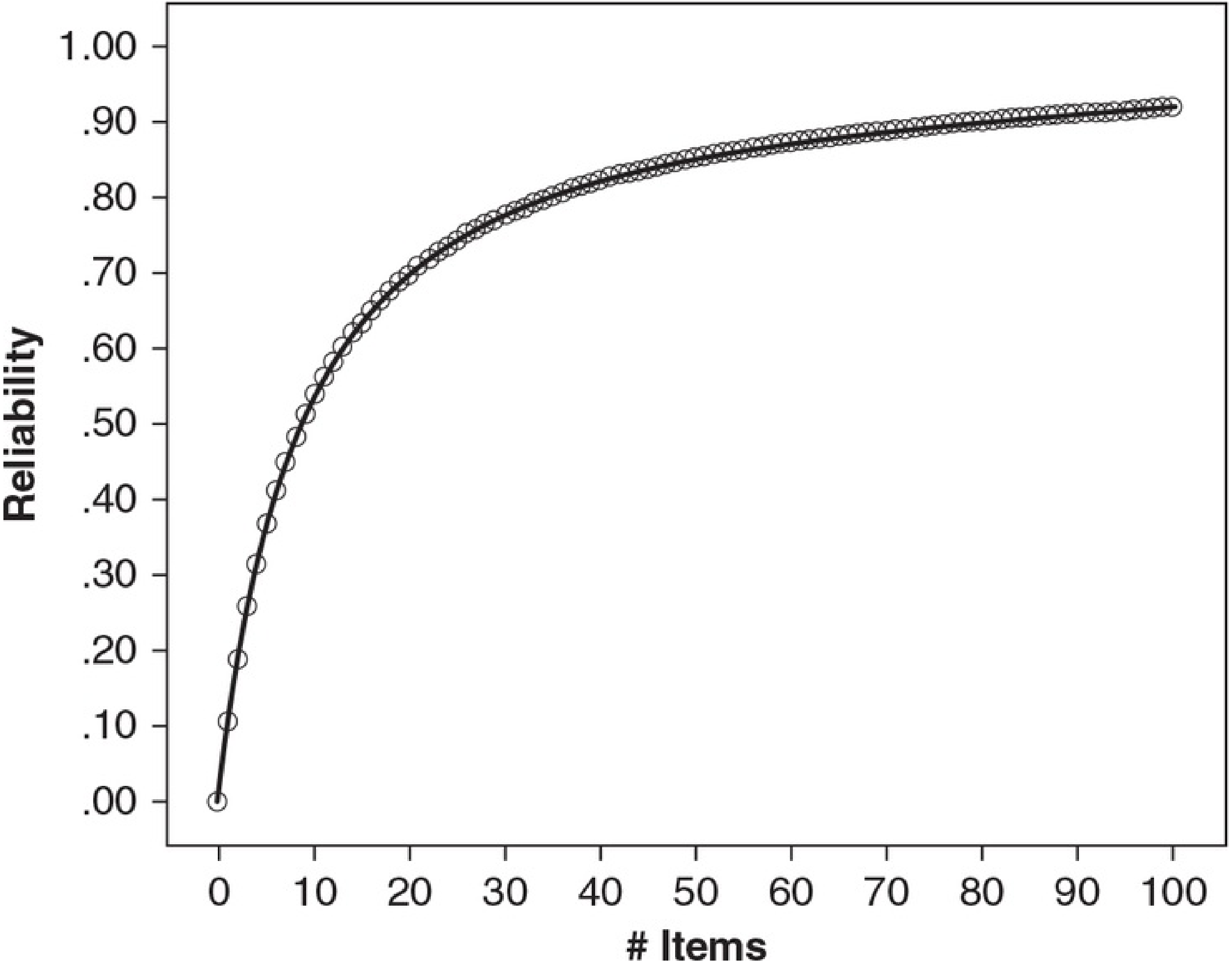
If the tester finds that the .86 level of reliability is adequate, a 50-item test would certainly involve less time and effort for examinees, proctors, test scorers, and so forth.

Just to bring things full circle, let’s check to see if this 50-item test with a reliability estimate of .8537 would turn out to be reliable at .9211 if adjusted back to the 100-item length. Using the simpler S-B formula from this entry, it turns out that:

$$r_{xx'} = \frac{2 \times r}{1+r} = \frac{2 \times .8537}{1+.8537} = \frac{1.7074}{1.8537} = .9211 \approx .92.$$

Thus, the S-B formula can be used to estimate the reliability of a test that is 2 times, 3 times, 4 times, 2.5 times, or even half (.5 times) or 20% (.2 times) as long, and so forth. Indeed, Figure 1 shows the resulting S-B formula reliability estimates calculated for 0–100 item test lengths. These calculations, like all of the example calculations in this entry, are based on the 20-item split-half correlation of .70 in the first two examples. Notice that as the number of items increases up to about 20 or 30, there is considerable gain in reliability, but that the increases taper off considerably after that, meaning that there is less bang-for-the-buck in terms of reliability gained by adding more items after that point. Although this curve will be different for every test, a similarly shaped curve will always occur. Indeed, that curve is described mathematically as follows: ; both the curve and the formula describe the estimated relationship between reliability increases or decreases as the number of items added or subtracted form a test.

Figure 1 Spearman-Brown prophecy formula reliability estimates for 0–100 items (based on 20-item split-half correlation of .70).



The S-B formula is not limited to applications involving half-test correlations. Indeed, it is often applied to what-if adjustments of Kuder-Richardson formulas 20 and 21, Cronbach’s α coefficient, and interrater and intrarater reliability estimates, among others. However, anyone applying S-B formula must keep in mind (as pointed out at the beginning of this entry) that the S-B formula only “provides a *rough* estimate of how much the reliability of scores would increase or decrease if the number of observations or items in a measurement instrument were increased or decreased.”

James Dean Brown

See also Coefficient Alpha; Internal Consistency; KR-20; Reliability; Split-Half Reliability

Further Readings

Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), Routledge handbook of language testing (pp. 303–315). New York, NY: Routledge.

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnan (Ed.), The companion to language assessment (pp. 1165–1181). Oxford, UK: Wiley-Blackwell.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

Jenny C. Wells Jenny C. Wells Wells, Jenny C.

Bryan G. Cook Bryan G. Cook Cook, Bryan G.

Special Education Identification Special education identification

1560

1562

Special Education Identification

Special education identification refers to the process for determining whether children and youth are eligible to receive services under the Individuals with Disabilities Education Improvement Act, a federal law commonly referred to as IDEA. The primary intent of the law is to ensure that all children with disabilities receive a free and appropriate public education. The IDEA includes a Child Find mandate that requires states to develop and implement a plan to seek, screen, and identify all children and youth between the ages of 3–21 years who may have a disability.

Children suspected of having a disability are referred for special education evaluation to determine whether they meet the federal criteria for eligibility for specialized instruction and related services. Although state regulations may vary, they must match or exceed the scope and intent of the federal regulations for the state to be eligible for federal funding to support the delivery of special education services to identified children. This entry discusses procedures used to identify students who are eligible to receive special education and what happens when parents disagree with the findings of the team that evaluates a student for special education eligibility.

Evaluation Process

The special education evaluation process begins with a referral for a special education evaluation. The referral may be made by the child’s parents or legal guardians, school personnel, or other community members. However, the evaluation may only be conducted once parents or legal guardians give verbal or written consent. Once consent for evaluation is received, there is a 60-day period in which the school must complete a comprehensive, individualized, initial evaluation of the child in the areas of suspected disability.

A variety of assessment tools and procedures must be used to gather relevant information on academic, developmental, and functional areas and must be comprehensive enough to provide an in-depth assessment of all areas of suspected disability and educational concern. The evaluation usually includes reports based on observation of the child and a parent interview. Although assessment tools used in the evaluation will vary depending on the age of the child and the nature of the suspected disability and educational concerns, standardized tests, curriculum-based assessments, and checklists are commonly completed as part of the evaluation process. All instruments used in the evaluation must be (a) technically sound, (b) validated for the specific purpose for which intended, (c) administered by trained personnel, (d) nondiscriminatory, and (e) administered in the child’s native language or communication mode.

The results of the evaluation are then considered by a multidisciplinary team that includes the parent or legal guardian of the child, a teacher familiar with the general education curriculum, individuals with expertise in the suspected area of disability, and individuals who are knowledgeable in the assessment instruments used in the evaluation. Information provided by the parents/guardians must be included and considered in the determination, and eligibility may not be determined based on a single assessment procedure. Team members collaboratively make decisions related to the determination of eligibility for special education services and classification based on all evaluation information available as well as professional judgment.

For a child to be deemed eligible for special education, the multidisciplinary team must conclude that, based on all of the evaluation information, the child has a disability as defined by IDEA and that as a result of the disability the child requires specialized instruction to receive an appropriate public education. The information gained through the initial evaluation is also used in the development of the child’s individualized education program.

The IDEA includes 13 categories of disability under which a child may be found eligible for special education services. Federal definitions for each disability category include descriptions of the characteristics of the disabilities and may include exclusion criteria for some categories. States may craft their own definitions, but these definitions are reviewed by the U.S. Department of definitions, but these definitions are reviewed by the U.S. Department of Education to determine whether they are equivalent to the IDEA definitions. The 13 disability categories, in alphabetical order, are (1) autism, (2) deaf–blindness, (3) deafness, (4) emotional disturbance, (5) hearing impairment, (6) intellectual disability, (7) multiple disabilities, (8) orthopedic impairment, (9) other health impairment, (10) specific learning disability, (11) speech or language impairment, (12) traumatic brain injury, and (13) visual impairment (including blindness).

Response to Intervention (RTI)

In the reauthorization of IDEA in 2004, an additional process for determining eligibility for special education for students suspected of having a learning disability was included. This process has been termed RTI. The RTI process examines a student’s response (in terms of academic performance or behavior) to increasingly intensive, research-based interventions. The RTI process was not specifically prescribed in IDEA, resulting in states and localities developing a variety of approaches.

In the RTI model, only after a student has been nonresponsive to quality, evidence-based instruction in a general education classroom would the student be formally evaluated for special education. The data generated during the process of providing progressively more intensive levels of instruction in the general education setting is included in the consideration for special education eligibility. This process was established to prevent inaccurate placement in special education where the student’s learning problem was due to inadequate instruction. RTI is not intended to delay a student’s access to an evaluation for special education, although this issue has become controversial in practice.

Parental Disagreement With Determination of Eligibility

If a child’s parents/guardians disagree with the team’s determination of eligibility, due process procedures are available for the finding to be reviewed by a neutral hearing officer. In addition, parents may also request that the school district provides information on where they may obtain an independent educational evaluation if they disagree with the findings of the evaluation conducted by the school district. The school district must then consider the findings of any independent evaluations that are provided to it by the parent or guardian when determining eligibility.

Jenny C. Wells and Bryan G. Cook

See also Curriculum-Based Assessment; Evaluation; Individuals with Disabilities Education Act; Response to Intervention; Standardized Tests

Further Readings

Pierangelo, R., & Giuliani, G. A. (2017). Assessment in special education: A practical approach (5th ed.). Upper Saddle River, NJ: Pearson.

U.S. Department of Education. (n.d.). Building the legacy: IDEA 2004. Retrieved from http://idea.ed.gov/

Yell, M. L. (2016). The law and special education (4th ed.). Upper Saddle River, NJ: Pearson.

Rachel M. Stein Rachel M. Stein Stein, Rachel M.

Special Education Law

Special education law

1562

1565

Special Education Law

Special education law consists of a series of legal dictates to monitor and protect students with specialized educational needs due to their disability status. Although these laws were designed to safeguard students’ educational rights in general, stipulations regarding appropriate measurement, evaluation, and research are embedded within this body of legislation. Understanding and following the laws and regulations pertaining to special education ensures both best practice and legal compliance when working with students with special education needs. This entry provides an overview of legislation relevant to research, measurement, and evaluation involving students in special education. Specifically addressed are the Individuals with Disabilities Education Improvement Act of 2004, Section 504 of the Rehabilitation Act of 1973, and the Every Student Succeeds Act (ESSA) of 2015.

Evolution of Special Education Law

Between 1852 and 1918, as compulsory education was enacted state by state, a number of laws were passed that excluded students with disabilities from general education. Alongside these laws restricting access to education was the creation of numerous special education classrooms to accommodate these newly displaced students. However, these special classrooms were often just holding spaces for students, motivating advocacy groups and subsequent law to argue that students with disabilities have the right to specialized protection of their right to education.

In 1975, Congress approved the Education for All Handicapped Children Act, requiring that states and their school districts provide individuals with disabilities a free and appropriate education. The law was revised in 1990 and renamed the Individuals with Disabilities Education Act (IDEA) and subsequently reauthorized in 2004 as the Individuals with Disabilities Education Improvement Act, although the acronym IDEA is still used to refer to the law. In recent years, legal opinions have consistently defended special education populations as vulnerable to educational and civil liberty violations. Actions by Congress and the executive branch, including the executive order that created the 2002 President's Commission on Excellence in Special Education, have moved beyond stating that students have a right to access to education to ensuring the quality of educational practices for students with disabilities.

The IDEA dictates the rules that schools must adhere to when providing education to students with disabilities. Included within the law are eligibility and identification requirements for special education, an explanation of a free appropriate public education, a definition of a least restrictive environment, and procedural safeguards to protect the educational rights of students and their caregivers. Under these principles, students with disabilities have a legally protected right to an appropriate education in the environment with the fewest restrictions possible that would limit their participation alongside general education students. For example, if a student is able to learn math in a general education classroom with an aide or with modified assignments, that is considered more appropriate than having the student learn math in a special education classroom.

Although the law requires that schools make an effort to provide opportunities for students with disabilities, it also acknowledges that some of the law's provisions are open to different interpretations. Therefore, the law includes procedures (e.g., mediation) that schools, parents, and students can use to resolve disagreements that arise about the implementation of the law's principles.

Measurement, Research, and Evaluation and IDEA

Evaluation and measurement are inherently part of special education law, both at the structural and individual levels. In other words, systemic programming and individual student education plans need to use evidence-based practice to support their implementation over time. Current law mandates that programs have a demonstrated record of effectiveness and that data are collected to evaluate eligibility for special education and to track and monitor students' progress toward goals. Therefore, measurement and evaluation are structurally mandated components of special education designed to ensure educational progress.

In addition to the evaluation and measurement inherent in special education, there is a need to adhere to legal mandates. For example, educators need to ensure that students in special education receive all of the supportive services that are included in their individualized education programs (IEPs). When engaging in measurement, research, and evaluation with students in special education, it is essential to have conversations with educational staff to make sure that the integrity of the students' IEPs, such as the number of educational minutes they receive, are not violated under IDEA.

Assessment and Evaluation to Determine Special Education Eligibility

Special education law requires a series of steps to ensure that data-based decision making is used to determine whether a student meets special education eligibility requirements. A team of qualified individuals, often as part of a prereferral problem-solving team, examines student data (e.g., academic performance) to determine whether a student should be evaluated for special education. A similar team of individuals then completes an assessment to determine special education eligibility.

Although the assessment process may vary depending on the reasons for the referral, federal law stipulates that the team must consist of individuals with expertise in the area of referral and related assessment tools. Further, the assessment must include structured norm-referenced tools appropriate to the question at hand (e.g., behavior rating scales) as well as an observation or other relevant qualitative measures. This might include general education teachers, special education teachers, administrators, school psychologists, social workers, and school counselors as well as any other relevant individuals. Parents or guardians must be part of the assessment process.

Further, the tools used in the assessment must be considered technically sound and administered according to standardized procedures. This includes using up-to-date assessments that are appropriate for use with the demographic groups represented within the student body. Evaluators should be mindful of the risks of using outdated test materials and old norming samples and also be aware of laws and court rulings specific to certain states. Use of improper assessment tools may result in students being misidentified or improperly disqualified for special education services.

Measurable Goals and IEPs

When students are found to meet the eligibility requirements for special education, an IEP is created. The initial IEP document relies on existing student information to serve as baseline data and establish goals. The goals must be measureable and student progress must be monitored. Throughout the process, progress monitoring data must be collected to facilitate the evaluation of student progress and ensure that any decisions regarding modifications or maintenance of student services are data based.

Every year, or more frequently if needed, the team must reconvene to update student goals and ensure that the current plan provides the support the student needs to try to meet the articulated goals. Parents must be notified of their children’s progress toward their goals at least as often as report cards are sent home. Further, every 3 years, the IEP team must reassess whether a student still requires special education to receive a free and appropriate education and make academic progress. For students who continue to meet the eligibility requirements for special education, the team needs to determine whether a student’s needs are being met with the current educational plan or if there are necessary modifications based on progress monitoring data.

Section 504 of the Rehabilitation Act of 1973

Section 504 of the Rehabilitation Act of 1973 is civil rights legislation that protects individuals with disabilities. Under Section 504, individuals with disabilities are defined as those with an impairment (physical or mental) that limits one or more major life activities. Impairments that fall under Section 504 are wide ranging and can include both physical (e.g., neurological, sensory, reproductive, endocrine) or mental (e.g., intellectual disability, learning disabilities, mental illness) disabilities. Individuals may also qualify for protection under Section 504 if they have a history of a limiting impairment or are regarded as impaired and are, therefore, limited by how others treat them.

To be protected under Section 504, an individual’s impairment must impact a major life activity including self-care, movement, seeing, hearing, speaking, breathing, learning, or working. Students who do not meet the requirements for special education may qualify for protections under Section 504. Section 504 is focused on preserving the rights of all individuals with disabilities, including protections for students in publicly funded educational environments (e.g., schools, after-school programs, recreational programs). All programs that fall under this umbrella must make reasonable accommodations that allow individuals to fully participate in activities. Thus, evaluation, measurement, and research activities need to provide reasonable accommodations (e.g., wheelchair accessible classrooms) to allow students with disabilities to participate without violating their protected civil rights under Section 504.

Section 504 of the Rehabilitation Act of 1973 dictates that qualified students with disabilities must receive a free appropriate public education that includes (a) an educational program designed to adequately meet a student’s needs, (b) within a setting that to the maximum extent possible is with nondisabled peers, (c) that uses an evaluation process to ensure appropriate identification, and (d) that ensures notification of parental rights and procedural safeguards are made available to students’ guardians.

Evaluation for Section 504 eligibility must include appropriate evaluation materials that are administered by trained professionals in the intended manner. Additionally, the assessment process must be multidimensional, rather than rely on a single test or score, and can include both standardized (e.g., achievement tests) and unstandardized (e.g., interviews, classroom performance) data sources.

The assessment should consider a variety of factors that may impact a student’s learning including cognitive functioning, achievement, teacher input, physical and health conditions, cultural background, and adaptive functioning.

ESSA

In 2015, ESSA, which updated and replaced the No Child Left Behind Act of 2001, was signed into law. ESSA legislation was written to hold schools accountable for student educational achievement. Although ESSA encompasses education generally, it also provides some relevant guidance for students with disabilities.

ESSA requires that students with disabilities be assessed and monitored using data. Further, schools are held accountable for student performance, including the performance of students with exceptional needs. Additionally, ESSA articulates that individuals must hold appropriate certification and licensure to work as a special education teacher. Schools must also provide the federal government with information about student participation in special education by reporting on the number of students with IEPs and their academic proficiency. ESSA provides additional guidance and mandates to ensure that schools are using educational approaches, including in special education programs, that are supported by data, evidence, and research.

Participation in special education does not exempt students from taking district and state assessments. Schools and districts must provide appropriate accommodations for students to participate in these evaluation activities. Further, ESSA mandates that schools and districts provide information to the federal government regarding student assessment performance, including the performance of students who receive special education services.

Research Considerations

The 2004 law reauthorizing IDEA created the National Center for Special Education Research as part of the Institute of Education Sciences of the U.S. Department of Education. This center is tasked with evaluating programs for individuals with disabilities and ensuring their effectiveness. Despite describing a range of research areas that warrant additional investigation (e.g., literacy skills, personnel preparation, early intervention), IDEA does not provide information about particular approaches to research.

Students in special education, in addition to being minors, may also be considered a vulnerable population because their disability status may put them at increased risk for violation of their rights. Therefore, researchers must proceed with the utmost caution to secure research consent and assent and protect students' rights throughout the research process. Furthermore, researchers working with special education populations need to ensure that special education laws are upheld and that students still have access to the educational services outlined in their IEPs.

Rachel M. Stein

See also Americans with Disabilities Act; Every Student Succeeds Act; Individualized Education Program; Individuals with Disabilities Education Act; Institute of Education Sciences; Special Education Identification; U.S. Department of Education

Further Readings

Jacob, S., Decker, D. M., & Hartshorne, T. S. (2011). Ethics and law for school psychologists (6th ed.). Hoboken, NJ: Wiley.

U.S. Department of Education, Office of Special Education and Rehabilitative Services. (2017, February 16). Individuals with Disabilities Education Act. Retrieved from https://www2.ed.gov/about/offices/list/osers/osep/osep-idea.html

Yell, M. L. (2012). The law and special education (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Kathleen Lynne Lane Kathleen Lynne Lane Lane, Kathleen Lynne

Eric Alan Common Eric Alan Common Common, Eric Alan

Specificity

Specificity

1565

1567

Specificity

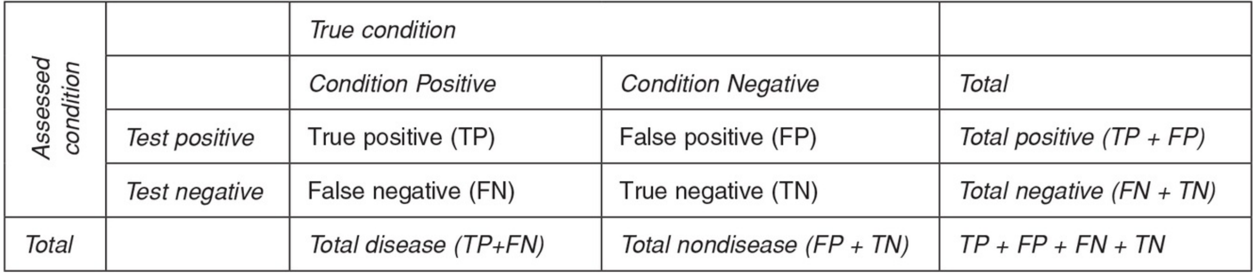
Specificity refers to a test’s accuracy at identifying those who *do not* have a condition or characteristic. It is the proportion of truly not at-risk or without condition (e.g., trait, disease, classification, and label) who are correctly identified as such through a diagnostic tool. Specificity describes the characteristic of a test in terms of how well the test correctly identifies *true negatives* (TNs) or those who do not have the predicted condition. Mathematically, it is expressed as the proportion of TN results to the sum of both true-negative and false-positive results. Mathematically, this can be expressed as:

$$px = \text{number of true negatives} / \\ (\text{number of true negatives} + \text{false positives}).$$

To better understand specificity, imagine describing a test along two dimensions depicting the relation between the predicted conditions. These dimensions can be further divided along four quadrants (see Figure 1). Quadrant 1, true positive (TP), is the number of persons with a disease who test positive. Moving clockwise, Quadrant 2 is the number of false positives (FPs) or number of well persons who test positive. Quadrant 3, TNs, depicts the number of well persons who test negative. Quadrant 4, false negatives, depicts the number of persons with a disease who test negative. Specificity is the sum of TNs (Quadrant 3) divided by the sum of TNs (Quadrant 3) and FPs (Quadrant 4).

Specificity is one of many test indices used to characterize the utility of a diagnostic or screening tool. Other useful indices include sensitivity, classification accuracy, and positive and negative predictive values (PPV and NPV, respectively). Sensitivity is the proportion of truly with or at-risk (TPs) and describes how well a test correctly identifies TPs. Sensitivity and specificity are always reported together. Classification accuracy is the proportion of TPs and TNs to the whole sample. PPV and NPV are the chance proportions that diagnostic results will be correct. Unlike specificity and sensitivity, which are characteristic of the diagnostic test and are not influenced by the population, both PPV and NPV are influenced by the proportion of a population found to have a condition. These are estimated using data from cross-sectional or other population-based studies in which valid prevalence estimates can be obtained. That is, specificity will remain unchanged as prevalence of the disease changes, whereas PPV and NPV will change as prevalence rate varies.

Figure 1 Relation between true conditions and test results



Assessed condition		*True condition*		
		Condition Positive	*Condition Negative*	*Total*
	Test positive	True positive (TP)	False positive (FP)	*Total positive (TP + FP)*
	Test negative	False negative (FN)	True negative (TN)	*Total negative (FN + TN)*
Total		*Total disease (TP+FN)*	*Total nondisease (FP + TN)*	*TP + FP + FN + TN*

Sensitivity and specificity of a diagnostic test are determined by comparing test results in a representative sample of individuals against a gold standard diagnostic where true rate of both negative test results and positive test results is known. Although the ideal test would be 100% sensitive and 100% specific in its classification, in practice, context matters. For almost any condition, there are two distributions: one for the population without condition and one for the population with condition. In most circumstances, these two distributions have overlapping scores. Unless there is perfect separation between two distributions on a particular diagnostic measure, a trade-off will have to be made in terms of improved sensitivity or specificity. In medicine, a test result indicating no disease when in fact the person carries a disease (false negative) are of primary concern. Conversely, in law the opposite is true and the primary concern is preventing false convictions (false negatives) as opposed to false acquittals (FPs). In prevention research, we tend to favor FPs, as the intervention is assumed to carry no or low risk (e.g., no harm is done to a student who receives extra math tutoring).

For interpretation, positive results from a test with high specificity are useful to diagnose the presence of a condition because the test rarely gives positive results when the condition is absent. Tests with high specificity have low type I error or false-positive rates. In contrast, negative results from tests with high sensitivity would not be useful in ruling out the condition because the test provides many FPs. In practice, however, the test result is usually all that is known. Therefore, it is important to also understand how accurate the test is at predicting whether the individual does or does not have the condition of interest. As noted earlier, neither specificity nor sensitivity are influenced by the population and will remain unchanged as the prevalence of the disease changes.

The value of a test as a diagnostic tool depends on the sensitivity and specificity of the instrument. A perfect measure would have sensitivity, specificity, and classification accuracy all equal 100%, but in reality, no diagnostic tool achieves 100% classification. As a result, there are trade-offs and as the specificity of a measure improves there is a loss in some of its sensitivity.

Kathleen Lynne Lane and Eric Alan Common

See also Sensitivity

Further Readings

Bennett, K. J., & Offord, D. R. (2001). Screening for conduct problems: Does the predictive accuracy of conduct disorder symptoms improve with age? Journal of the American Academy of Child & Adolescent Psychiatry, 40(12), 1418–1425. Retrieved from http://dx.doi.org/10.1097/00004583–200112000-00012

Gilbert, J. K., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2012). Early screening for risk of reading disabilities: Recommendations for a four-step screening system. Assessment for Effective Intervention: Official Journal of the Council for Educational Diagnostic Services, 38(1), 6–14. doi:10.1177/153450841245149

Kauffman, J. M. (1999). How we prevent the prevention of emotional and behavioral disorders. Exceptional Children, 65(4), 448–468.

Loong, T. W. (2003). Understanding sensitivity and specificity with the right side of the brain. BMJ, 327(7417), 716–719.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science, 240, 1285–1293.

Vidakovic, B. (2011). Sensitivity, specificity, and relatives. In Statistics for bioengineering sciences: With MATLAB and WinBUGS support. New York, NY: Springer.

Sarah Lockenvitz Sarah Lockenvitz Lockenvitz, Sarah

Julie Masterson Julie Masterson Masterson, Julie

Speech-Language Pathology Speech-Language pathology

1567

1570

Speech-Language Pathology

Speech-language pathology is a diverse field that encompasses communication and swallowing disorders, including impairments of speech production, fluency, voice/resonance, language, and cognition. Clinicians and professionals who specialize in the assessment and treatment of speech-and language-based communication disorders and swallowing disorders are known as speech-language pathologists (SLPs); SLPs work in a variety of settings, including medical and educational settings. Practitioners of speech-language pathology who work directly with patients are responsible for assessing for the presence of disorders, diagnosing disorders, developing and implementing goals and treatment plans, and discharging and following up with patients when appropriate.

Speech-language pathology has progressed from basing clinical methods on principles that seem intuitive and anecdotally supported to emphasizing the importance of evidence-based practice. Clinical experience plays a role in evidence-based practice, but equally important are careful consideration of the best research evidence and the values of the patients and their families. The evidence can be both experimental research (emphasis on controlled and replicable studies) and qualitative (emphasis on descriptive, contextualized social experiences). Through an increasing body of strong evidence underlying clinical principles, speech-language pathology has established itself as a reputable field of scientific study grounded in evidence-based practice.

This entry discusses speech-language pathology with regard to service areas and settings and reviews SLP credentialing requirements.

Service Delivery Areas

Clinicians and professionals who practice speech-language pathology may complete screenings, assessments, treatment, and research related to disorders in the following service delivery areas.

Speech Production

Disorders of speech production affect articulation of sounds or classes of sounds and/or the planning and execution of fluent sound production. The term *phonology* refers to the rules and patterns governing the speech sound system in a language; thus, errors in speech production are considered to fall under the umbrella term *phonological disorders*. These problems can result from motor-based difficulties, structural abnormalities, or no clear etiology.

In early development, children may struggle to produce a component of the word structure (e.g., final sounds). Later, the errors tend to occur on sounds that are harder to articulate, and a child may substitute a related sound that is not quite so challenging. For example, children often may use a “w” for an “r” (e.g., WABBIT for *rabbit*). In some cases, speech production errors occur due to an individual’s inability to use the appropriate linguistic pattern even though the ability to articulate a sound or sound class is intact. For example, a child might substitute a “th” sound for an “f” sound (FUM for *thumb*) yet use the “th” sound when attempting to produce words containing an “s” sound (THUM for *some*). It is this linguistic and representational nature of speech sounds that results in phonology spanning both speech and language.

Fluency

Disorders of fluency involve the disruption of the flow of speech. Stuttering is characterized by atypical disfluencies found in the speech flow, such as sound repetitions (e.g., “I have a *b-b*-ball”), syllable repetitions (e.g., “I’m watching *tel-tel*-television”), sound prolongations (e.g., “That’s my *ssssss*sister”), and blocks or difficulty with initiating words. Cluttering is characterized by a perceived rapid speech rate and/or syllable deletions, syllable collapses (e.g., “I’ll doitmorrow”), or deletions of word endings.

Voice/Resonance

Voice/Resonance

Voice disorders may affect phonation, the process by which airstream flow causes the vibration of the vocal folds and results in the sustaining of vocalizations. Disorders of voice may also affect pitch (i.e., perceptual highness or lowness based on the fundamental frequency of the voice of a given individual) and loudness (i.e., perceptual loudness or quietness based on the amplitude of the voice of a given individual). Resonance disorders involve either an excess (hypernasality) or a deficit (hyponasality) of airflow through the nasal cavity.

Language and Literacy

Disorders of language involve difficulties with form or function (or both) within a given language. Individuals may struggle with morphology, which is the system that governs the combination of elements to form words (e.g., using the “s” to signal plural or the “ed” to indicate past tense). Examples of difficulties in syntax (sentence structure) include failure to include a word (“She going” instead of “She is going”) or errors in word order (“No mommy go” for “Mommy don’t go”). Pragmatics refers to the social use of language in a broader context such as conversation and other interactive situations. Problems in this area may include lack of eye contact or the inability to consider the appropriateness of conversational topics. Paralinguistic communication involves communication through means other than spoken language, such as gestures and signs.

SLPs consider not only spoken language abilities but also written language skills. This includes single word reading (decoding) and comprehension skills. It also includes writing composition and spelling. Both decoding and spelling involve appreciation for the sound composition in words (phonological awareness), awareness of meaning elements in words (morphological awareness), the mapping of spellings to sounds and meaning units (orthography), and the ability to store representative “mental pictures” of words.

Cognition

Disorders of cognition may affect any combination of the following cognitive processes: (a) orientation, the awareness of person, place, and time; (b) attention, the select concentration on a particular piece of information; (c) memory, the

the select concentration on a particular piece of information; (c) memory, the encoding, storage, and retrieval of declarative, procedural, and experiential information; (d) problem solving, the ability to reason through a particular problem and determine appropriate solutions; and (e) executive functioning, the ability to monitor, plan, and organize information to facilitate the achievement of cognitive goals. Examples of cognitive disorders that may disrupt these processes include dementia and traumatic brain injury.

Swallowing

Swallowing disorders, collectively known as *dysphagia,* are concerned with challenges of safely and/or efficiently moving nutritional material from the lips to the stomach. Disorders in swallowing can result in respiratory complications such as pulmonary infection, safety concerns such as choking, and nutrition and hydration deficiencies. Contributing to these disorders are factors such as decreased strength, range of movement, poor coordination, impaired sensory awareness/alertness, and abnormalities of the involved musculature and structures. Management may involve compensatory techniques and postural adjustments, and rehabilitation often includes strategies to strengthen relevant muscles, develop coordination, and improve sensory feedback of the alimentation mechanism.

Additional Information on Service Delivery Areas

Each of the areas previously described is not necessarily a discretely isolated aspect of speech-language pathology. An unresolved speech production disorder in a child puts that child at risk of encountering difficulties with literacy. Aphasia, a neurogenic language disorder caused by injury to the brain, frequently is associated with dysphagia in victims of cerebrovascular accidents. Traumatic brain injury patients may require intervention to target both cognitive functioning and pragmatic language. It is critical for the SLP to be part of an interdisciplinary team that assesses a potential client thoroughly and comprehensively, staying alert for concomitant disorders with varying degrees of relatedness.

SLPs work with patients with communication and swallowing disorders stemming from a number of etiologies. These include but are not limited to neonatal damage or dysfunction (e.g., cerebral palsy), developmental disabilities (e.g., specific language impairment, autism spectrum disorders), orofacial (e.g., specific language impairment, autism spectrum disorders), orofacial anomalies (e.g., cleft palate), laryngeal anomalies (e.g., vocal polyps or nodules), neurological dysfunction (e.g., cerebrovascular accident, traumatic brain injury, dementia), genetic disorders (e.g., Down syndrome), and unknown etiologies (e.g., functional articulation disorders).

Other services falling within the scope of practice for speech-language pathology include modification of dialect, accent, and other speech patterns. These services are elective instead of therapeutic and may be sought out by individuals wishing to improve their proficiency with Standard American English for professional reasons, as Standard American English is the dialect used by such entities as the government and the media in the United States. The SLP must be able to distinguish between typical individual speech and language variation and diagnosable disorders. While individuals may request intervention from an SLP for dialect modification, dialect and accent differences must not be interpreted to be disorders. Dialect features are systematic and rule based, and they carry no inherent social value. A true disorder of speech or language cannot be diagnosed as such until the possibility of influence from dialect or accent on the communication patterns in question is rejected. For patients whose speech and/or language is affected by both dialect influences and a disorder, some communication patterns may be a result of the difference and some may be a result of the disorder. Unless elective services for dialect modification are requested, the SLP should determine the source of all communication patterns and treat only those that are a result of the disorder.

Another example of a preventive or elective service area is the provision of education regarding vocal hygiene and appropriate use of voice, which may be implemented for individuals such as public speakers or professional vocalists who are at risk for vocal trauma. They may be educated on appropriate use of strategies including vocal rest and hydration.

SLPs also may have undergone instruction in basic audiology and audiologic habilitation and rehabilitation and thus may be qualified to engage in related services such as conducting hearing screenings, working with hearing impaired individuals to improve articulation, and teaching communication through sign language. For patients who rely on nonoral communication, SLPs may offer training for using devices and systems of augmentative and alternative communication, either unaided (such as gestures and sign) or aided (such as communication boards or speech synthesizers).

Other Areas of Practice

The practice of speech-language pathology is not limited to clinical assessment and treatment of speech-and language-based communication disorders and swallowing disorders. SLPs may engage in advocacy for either individual patients or for the entire discipline through the provision of education and training. The purpose of advocacy can range from increasing awareness to influencing policy makers to enact change at local, state, and national levels to reduce or eliminate communicative barriers. Through counseling and appropriate referral, SLPs may work to educate, guide, support, and empower patients and their families and caregivers. The SLP may serve as an academic educator at higher education institutions and may conduct research related to communication disorders and swallowing. Mentoring, supervising, and training of beginning clinicians may be a responsibility of the credentialed SLP, and similar supervisory roles may include overseeing other support personnel (i.e., speech-language pathology assistants [SLPA]). In addition, SLPs may hold administrative positions in various settings, and consequently their service may extend beyond the field of speech-language pathology.

Settings and Collaboration

Speech-language pathology is practiced in a number of different settings. The SLP can find employment in education settings from early intervention to higher education and in medical settings, such as intensive care units and inpatient facilities located in hospitals, long-term acute care units, rehabilitation centers, skilled nursing facilities, outpatient clinics, and home health. SLPs also deliver services through private practice, psychiatric centers, health departments, research agencies, the military, group homes, and telepractice. Some settings necessitate the specialization of working with certain populations, which may be general (e.g., children or adults) or more specific (e.g., craniofacial teams providing ongoing care for cleft palate). A patient may demonstrate the need for a care team, which will require the SLP to collaborate with other professionals. Such professionals may include but are not limited to general education teachers, special education teachers, principals, psychologists, physicians, radiologists, nurses, social workers, audiologists, physical therapists, occupational therapists, and other SLPs.

Credentialing of the SLP

In the United States, the American Speech-Language-Hearing Association (ASHA) is the professional association for SLPs. The Council for Clinical Certification is affiliated with ASHA, but it autonomously sets the standards for individual certification, the Certificate of Clinical Competence in Speech Language Pathology (CCC-SLP). Attainment of the CCC-SLP requires the completion of a postbaccalaureate degree from an institutional program that is accredited by the Council for Academic Accreditation, which also is affiliated with ASHA but autonomously sets the standard for accrediting master's programs in speech-language pathology. Candidates for the CCC-SLP also must pass a national examination and complete a postgraduate professional experience supervised by an SLP who holds the CCC-SLP. This certification must be maintained periodically by participating in continuing education experiences. Individual states require a license to practice speech-language pathology, and the license is awarded based on each state's own initial and maintenance requirements. In most states, the requirements for licensure mirror those in place to earn the CCC-SLP.

SLPs are sometimes supported by an SLPA. The duties that SLPAs are allowed to perform vary somewhat across states, but they always must be implemented under the supervision of a licensed SLP. Educational requirements to practice as a supervised SLPA also vary from state to state. Certification and licensure may be suspended or revoked if an SLP does not abide by ASHA's code of ethics, which requires SLPs to prioritize the welfare of their patients, to perform professionally and competently, to present information accurately, and to promote collaborative relationships between ASHA's overseen professions (speech-language pathology; audiology; speech, language, and hearing science; related support personnel; and students).

Sarah Lockenvitz and Julie Masterson

See also Autism Spectrum Disorder; Developmental Disabilities; Evidence-Based Interventions; Literacy; Qualitative Research Methods, Quantitative Research Methods

Further Readings

Clinical Topics and Disorders in Speech-Language Pathology. (2016). The American Speech-Language-Hearing Association. Retrieved from http://www.asha.org/slp/clinical/

National Institute on Deafness and Other Communication Disorders. (2016). Statistics on voice, speech, and language. Retrieved from https://www.nidcd.nih.gov/health/voice-speech-and-language

Owens, R. E., Farinella, K. A., & Metz, D. E. (2015). Introduction to communication disorders: A lifespan evidence-based perspective (5th ed.). Boston. MA: Pearson.

Professional Practice Issues. (2016). The American Speech-Language-Hearing Association. Retrieved from http://www.asha.org/slp/practice-issues/

Scope of Practice in Speech-Language Pathology. (2016). The American Speech-Language-Hearing Association. Retrieved from http://www.asha.org/policy/SP2016–00343/

Zipoli, R. P., & Kennedy, M. (2005). Evidence-based practice among speech-language pathologists. American Journal of Speech-Language Pathology, 14, 208–220. doi:10.1044/1058–0360(2005/021)

Eduardo Estrada Eduardo Estrada Estrada, Eduardo

Speeded Tests

Speeded tests

Speeded Tests

In the context of educational measurement, the term *speeded test* (or *speed test*) refers to a measuring tool composed of a list of relatively easy items, intended to be answered in a very limited time. When applying a speeded test, it is common to ask (or even force) the test takers to solve the items sequentially from the first to the last one. If the difficulty level and time limit are correctly set, none of the test takers will be able to reach the last item before the time limit is reached. The total score is usually computed as the number of items correctly answered when the time limit is met, and the differences in the scores are mainly attributed to individual differences in speed. In 1950, Harold Gulliksen proposed the term in his book *Theory of Mental Tests*, together with the opposite concept of power test. This entry describes what a speeded test is, explains what speeded tests are typically used for, provides some examples of commercial speeded tests, and explains how speeded test are related to the concept of test speediness.

Speeded Tests for Measuring Basic Cognitive Abilities

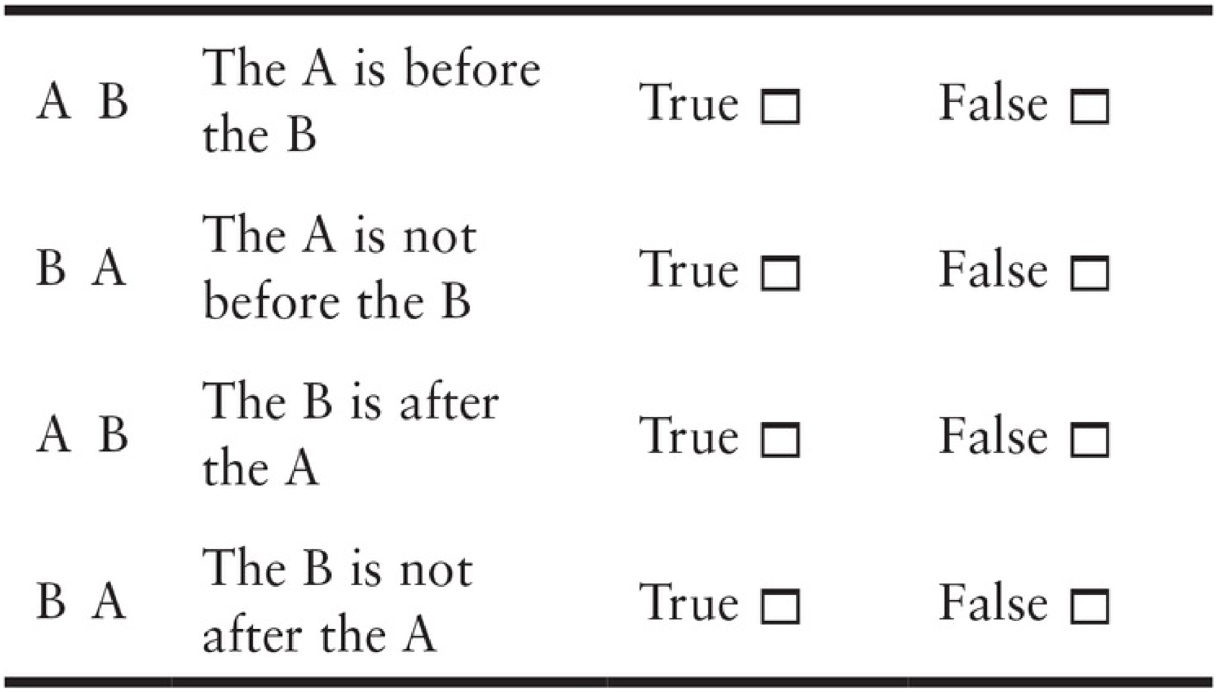
Speeded tests are often used to measure basic cognitive skills such as processing speed, reaction time, or visual search. Two examples of this are the symbol search test and the coding test of the Weschler Intelligence Scales (Wechsler Adult Intelligence Scale for adults and Intelligence Scale for Children for children). In the symbol search test, a visual pattern (or symbol) is provided as a reference. The test taker must search for exact copies of this symbol within a large set of similar symbols. The score is computed as the number of exact copies found when the time limit is reached. In the coding test, the examinee is provided with an arbitrary coding key pairing symbols and numbers. Then, a set of symbols is shown, and the examinee must translate them into the correct numbers, according to the coding key. The score is computed as the number of digits correctly decoded when the time limit is reached. Although these two tests are designed to measure speed of visual processing, most authors agree that they also tap other mental skills such as short-term memory and paired associates learning.

Speeded Tests for Measuring High-Level Cognitive Abilities

Despite speeded tests often being used for assessing basic mental skills such as processing speed, some speeded tests have been proposed for a quick evaluation of higher level complex mental abilities. One example of this is the Baddeley's three-minute reasoning test, created for efficiently measuring verbal intelligence in research contexts. Another example is the Wonderlic personnel test (WPT) widely used in the organizational field in the United States for a quick estimation of general intelligence.

The Baddeley's three-minute reasoning test is composed of a long list of sentences like that shown in Table 1: All the items are easy enough to be correctly answered by any competent English reader, although some sentences are more complex than others. The score of the test is computed as the number of questions correctly answered in 3 minutes. Despite its simplicity, Alan D. Baddeley found a correlation of 0.59 between the scores of this test and the scores of the British Army verbal intelligence test, which takes around 1 hour to be completed. Based on these data, some authors hold that this test can provide an efficient and nondetailed estimation of general verbal ability.

The WPT is a brief commercial battery composed of different type of items. It measures verbal ability (e.g., identifying the antonym of a word), mathematic ability (e.g., find out which number is the lowest one within a set), and logical reasoning (e.g., select the next element in a logical series). All the items are easy to solve for most adults, but the different types of items are mixed during the test administration, and the examinees have 12 minutes to solve 50 questions. Consequently, they are forced to change from one type of reasoning to another, and do it quickly. The score is computed as the number of questions correctly answered when the time limit is reached. WPT creators have found high correlations between the WPT's scores and the Full-Scale Intelligence Quotient scores from the Weschler Adult Intelligence Scales battery and other general intelligence tests. Based on these data, they hold that the WPT can be used for a quick and efficient evaluation of general intelligence.



A B	The A is before the B	True □	False □
B A	The A is not before the B	True □	False □
A B	The B is after the A	True □	False □
B A	The B is not after the A	True □	False □

Test Speediness

Although all speeded tests are designed to capture differences in respondents' speed, it is virtually impossible to leave out some skills considered as power abilities. For example, it seems clear that an antonym item from the WPT also measures vocabulary knowledge, and the questions from Baddeley's three-minute reasoning test measures grammatical ability (which is a part of the general verbal ability). Even the speed test intended to measure more basic cognitive skills tap some related power abilities. Some authors have proposed that most tests are partly power and partly speed test in unknown proportions. The concept of "test speediness" or "test speededness" has been defined as the extent to which the time restrictions on a maximum performance test have an impact on the test takers' achievement.

Various methods have been proposed for studying the speediness of a test. These methods try to isolate what proportion of the scores' variance is due to the speed and power components. Some methods use information external to the test such as response time measures, whereas others rely only on information provided by the test, such as the proportion of unreached items.

the test, such as the proportion of unreached items.

Eduardo Estrada

See also Power tests; Wechsler Intelligence Scales

Further Readings

Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. Psychonomic Science, 10, 341–342. doi:10.3758/BF03331551

Chadha, N. K. (2009). Speed test versus power test. In Applied Psychometry (pp. 39–48). New Delhi: Sage. doi:10.4135/9788132108221.n Estrada, E., Román, F. J., Abad, F. J., & Colom, R. (2017). Separating power and speed components of standardized intelligence measures. Intelligence, 61, 159–168. doi:10.1016/j.intell.2017.02.002

Gulliksen, H. (1950). Theory of mental tests (Vol. xix). Hoboken, NJ: Wiley.

Hunt, E. (2011). The tests. In Human Intelligence (pp. 31–78). New York, NY: Cambridge University Press.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. Educational Measurement: Issues and Practice, 26, 29–37. doi:10.1111/j.1745–3992.2007.00106.x

Wonderlic, E. F., & Wonderlic, C. F. (1992). Wonderlic personnel test user’s manual. Libertyville, IL: Wonderlic Personnel Test.

Qingqing Zhu Qingqing Zhu Zhu, Qingqing

Patricia A. Lowe Patricia Lowe Lowe, Patricia

Split-Half Reliability Split-Half reliability

1572

1574

Split-Half Reliability

Split-half reliability is a statistical method used to measure the consistency of the scores of a test. It is a form of internal consistency reliability and had been commonly used before the coefficient α was invented. Split-half reliability is a convenient alternative to other forms of reliability, including test–retest reliability and parallel forms reliability because it requires only one administration of the test. As can be inferred from its name, the method involves splitting a test into halves and correlating examinees' scores on the two halves of the test. The resulting correlation is then adjusted for test length using the Spearman-Brown prophecy formula. This entry introduces the basic principles and estimation procedures for this method and discusses its limitations.

Basic Principles and Estimation Procedures

According to classical test theory, variations in examinees' test scores are due to (a) variations in the test takers' true ability or trait and (b) error. The proportion of variation in the total score, resulting from variations in the examinees' true ability, is defined as test reliability. Traditionally, researchers have estimated reliability by administering a test twice to the same examinees and correlating their scores obtained at the 2 times (test–retest reliability) or administering two parallel forms of a test to test takers and correlating their scores on the two forms (parallel forms reliability). Both of these methods have limitations because it is not always feasible to administer a test multiple times and not all tests have multiple forms. One convenient alternative is to split a test in half and use each half as a parallel form of the other. Comparing scores on the two halves is then another way to measure test reliability. This method, referred to as split-half reliability, is considered a measure of the internal reliability of a test or how consistently the items perform within a test. The underlying assumption is if a test measures a single construct, then individuals should perform equally well on both halves of the test.

To estimate split-half reliability, the first step is to split the test in half and administer the two halves to the examinees. If there are multiple subscales or content areas assessed within a single test, split-half reliability should be calculated for each subscale or content area separately. When a test has more than 2 items, there are apparently multiple ways to split it. The principle is to obtain two halves as equivalent as possible. One could consider using the middle item (e.g., the fifth item on a 10-item scale) as the dividing point or randomly divide the test items into two groups which would represent the two halves of the test. However, because many tests organize items by difficulty level, these methods could easily lead to nonequivalent halves, which would result in an underestimation of the test reliability. Furthermore, attention and fatigue may affect individuals' performance differently at the beginning and at the end of the test. That is, individuals may be more alert when they take the first half of a test, but more tired when they take the second half of the test. For these considerations, a commonly used approach is to split the test by even-and odd-number items. Another approach is to manually balance the difficulty level in the two halves of the test. Once the test is split in half and administered to the examinees, a Pearson correlation coefficient is calculated between the scores of the two halves of the test.

A simple correlation of the two halves of the test, however, is not comparable to the parallel forms reliability because everything else being equal, a longer test results in higher reliability. Therefore, the split-half reliability estimation, which was calculated between the scores of the two halves of the test, involves an additional step in which the correlation is corrected for test length using the Spearman-Brown prophecy formula , where r is the reliability of the current test, N is the number of times the current test is lengthened, and $r_{\text{predicted}}$ is the predicted reliability of the new test. In 1910, both Charles Spearman and William Brown independently published this formula to predict the reliability of a test when it is expanded N times by adding new items with the same psychometric properties as the items that appeared in the original (i.e., old) test. When estimating split-half reliability today, the half tests are considered to be the “current” test that needs to be expanded twice to reach the original length of the test before it was split in two. Therefore, in this process, N equals 2 and the formula can be simplified as , with r representing the correlation between the two halves and $r_{\text{predicted}}$ being the split-half reliability.

Limitations

The most common limitation discussed about the split-half reliability is that depending on the method used to split the test, one can obtain different values for the reliability coefficient. Although efforts can be taken to ensure the two halves are as equivalent as possible, such a result is not guaranteed, and there is no splitting method that is statistically optimal. Because of this limitation, other methods, such as Cronbach's coefficient α, are often considered more appropriate measures of internal consistency reliability of the scores of a test. Also, because coefficient α is mathematically the average of the correlations between all possible halves of a test, it has replaced the split-half reliability in most cases.

Qingqing Zhu and Patricia A. Lowe

See also Classical Test Theory; Coefficient Alpha; Internal Consistency; Pearson Correlation Coefficient; Reliability; Spearman-Brown Prophecy Formula; Test–Retest Reliability

Further Readings

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Brownell, W. A. (1933). On the accuracy with which reliability may be measured by correlating test halves. The Journal of Experimental Education, 1, 204–215.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York, NY: Macmillan.

Spearman, C. (1910). Correlation calculated from faculty data. British Journal of

Psychology, 3, 271–295.

Heidi Arnouts Heidi Arnouts Arnouts, Heidi

Split-Plot Design Split-Plot design

1574

1576

Split-Plot Design

A split-plot design is an experimental design in which the levels of one or more experimental factors are held constant for a batch of several consecutive experimental runs, which is called a whole plot. The levels of the remaining factors are varied during these experimental runs, and each level combination is considered as a subplot within the whole plot. Split-plot designs therefore consist of two types of experimental units: whole plots and subplots, where the subplots are nested within the whole plots. A split-plot design results in correlated responses for the experimental runs in the same whole plot. A correct analysis of the data from split-plot designs should take this correlation into account. In practice, split-plot designs are often used inadvertently, thereby often ignoring the typical split-plot correlation. The resulting statistical analysis is then inappropriate. This entry discusses the agricultural origin of split-plot designs as well as the traditional industrial applications of this type of design. Finally, the difficulties of using this powerful design in educational settings are presented.

Agricultural Split-Plot Designs

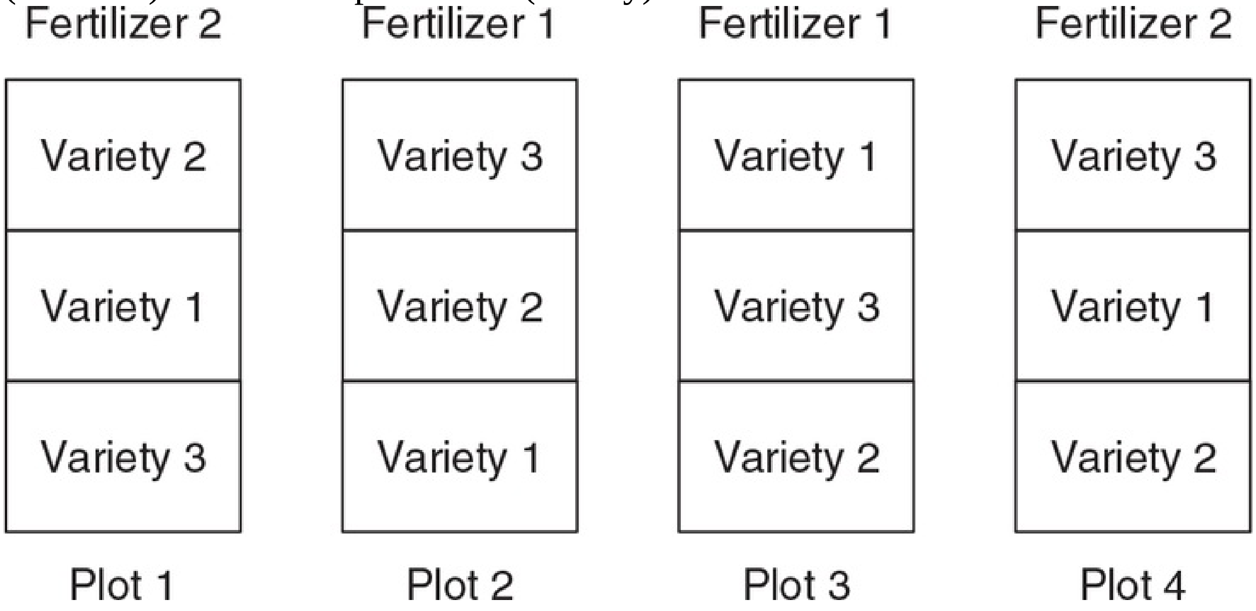
The terminology used in the context of split-plot designs comes from its initial agricultural applications, where experiments were performed on different plots of land. In these experiments, the levels of one or more experimental factors were allocated to large plots of land, also referred to as whole plots. The levels of the remaining experimental factors were assigned to smaller plots of land or subplots.

A typical example of a split-plot design in agriculture is an experiment for investigating the effect of different fertilizers and varieties on the yield of crops.

Because these fertilizers are often sprayed from planes, a whole plot of land must be treated with the same type of fertilizer. Next, each large plot of land can be divided into smaller plots on which the crop varieties can be planted. The larger plots of land are called whole plots, whereas the smaller plots are referred to as subplots. In this experiment, the factor fertilizer is the whole-plot factor because the levels of the factor fertilizer are applied to whole plots. The second factor, crop variety, is named the subplot factor because it is applied to subplots. Figure 1 provides a graphical representation of a split-plot design in which the whole-plot factor, fertilizer, has two possible levels, and the subplot factor, variety, has three levels. This specific design involves four whole plots (labeled plot 1–4), each of which is divided into three subplots.

In this specific example, the split-plot design involves one whole-plot factor and one subplot factor; however, split-plot designs involving more than one whole-plot factor and/or subplot factor are also possible. In the split-plot design in Figure 1, a complete random assignment of the factor-level combinations to the subplots is impossible because all subplots within the same whole plot should be treated with the same fertilizer. Typically, there are two levels of randomization in a split-plot design. First, the level combinations of the whole-plot factors are randomly assigned to the whole plots. Next, the level combinations of the subplot factors are randomly assigned to the subplots within each whole plot.

Figure 1 Agricultural split-plot design involving one whole-plot factor (fertilizer) and one subplot factor (variety)



The split-plot designs in agricultural experiments are usually classical split-plot designs (i.e., all combinations of levels of the subplot factors occur within each combination of levels of the whole-plot factors). Moreover, the whole-plot and subplot factors in a classical split-plot design are traditionally treated as categorical, allowing for an analysis of variance–based approach to analyze the data. The analysis should, however, take into account that the experimental runs performed in the same whole plot are correlated. Therefore, the analysis of variance model should contain a random whole-plot effect that represents the variation in the response due to the fact that an experimental run is performed in a certain whole plot.

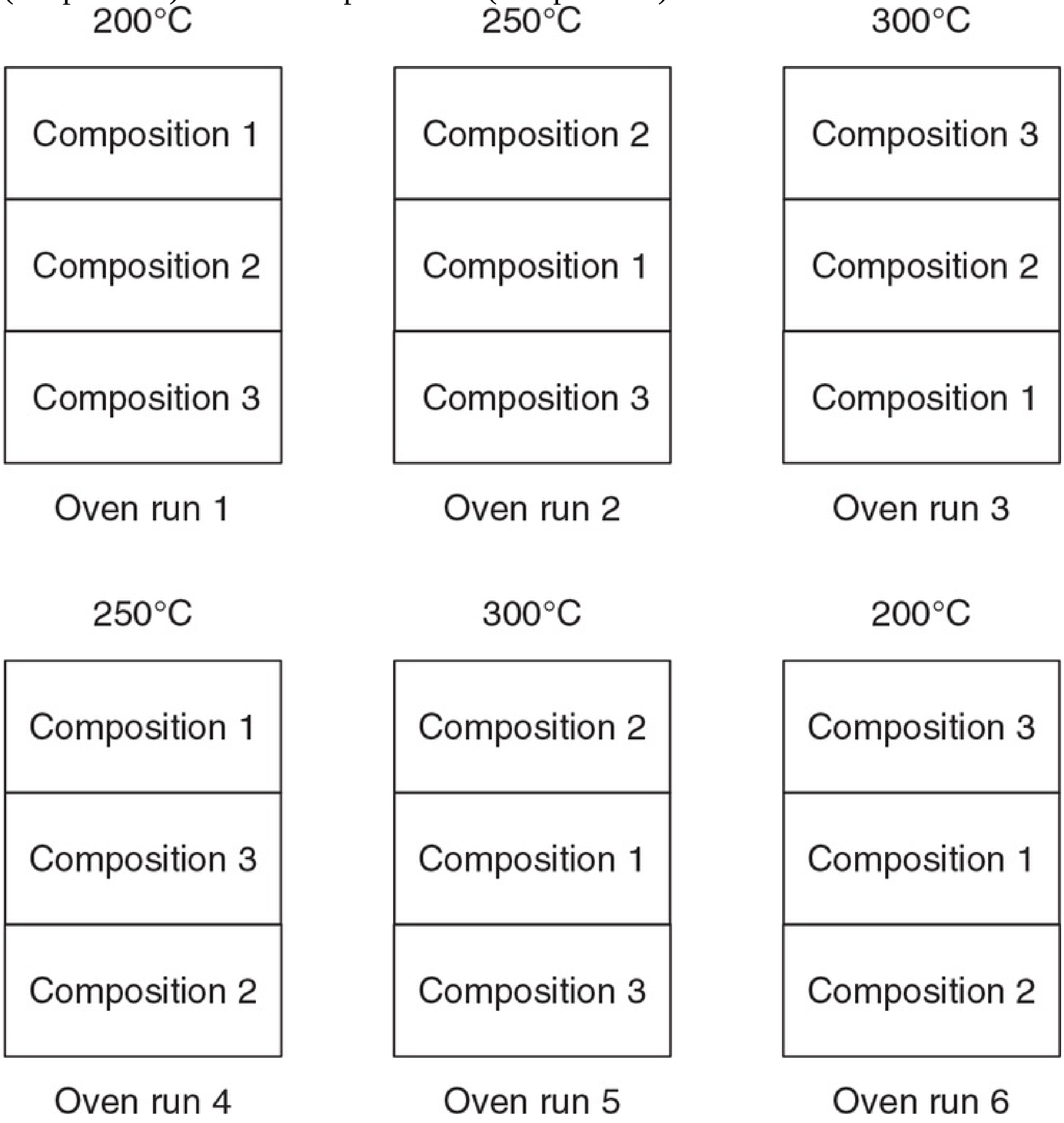
Industrial Split-Plot Designs

In industrial experiments, it frequently happens that some of the experimental factors are held constant for a number of successive runs and a split-plot design is applied. This may be due to the fact that these experimental factors are expensive or time-consuming to change, or it may be due to the fact that the experiment is run in large batches and the batches can be subdivided later for additional treatments.

A typical example of an industrial split-plot design is an experiment to check the influence of oven time and material composition on the strength of a certain component. Assume that the experimental factor oven temperature has three different levels (200°C, 250°C, and 300°C, respectively), and there are also three possible material compositions available. Because it is time-consuming to reset the oven and to reach the required temperature, several components are heated simultaneously in the same oven. More particularly, for each temperature, three components with randomly assigned material compositions are randomly arranged in the oven and heated together. A graphical representation of the industrial split-plot design is given in Figure 2. In the given industrial split-plot design, each temperature level is replicated twice. The six different oven runs are the whole plots of the split-plot experiment and the three positions in the oven are the subplots. Consequently, in this design, oven temperature is the whole-plot factor and material composition is the subplot factor.

The split-plot design in Figure 2 is again an example of a classical split-plot design. This is, however, usually not the case for industrial split-plot designs because they often involve several quantitative whole-plot and subplot factors and, due to time and cost constraints, are limited in size. As a result, in many industrial split-plot designs, only a fraction of all combinations of subplot factor levels appear in every whole plot and a more general regression-based modeling approach is necessary. Again, a correct statistical analysis will demand for a random whole-plot effect in the regression model, and the generalized least-squares estimator is the best linear unbiased estimator of the model parameters.

Figure 2 Industrial split-plot design involving one whole-plot factor (temperature) and one subplot factor (composition)



Educational Split-Plot Designs

Educational Split-Plot Designs

In educational applications, split-plot designs often use the same grouping schemes as shown in the industrial and agricultural approaches. The “whole plot” might be a school or all the students who have a certain teacher. Often the broader group comparison involves one teaching method versus another or a demographic characteristic. Repeated measures or observations at different times might involve a change in conditions or a growing increase in “dose” or exposure to a teaching method, or it could be simply a developmental change over time. Correct applications of a true split-plot design are difficult in educational research because it is often difficult or impossible to randomly assign participants to different groups or levels of a categorical independent variable. Consequently, statistical approaches that attempt to control for, or account for, confounding variables are more likely to be used, especially when researchers are interested in accounting for growth or change across time.

Heidi Arnouts

See also Analysis of Variance; Correlation; Experimental Designs; Random Assignment

Further Readings

Bisgaard, S. (2000). The design and analysis of 2^{k-p} x 2^{q-r} split-plot experiments. Journal of Quality Technology, 45, 39–56.

Box, G. E. P., & Jones, S. P. (1992). Split-plot designs for robust product experimentation. Journal of Applied Statistics, 19, 3–26. Retrieved from http://dx.doi.org/10.1080/02664769200000001

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, Scotland: Oliver & Boyd.

Goos, P., & Jones, B. (2011). Optimal design of experiments: A case study approach. New York, NY: Wiley.

Matthew Gordon Ray Courtney Matthew Gordon Ray Courtney Courtney, Matthew Gordon Ray

SPSS

SPSS

1576

1583

SPSS

IBM SPSS Statistics is a statistical software package commonly used for statistical analysis in the social sciences. It eliminates time-consuming data preparation tasks and provides predictive and comparative data insights via a menu-driven (and syntax-driven), user-friendly interface. The program is not only ubiquitous in academic settings, such as on education and social science campuses, but also widely used in various other sectors such as biology, economics, and business. After presenting the history of this product, this entry reviews various features and uses of SPSS.

History

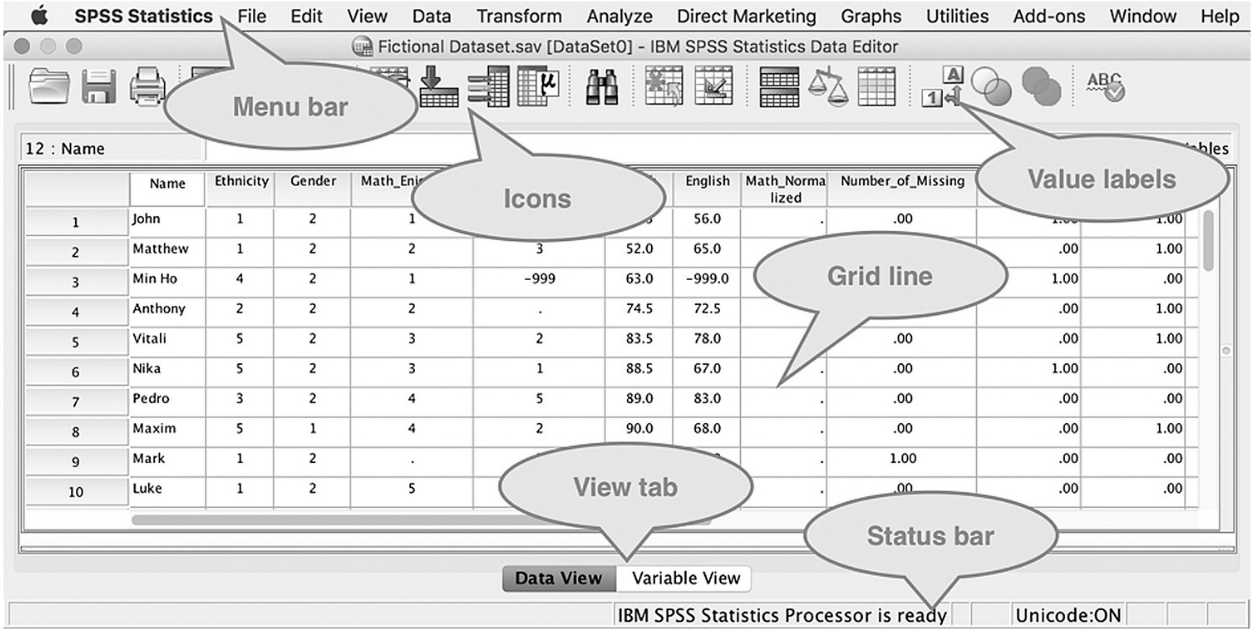
SPSS can be traced back to 1967 when, according to the *Chicago Tribune*, Stanford University PhD candidate Norman Nie became “frustrated trying to use a computer to analyze data.” After Nie took detailed technical notes to fellow academics Dale Bent (an expert in file structures) and Hadlai Hull (an expert in coding), the 1968 version of SPSS was born.

Throughout the 1970s, SPSS was available for use on a variety of mainframe computer systems and the developers sought to ensure that the program was easy to use for academics who were not computer savvy. By the mid-1980s, SPSS became the first software package of its kind to become available on personal computers, and use on U.S. and overseas university campuses became more widespread. In the 1990s and early 2000s, the company diversified from academic settings into the business intelligence software market. In 2009, IBM acquired SPSS and has further developed and extended its suite of functions. As of 2017, IBM SPSS Statistics is available in all operating systems.

The Data Editor, Viewer, and Syntax Windows

Upon opening IBM SPSS, the user will be presented with the *Data Editor* window (Figure 1). The data are presented in the center of the window, whereas the window also presents the menu bar, icons, grid lines (between cells), view tab, and status bar. The data in the Data Editor can be window is the *Viewer* window, which displays output from analyses carried out. The Viewer window can archive all outputs and is saved with the “.spo” extension. Finally, the *Syntax* window displays the command language. Generally researchers use the dialogue boxes to set up commands and do not see the operating syntax running “under the hood.” If a researcher wishes to see or manipulate the syntax underlying any procedure, the researcher can click on the *Paste* option provided at the bottom of each operating dialogue box and run the procedure from there.

Figure 1 The IBM SPSS Data Editor



Fictional Dataset.sav [DataSet0] - IBM SPSS Statistics Data Editor

12 : Name

	Name	Ethnicity	Gender	Math_Eni			English	Math_Normalized	Number_of_Missing		
1	John	1	2	1	[illegible]	[illegible]	56.0	.	.00	[illegible]	1.00
2	Matthew	1	2	2	3	52.0	65.0	[illegible]	[illegible]	.00	1.00
3	Min Ho	4	2	1	−999	63.0	−999.0	[illegible]	[illegible]	1.00	.00
4	Anthony	2	2	2	.	74.5	72.5	[illegible]	[illegible]	.00	1.00
5	Vitali	5	2	3	2	83.5	78.0	[illegible]	.00	.00	1.00
6	Nika	5	2	3	1	88.5	67.0	.	.00	1.00	.00
7	Pedro	3	2	4	5	89.0	83.0	.	.00	.00	.00
8	Maxim	5	1	4	2	90.0	68.0	.	.00	.00	1.00
9	Mark	1	2	.	[illegible]	[illegible]	[illegible]	.	1.00	.00	.00
10	Luke	1	2	5	[illegible]	[illegible]	[illegible]	.	.00	.00	.00

IBM SPSS Statistics Processor is ready | Unicode:ON

Source: IBM Corporation. (2012).

The SPSS Menu Bar

On the Data Editor window, the menu bar consists of 12 go-to options that help the user make use of the SPSS program. *File* on the menu bar includes typical user functions such as Open, Save, Export, Exit, and Print. *Edit* includes the Undo/Redo (with limited memory), Cut, Copy, Paste, Clear, and Find options. *View* enables researchers to select or deselect the icons, grid lines, and status saved in a file with the extension ".sav." Besides the Data Editor window, the other commonly used bar on the Data Editor window. *View* also allows the user to adjust font size and display raw data or value labels.

Procedures related to the manipulation of data (and, by association, the generation of command language) can be carried out by following the prompts associated with the *Data*, *Transform*, *Analyze*, *Graphs*, *and Utilities* options on the menu bar. *Data* allows the researcher to identify duplicate cases, merge files, perform propensity score matching, split the file for separate analyses, and weight cases, among other functions. *Transform* allows the user to compute and anonymize variables, rank cases, and recode variables (e.g., recoding the word *male* to 1 and *female* to 2), among a host of other options. *Analyze* includes the various commands used to carry out descriptive and statistical analysis, whereas *Graphs* includes the various options to create graphs, charts, and plots. *Utilities* provides for an efficient way to view all information about a variable (through the *Utilities* → *Variables* option). Among other options, the *Utilities* → *Custom Dialogues* and *Extension Bundles* options enable the user to trial more sophisticated procedures made available through plug-ins with other statistical programs, such as R.

On the far right of the menu bar, *Add-ons* provides programs that can be added to the base package. *Window* allows the user to quickly toggle between the Data Editor, Viewer (output), and Syntax windows. Finally, *Help* provides tutorials and assistance.

The Basics

The suite of functions available in SPSS are broad and can assist in the entire analytical process from data collection, preparation, transformation, analysis, and reporting of educational measurement and evaluation data. The program can import data from other programs such as Microsoft Excel, making the transfer of information generated from survey platforms, such as www.surveymonkey.com, quite easy.

For the purpose of illustrating some basic functions of the program, let us conceive a fictional educational data set that includes 10 students' names, ethnicity (coded 1 = *White*, 2 = *Black*, 3 = *Hispanic*, 4 = *Asian*, and 5 = *other*), socioeconomic status (coded 1 = *low*, 2 = *medium*, 3 = *high*), gender (coded 1 = *male*, 2 = *female*), math and English enjoyment levels (coded 1 = *none*, 2 = *a little*, 3 = *moderate level*, 4 = *high*, 5 = *very high*), and percentile scores on a math and English test. To define and visually inspect the data, the Data Editor provides two general user views, the *Data View* and *Variable View* (via the View tab, Figure 1). In *Data View* (Figure 2), the researcher can easily cut and paste data directly from other programs (such as Excel) into the SPSS program. The *Variable View* (Figure 3) defines each variable's *type*, *measure*, and other relevant aspects. From within both views, copy and paste functionality works for users to quickly and easily input and set up data sets.

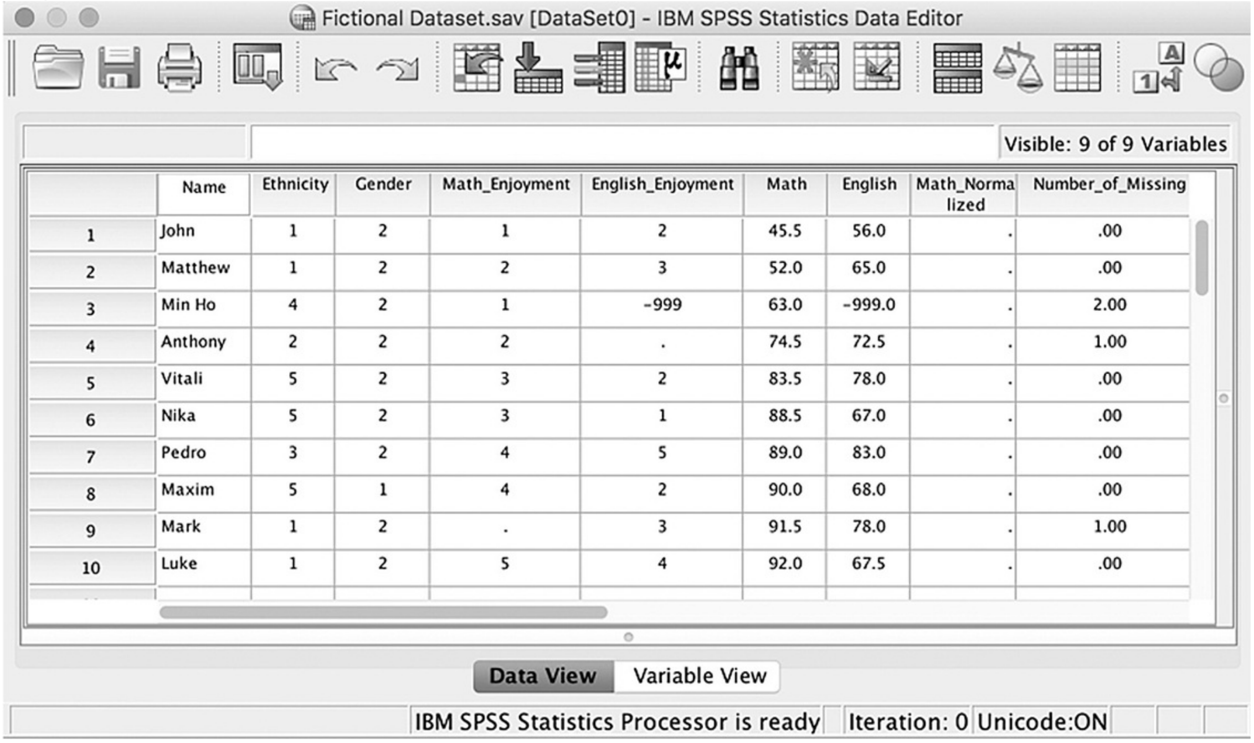
SPSS defines variables that represent text (rather than numbers) as the *String* type (exemplified by the *Name* variable in Figure 2). Such variables are automatically classified as *Nominal* under the *measure* column. More commonly, SPSS makes use of *Numeric* types of data. Numeric type variables can be classified by the following three measures (examples from the fictional data set given):

1. *Scale*: math and English percentile variable,
2. *Ordinal*: math and English enjoyment variable (anchored by five ordered categories), and
3. *Nominal*: ethnicity and gender variables.

It is important that the user carefully defines the appropriate *type* and *measure* to each variable prior to performing statistical analyses (certain procedures can only be performed with particular variable types).

Other aspects of each variable that are principally useful to define are the *Values* and *Missing* columns. As mentioned, the nominal variable, *Gender*, has two categories. The *Values* option allows the researcher to assign a value of 1 to males and 2 to females. Similarly, the *Ordinal* variable, math enjoyment, can be assigned a value of 1 to represent none, 2 to represent a little, and so forth. By clicking on the *Value Labels* icon (Figure 1), users are able to toggle between the numeric and raw value labels associated with the variables in the Data Editor window (Data View).

Figure 2 IBM SPSS Statistics Data View

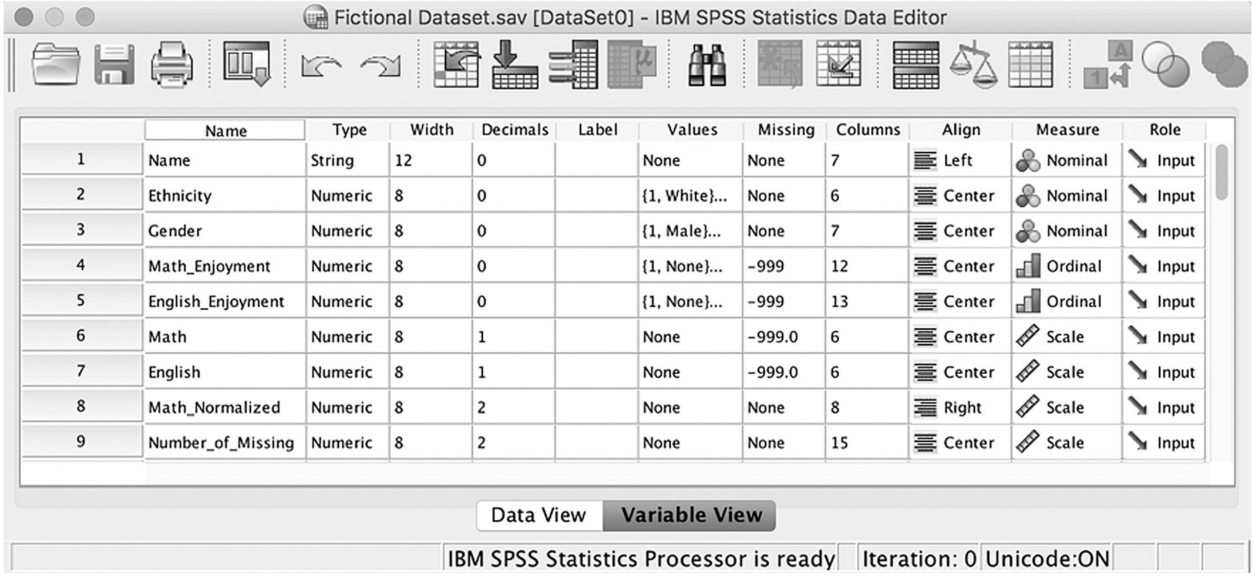


Fictional Dataset.sav [DataSet0] - IBM SPSS Statistics Data Editor
Visible: 9 of 9 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Iteration: 0
Unicode:ON

	Name	Ethnicity	Gender	Math_Enjoyment	English_Enjoyment	Math	English	Math_Normalized	Number_of_Missing
1	John	1	2	1	2	45.5	56.0	.	.00
2	Matthew	1	2	2	3	52.0	65.0	.	.00
3	Min Ho	4	2	1	-999	63.0	-999.0	.	2.00
4	Anthony	2	2	2	.	74.5	72.5	.	1.00
5	Vitali	5	2	3	2	83.5	78.0	.	.00
6	Nika	5	2	3	1	88.5	67.0	.	.00
7	Pedro	3	2	4	5	89.0	83.0	.	.00
8	Maxim	5	1	4	2	90.0	68.0	.	.00
9	Mark	1	2	.	3	91.5	78.0	.	1.00
10	Luke	1	2	5	4	92.0	67.5	.	.00

Source: IBM Corporation. (2012).

Figure 3 IBM SPSS Statistics Variable View



Fictional Dataset.sav [DataSet0] - IBM SPSS Statistics Data Editor
Data View
Variable View
IBM SPSS Statistics Processor is ready
Iteration: 0
Unicode:ON

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Name	String	12	0		None	None	7	Left	Nominal	Input
2	Ethnicity	Numeric	8	0		{1, White}...	None	6	Center	Nominal	Input
3	Gender	Numeric	8	0		{1, Male}...	None	7	Center	Nominal	Input
4	Math_Enjoyment	Numeric	8	0		{1, None}...	-999	12	Center	Ordinal	Input
5	English_Enjoyment	Numeric	8	0		{1, None}...	-999	13	Center	Ordinal	Input
6	Math	Numeric	8	1		None	-999.0	6	Center	Scale	Input
7	English	Numeric	8	1		None	-999.0	6	Center	Scale	Input
8	Math_Normalized	Numeric	8	2		None	None	8	Right	Scale	Input
9	Number_of_Missing	Numeric	8	2		None	None	15	Center	Scale	Input

Source: IBM Corporation. (2012).

IBM SPSS also accounts for missing variables. Blanks in the data are automatically presented as periods—these missing data are defined as *System Missing*. In the fictional data set, both Mark and Anthony did not respond to a question related to their academic enjoyment (Figure 1). If it is difficult to determine the reason why the blank exists, it is more common for the researcher to define such variables as System Missing. However, in some studies, researchers may wish to distinguish between such instances and instances in which a question was not applicable. In this case, the researcher typically defines these *User Missing* values as large negative numbers (e.g., −999) as distinct from the other values in the data set. Such definitions can be made via the prompts provided in the Missing column. This missing data condition is exemplified by the −999 in which the student Min Ho did not sit the English test because he is enrolled in an English as a Second or Other Language class (not mainstream English).

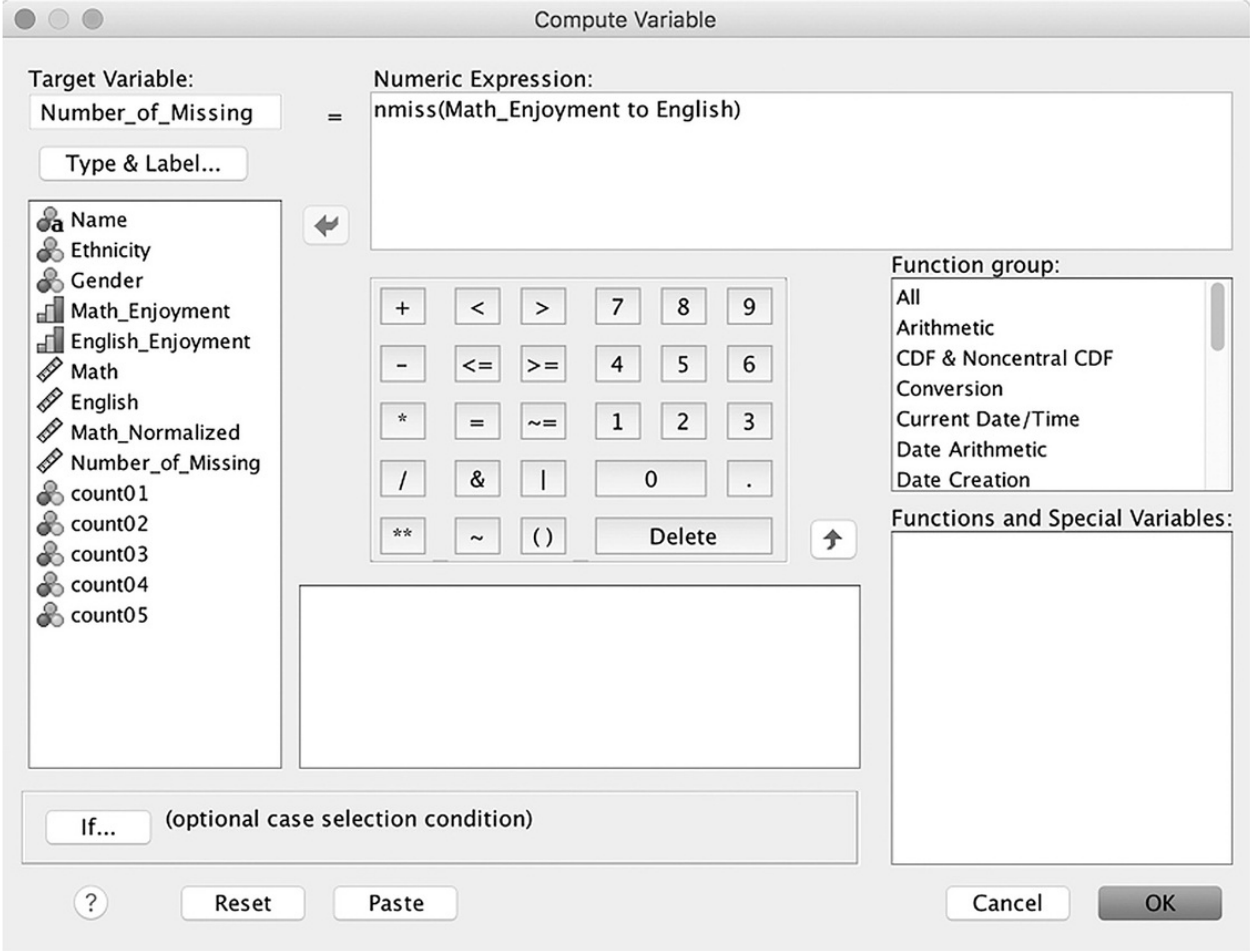
Utility for Data Preparation

SPSS Statistics is versatile in its capacity to assist users in data preparatory procedures. With an understanding of basic commands, users can ascertain important information about potential risks to the validity of their data set. For example, if a user wants to tally the number of missing blanks for each observation (participant), the user can make use of the *nmiss* function (Figure 4) provided by selecting the *Transform → Compute Variable* option.

In the example data set, the result of the nmiss procedure provides the user with a tally column, *Number_of_Missing* (see far right column in Figure 1). For the purpose of reducing risk associated with the data set, the researcher might remove cases that are missing more than a certain percentage of blanks.

As part of the data preparatory process, assessments of the degree to which participants gave the same response can also be carried out with ease. With large sample surveys, certain “anomalous” respondents may choose to give the same response to each question throughout the survey, perhaps in an attempt to finish as soon as possible. Counts of the frequency with which each participant gives the same response can be generated easily through SPSS’s syntax function (*File → Open → Syntax; Run → All*). The syntax used to generate this query is presented in Figure 5.

Figure 4 The SPSS nmiss procedure



Target Variable: Number_of_Missing

Numeric Expression: nmiss(Math_Enjoyment to English)

Source: IBM Corporation. (2012).

In Figure 1, the two far right columns present the counts of 1s (Count 1) and 2s (Count 2s) for each case, respectively (variable names obscured by speech bubbles).

This functionality would be especially useful in large data sets that involve multiple ordinal (Likert type) questions. In such cases, participants who give the same number of responses to a very high percentage of questions could be considered a threat to validity and removed from the data set.

Assessments of item-level normality (skewness and kurtosis) can be made quickly via the *Analyze* → *Descriptive Statistics* → *Descriptives* → *Options* function. Thereafter, if item normality was deemed necessary for analysis, exponential transformations of violating variables can easily be performed via the *Transform* → *Compute Variable* function (Figure 6).

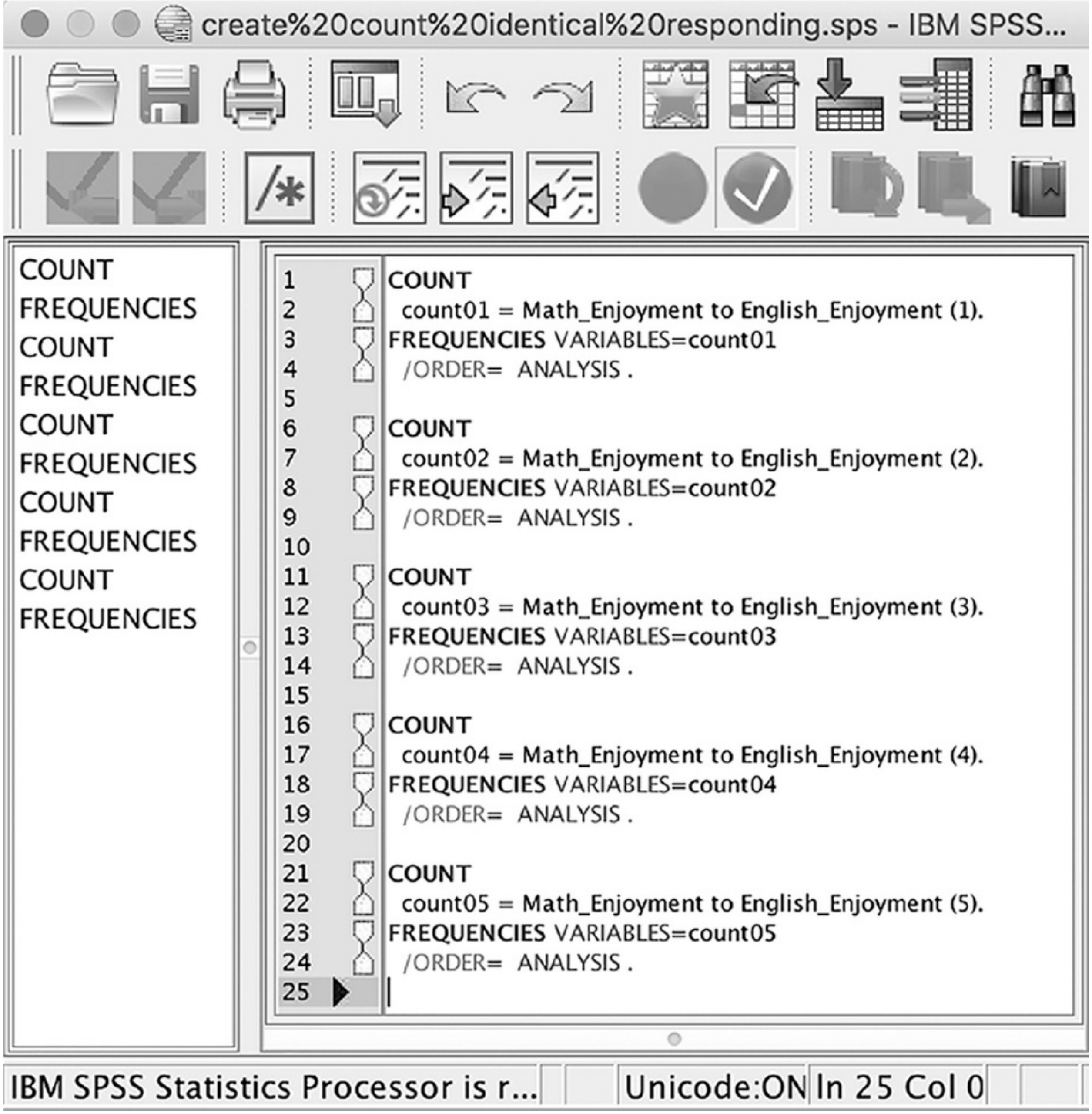
Note that the math variable is first anchored at 1 via (math-44.5) operation; thereafter, each value of that variable is squared (**2) resulting in the normalized variable, *Math_Normalized*; by clicking the *Paste* option at the bottom of the dialogue box, a Syntax window with command language is generated.

The SPSS Statistics program also employs a suite of utilities that can help to deal with missing data via its *Analyze* → *Missing Value Analysis* menu. In this case, the user is guided to perform appropriate tests, such as Roderick Little's missing completely at random test; pending results of such tests may use the *Impute* function to generate robust estimates of the missing values via the expectation–maximization algorithm.

Descriptive Analysis

IBM SPSS enables users to generate descriptive statistics with ease. If a researcher wants to ascertain the number of females and males in the sample, the researcher can make use of the *Analyze* → *Descriptives* → *Frequencies* function. Following the prompts, the user would be provided with an output window detailing results of the query (Figure 7).

Figure 5 Using SPSS syntax to generate counts of repeated responses

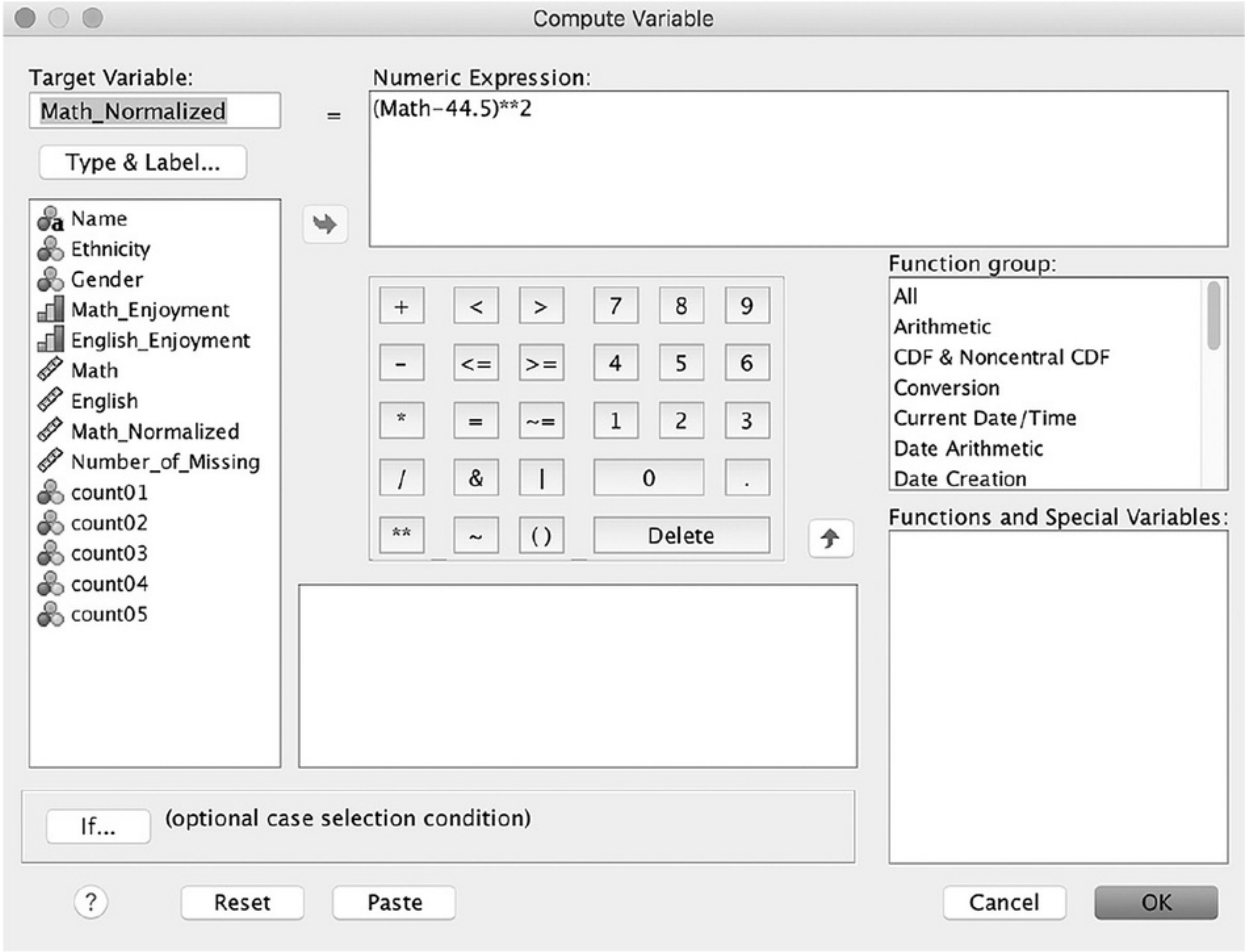


```
COUNT
FREQUENCIES
COUNT
FREQUENCIES
COUNT
FREQUENCIES
COUNT
FREQUENCIES
COUNT
FREQUENCIES
```

```
COUNT
  count01 = Math_Enjoyment to English_Enjoyment (1).
FREQUENCIES VARIABLES=count01
  /ORDER=  ANALYSIS .

COUNT
  count02 = Math_Enjoyment to English_Enjoyment (2).
FREQUENCIES VARIABLES=count02
  /ORDER=  ANALYSIS .

COUNT
  count03 = Math_Enjoyment to English_Enjoyment (3).
FREQUENCIES VARIABLES=count03
  /ORDER=  ANALYSIS .

COUNT
  count04 = Math_Enjoyment to English_Enjoyment (4).
FREQUENCIES VARIABLES=count04
  /ORDER=  ANALYSIS .

COUNT
  count05 = Math_Enjoyment to English_Enjoyment (5).
FREQUENCIES VARIABLES=count05
  /ORDER=  ANALYSIS .

```

Source: IBM Corporation. (2012).

Figure 6 Variable normalization of math via exponentiation of 2

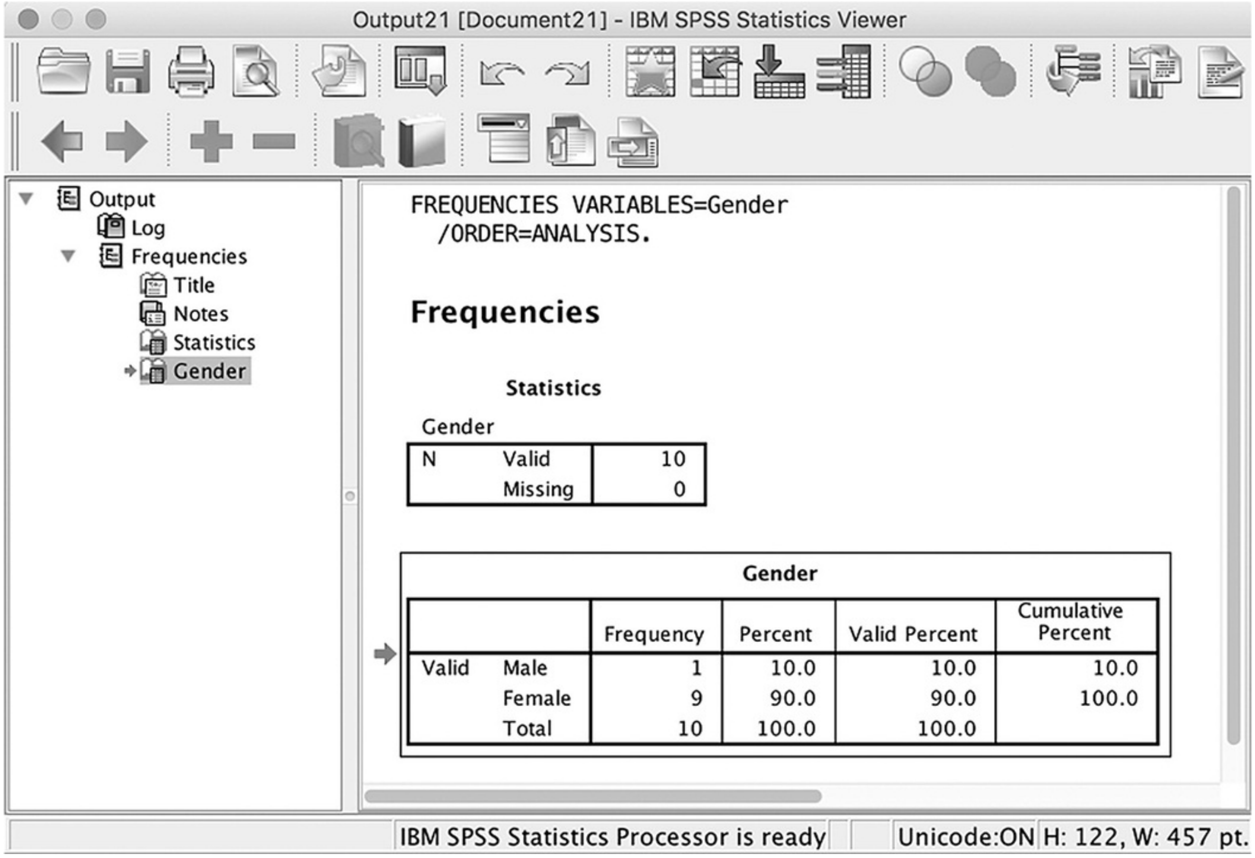


Compute Variable

Target Variable: Math_Normalized = Numeric Expression: (Math-44.5)**2

Source: IBM Corporation. (2012).

Figure 7 Output from SPSS frequency query


Output21 [Document21] - IBM SPSS Statistics Viewer
Output
Log
Frequencies
Title
Notes
Statistics
Gender
IBM SPSS Statistics Processor is ready
Unicode:ON
H: 122, W: 457 pt.

```
FREQUENCIES VARIABLES=Gender
  /ORDER=ANALYSIS.
```

Frequencies

Statistics

Gender

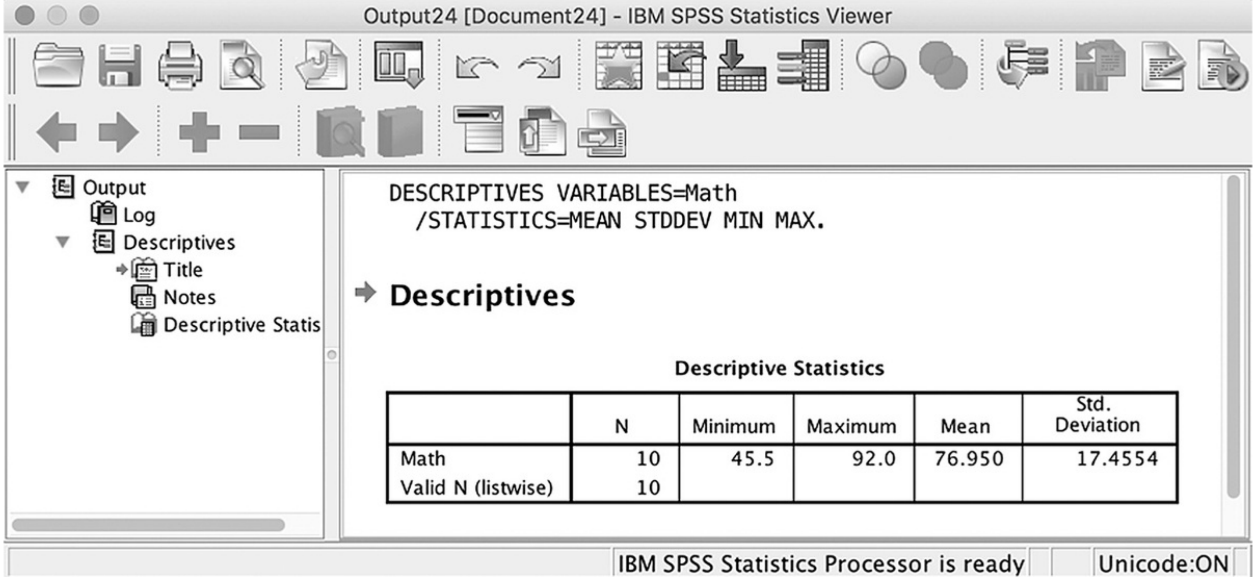
N	Valid	10
	Missing	0

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	1	10.0	10.0	10.0
	Female	9	90.0	90.0	100.0
	Total	10	100.0	100.0	

Source: IBM Corporation. (2012).

Figure 8 Output from SPSS descriptive query


Output24 [Document24] - IBM SPSS Statistics Viewer
Output
Log
Descriptives
Title
Notes
Descriptive Statis
IBM SPSS Statistics Processor is ready
Unicode:ON

```
DESCRIPTIVES VARIABLES=Math
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Math	10	45.5	92.0	76.950	17.4554
Valid N (listwise)	10				

Source: IBM Corporation. (2012).

In addition to frequency counts for each group, by default the associated percentages are also given. If a researcher is interested in measures of a variable's centrality and dispersion, the *Analyze → Descriptive Statistics → Descriptives* function can be used. Following the prompts (and manipulation of the Options presented in the Descriptives dialogue box), the researcher is able to generate the math score mean and standard deviation values presented in Figure 8.

Various Statistical Procedures

One of the strengths of IBM SPSS is its utility to clean and prepare data for subsequent analysis. Beyond this, it provides the researcher with a broad suite of statistical procedures that encompass many of the lines of analysis followed in quantitative education research, measurement, and evaluation. The following are some of the procedures available in SPSS: contingency tables, reliability tests, correlation coefficients, *t* tests, analysis of variance, multivariate analysis of variance, general linear modeling, regression, nonlinear regression, multiple linear regression, logistic regression, log-linear regression, cluster analysis, exploratory factor analysis, discriminant analysis, multidimensional scaling, survival analysis, probit analysis, forecasting/time series, nonparametric analysis, and neural network analysis.

Matthew Gordon Ray Courtney

See also R; SAS; Stata

Further Readings

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 39(1), 1–38. Retrieved from http://web.mit.edu/6.435/www/Dempster77.pdf

IBM Corporation. (2012). IBM SPSS Statistics 21. Armonk, NY: IBM.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202. Retrieved from http://www.jstor.org/stable/2290157

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research & Evaluation, 15(12). Retrieved from http://pareonline.net/pdf/v15n12.pdf

Rose, B. E. (2001, April 24). "Chicago-based software firm cuts 90 jobs. Chicago Tribune.

Wessa, P. (2013). Box-Cox normality plot—Free statistics software. Office for research development and education, version 1.1.23-r7. Retrieved from http://www.wessa.net/rwasp_boxcoxnorm.wasp

Michael Quinn Patton Michael Quinn Patton Patton, Michael Quinn

Stakeholders

Stakeholders

1583

1585

Stakeholders

Stakeholders are individuals, groups, or organizations that can benefit from an evaluation or those who can affect or may be affected by an evaluation process or its findings. The word *stakeholder* originated in gambling in 16th-century England, where wagers were posted on wooden stakes. Later the term was broadened to refer to a neutral or trustworthy person who held a wager until the winner was decided. The term *stakeholders* was brought into evaluation from management consulting where it was adopted in 1963 at the Stanford Research Institute as a way of describing people who were not directly stockholders in a company but whose support was critical to the company’s success, for example, critically skilled employees and senior leadership.

Increasingly, stakeholders are seen as important in evaluation practice for both practical and ethical reasons. The emphasis in evaluation on identifying and engaging key stakeholders is based on the principle of intentionality, namely, that evaluation credibility, relevance, and use is enhanced by focusing on the intended uses of the evaluation by the primary intended users. This means that the evaluator does not alone determine the priority evaluation questions and methods but works with key stakeholders throughout the evaluation process.

Research has demonstrated that attention to and involvement of key stakeholders strengthens the design and implementation of evaluations and makes evaluation results more useful. This entry discusses the different types of stakeholders, stakeholder identification and analysis, and dealing with variations in stakeholder power.

Diversity of Stakeholders

Diversity of Stakeholders

Stakeholders in a program can be distinguished in four general groups:

1. Those in positions of authority who make major decisions about the program's funding and strategy;
2. Staff who have direct responsibility for the program, plus program developers, administrators in the organization implementing the program, and program managers;
3. The intended beneficiaries of the program, their families, and their communities; and
4. Others with a direct, or even indirect, interest in program effectiveness, including journalists and members of the general public, or, more specifically, taxpayers, in the case of public programs.

Essentially, then, stakeholders include anyone who makes decisions about or has an interest in the effectiveness of a program. Determining the priority stakeholders for any given evaluation varies by program area, evaluation purpose, and potential stakeholder interest and political influence.

Variations by Program Area

In the education arena, stakeholders can include teachers, parents, students, administrators, elected school officials, government educational policy makers and bureaucrats, philanthropic funders of educational programs, journalists who cover education issues, advocates for educational reform, curriculum developers, educational academics and scholars, taxpayers who may vote on educational referenda and school bonding proposals, and the general public. In the health arena, stakeholders can include patients, doctors, nurses, health system administrators, health insurers and insurance agents, health system policy makers and bureaucrats, philanthropic funders of health programs, journalists who cover health issues, advocates for health reform, medical device makers and pharmacists, taxpayers affected by the costs of public health, health academics and scholars, and the general public.

Similar distinct lists of stakeholders can be created for initiatives in criminal justice, international development, environmental sustainability, antipoverty programs, humanitarian assistance, and so forth. Thus, identifying stakeholders involves knowing who the important actors, participants, implementers, and intended beneficiaries are in any specialized arena of programming, intervention, and change.

Variations by Type of Evaluations

Different evaluation purposes serve the interests of different stakeholders. Variations in evaluation purpose imply asking and answering different evaluation questions and using targeted strategies to promote use among priority stakeholders. Most evaluation efforts include some early consultation with key stakeholders to help ensure an appropriate evaluation design and effective use of findings.

> Formative evaluations aimed at program improvement typically target program staff as primary stakeholders.
>
> Summative evaluations that render judgments of overall merit, worth, and significance of a program are aimed at major decision makers, policy makers, and those who fund programs.
>
> Accountability evaluations are aimed at those who regulate programs and specify what they are supposed to do and accomplish.
>
> Developmental evaluations support innovation and adaptation in complex dynamic environments, where the stakeholders are social innovators, change agents, people working toward major systems change, and social entrepreneurs.
>
> Knowledge-generating and theory-driven evaluations target scholars and academics who study the effectiveness of change processes as well as program designers who work from a conceptualization of a theory of change.

Variations by Potential Stakeholder Interest and Influence

Stakeholders can be subdivided into subgroups based on interest and influence. The most general category is the broad audience for an evaluation’s findings, but audiences are generally seen as anonymous and passive. To target an evaluation at the information needs of a specific person or a group of identifiable and interacting persons, called key stakeholders, is quite different from identifying the audience for an evaluation. *Key stakeholders* are a subset of people who have special connections to and influence over a program, although who is key will always be a judgment call and a matter for negotiation. *Primary intended users* are a subset of key stakeholders. They are those *specific* stakeholders selected to work with the evaluator throughout the evaluation to focus the evaluation, participate in making design and methods decisions, and interpret the results to assure that the evaluation is useful, meaningful, relevant, and credible. Primary intended users represent key and diverse stakeholder constituencies and have responsibility for transmitting evaluation findings to those constituencies for use.

Differentiating stakeholders is based on the understanding that the stakeholders of any particular evaluation will have diverse and often competing interests. No evaluation can answer all potential questions equally well. This means that some process is necessary for narrowing the range of possible questions to focus the evaluation. This means deciding whether to focus on and engage with just the subset of key stakeholders or to narrow further and concentrate on the subset of key stakeholders who are the primary intended users of the evaluation.

Stakeholder Identification and Analysis

Stakeholder identification and analysis involves mapping stakeholders and their interests in the program and the evaluation. Doing so will also surface or highlight some key evaluation issues and begin the process of identifying coalitions of either support for or opposition to the evaluation’s results. Stakeholder analysis should precede stakeholder engagement. This means that at least some stakeholders need to be engaged from the start to give the evaluator the information needed to fully understand stakeholders’ interests, expectations, powers, interrelationships, and the various roles they might need to play for a well-designed evaluation to serve its intended purpose for its intended users.

Determining who should be involved, how, and when in doing stakeholder analyses are important decisions. Key participants include those who have information that cannot be obtained through other means and those whose participation is needed to assure a successful evaluation. There is no rule governing when and how much stakeholders should participate in the stakeholder analysis; this involves trade-offs in broad representation of stakeholders; analysis quality, credibility, and legitimacy; and the ability to act based on the stakeholder analysis.

Dealing With Variations in Stakeholder Power

Power and status differences among stakeholders often come into play in evaluations, especially in large, complex evaluations with multiple stakeholder constituencies. In such cases, it can be useful to place stakeholders into four categories:

1. Those with substantial power over decision making about the program and high interest in the evaluation, sometimes called *power actors*;
2. Those with little power or interest, sometimes called *the disengaged;*
3. Those with power but little direct interest in the evaluation because they are preoccupied with other things but who will become engaged if something important emerges, sometimes called *the watchers*; and
4. Those with great interest in the evaluation but little power, sometimes called *the followers* because they closely follow what happens and the findings that emerge.

This analysis can help evaluators determine how to engage various stakeholder subgroups to produce a useful evaluation. For example, *power actors* are by definition key stakeholders. The interests of *watchers* should be assessed and anticipated. The *followers* can potentially become engaged and involved because of their high interest. In contrast, special efforts will be needed to reach *the disengaged*.

Michael Quinn Patton

See also Empowerment Evaluation; Evaluation, History of; Evaluation Versus Research; Program Evaluation; Utilization-Focused Evaluation

Further Readings

Bryson, J. (2000). What to do when stakeholders matter: A guide to stakeholder identification and analysis techniques. Public Management Review, 6, 21–53.

Bryson, J. M., & Patton, M. Q. (2015). Analyzing and engaging stakeholders. In

K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), Handbook of practical program evaluation (4th ed., pp. 36–61). Hoboken, NJ: Wiley.

Bryson, J. M., Patton, M. Q., & Bowman, R. A. (2011). Working with evaluation stakeholders: A rationale, step-wise approach and toolkit. Evaluation and Program Planning, 34, 1–12.

Cullen, A. E., Coryn, C. L., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluation. American Journal of Evaluation, 32, 345–361.

Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. Academy of Management Review, 22, 853–886.

Patton, M. Q. (2008). Utilization-focused evaluation (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2012). Essentials of utilization-focused evaluation. Thousand Oaks, CA: Sage.

Catherine O. Fritz Catherine O. Fritz Fritz, Catherine O.

Peter E. Morris Peter E. Morris Morris, Peter E.

Standard Deviation

Standard deviation

1586

1587

Standard deviation

A data set or sample is reasonably described by where the data points are centered (central tendency), how much spread or dispersion there is among the data points, and its frequency distribution (i.e., the shape of its histogram). This information allows interpretations and further calculations to be made from the data. Where the data set approximates a normal (bell-shaped) distribution, the mean is the best measure of central tendency, although medians are better central tendency measures for other distributions. If the distribution is approximately normal, then the standard deviation (*SD*) indicates the dispersion of the data.

It might be expected that the estimate of dispersion would be based on the average of the deviations of each data point from the mean, ignoring whether the deviations were positive or negative. However, some valuable statistical procedures (e.g., multiple linear regression, analysis of variance) rely on the square of the deviations rather than the absolute deviations. Therefore, the most commonly reported measures of dispersion—the variance and the *SD*—are also based on the square of the deviations.

The variance is calculated by first finding the deviation of each score (X) from the mean (M), $[X - M]$, squaring each deviation $(X - M)^2$, and then adding these squared deviations together to obtain the sum of squares (*SS*):

$$SS = X - M^2.$$

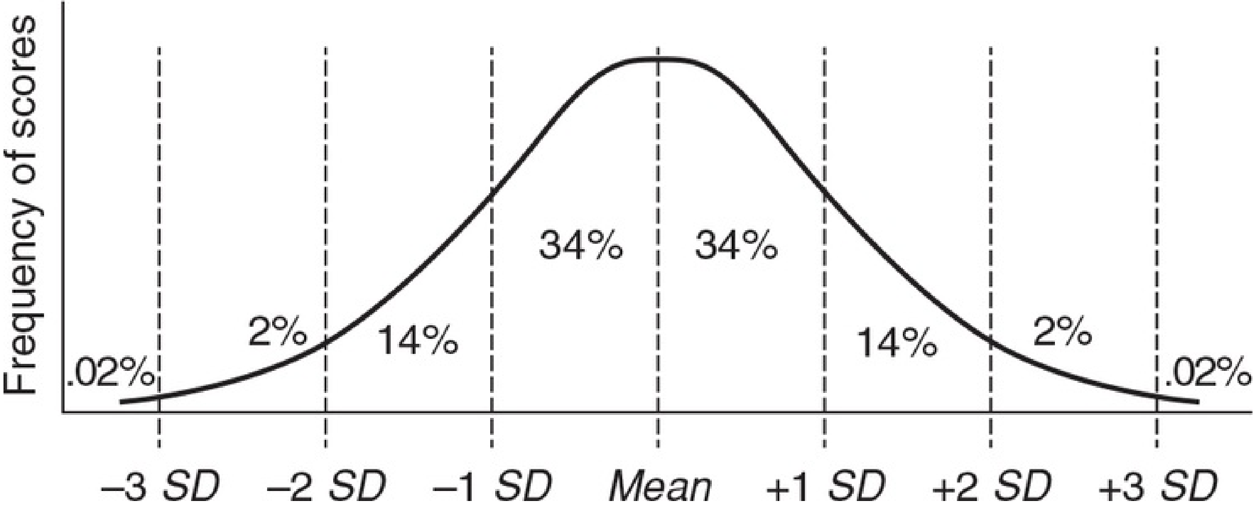
The variance of the sample is the average of this *SS*, obtained by dividing the *SS* value by the number of scores (*N*) in the sample. Thus, the variance is given by *SS* ÷ *N*. This measure of variance is very useful in many statistical calculations, but, because of the squaring of the deviations, it is out of scale with the original data. The problem of scale is addressed by taking the square root of the variance to give the *SD*, so compensating for the squaring of the deviations in the calculation of the variance. Thus, the *SD* is *SS* ÷ *N*.

When working with a sample of data, these formulae tend to slightly underestimate the population *SD* (or variance). The correction for this underestimate is to divide by *N* − 1, rather than *N*, yielding slightly higher values. Therefore, the formulae that are usually used to calculate these statistics are:

$$\text{Standard Deviation} = SS \div N - 1 \text{and Variance} = SS \div N - 1.$$

The *SD* should always be reported for reasonably normally distributed data because it provides a good idea of the data’s variability. Figure 1 illustrates the percentage of scores expected to occur within each *SD*. For example, 68% of the scores fall within one *SD* of the mean, 96% within two *SD*s, and 99.96% within three *SD*s. Data points more than three *SD*s from the mean are highly unlikely if the data are normally distributed, which is why these data points are often scrutinized and removed as outliers in a sample.

Figure 1 A normal frequency distribution with standard deviations (*SD*s) noted



The *SD* also provides the means of calculating further useful statistics, including standardized (*z*) scores, standardized effect sizes such as Cohen’s *d*, and, in combination with the sample size, the standard error of the mean and confidence intervals.

Catherine O. Fritz and Peter E. Morris

See also Descriptive Statistics; Distributions; Effect Size; Histograms; Normal Distribution; Variance; Z Scores

Further Readings

Aron, A., Coups, E., & Aron, E. N. (2014). Statistics for the behavioral and social sciences: A brief course (5th ed.). Upper Saddle River, NJ: Pearson.

Howell, D. C. (2010). Fundamental statistics for the behavioral sciences (7th ed.). Wadsworth.

Kranzler, J. H. (2010). Statistics for the terrified (5th ed.). Upper Saddle River, NJ: Pearson.

Dorothy J. Musselwhite Dorothy J. Musselwhite Musselwhite, Dorothy J.

Brian C. Wesolowski Brian C. Wesolowski Wesolowski, Brian C.

Standard Error of Measurement Standard error of measurement

1587

1590

Standard Error of Measurement

The term *standard error of measurement* indicates the spread of measurement errors when estimating an examinee’s true score from the observed score. Standard error of measurement is most frequently useful in test reliability. An observed score is an examinee’s obtained score, or raw score, on a particular test. A true score would be determined if this particular test was then given to a group of examinees 1,000 times, under identical conditions. The average of those observed scores would yield the best estimate of the examinees’ true abilities. Standard deviation is applied to the average of those scores across persons and administrations to determine the standard error of measurement. Observed score and true score can be used together to determine the amount of error:

$$\text{Score}_{\text{true}} = \text{Score}_{\text{observed}} + \text{Score}_{\text{error}}.$$

However, this true score is purely hypothetical and is not a practical way to estimate error. Therefore, other estimates of error must be used, including standard deviation and reliability.

Standard error of measurement applies to a single score and should be applied more frequently than a reliability coefficient to interpret individual score meaning. Standard error is used in conjunction with the normal distribution in order to make decisions about individual test scores. Accordingly, standard error can be used to estimate a range of scores around a specified cut point when determining an examinee’s ability or potential. The normal distribution can aid in the interpretation of scores that fall above, below, or between specific points on the distribution. This concept is particularly important, as it relates to standardized testing and promotion or retention criteria. For example, if the cut point for failing is a 50, and administrators want to be 68% sure of their decision, standard error of measurement indicates that examinees who are within one standard error (*SE*) of the cut point (i.e., 50 ± $SE_{\text{measurement}}$) may fluctuate above or below the cut point if the test were administered again. In situations such as this, it is imperative that more data be gathered, such as class performance indicators or growth scores, to determine promotion or retention.

A large standard error indicates a large amount of variability between different samples; therefore, the sample may not accurately represent the population. This occurs when sample means are spread far along the *y*-axis, in the tails of the normal distribution. When sample means are grouped closer to the population mean, standard error will be smaller.

This entry first discusses the distinction between standard error and standard deviation, as these concepts are often confused. Then, standard error is applied to confidence intervals and other assessment situations.

Standard Error Versus Standard Deviation

Standard deviation indicates how well the mean represents sample data. When considering a population, however, the mean of one sample does not necessarily represent the mean of every possible sample. If several samples were taken from one population, each sample mean may differ. Sampling variation is crucial to understanding the connection from standard deviation to standard error.

Standard deviation is a measure of spread, specifically as scores are situated around the mean. More specifically, standard deviation considers scores between examinees. Standard deviation is more closely related to range but is not as affected by outlying scores. A high standard deviation is an indication that scores have more variation or are widely distributed around the mean. A low standard deviation indicates the scores have less variation and are not widely distributed around the mean.

Standard deviation is most useful when determining where examinees are expected to fall within a range of scores. For example, as standard deviation aligns with the concept of normal distribution, it can be assumed that roughly 68% of examinees will attain scores within the range of one standard deviation above and below the mean score. Standard deviation can be calculated using just one test administration.

Standard error is, like standard deviation, a measure of spread. Standard error determines an individual examinee’s spread had that student been tested repeatedly. Unlike standard deviation, standard error must be calculated using a much larger data set. Using one sample numerous times would elicit very similar means among the test administrations. However, numerous random samples would yield many different means. These sample means would form their own normal distribution where the means would ultimately have a single mean. The “mean of means” would be the best approximation of the population mean. The standard deviation of the distribution of these means is called the standard error of the mean, which refers to the fluctuations, or errors, that occur when estimating the population mean from sample means. The distribution created by the statistics from these multiple samples is called the sampling distribution. Because it is not feasible to take 1,000 random samples, a formula is used to estimate standard error of the mean using one sample:

$$SE_{\text{mean}} = s \,/\, \sqrt{N},$$

where SE_{mean} refers to the standard error of the mean, s refers to the standard deviation of the mean, and N refers to the sample size. In terms of sample size, there exists an inverse relationship. Larger sample sizes will yield a smaller standard error, and smaller sample sizes will yield a larger standard error.

Standard error of measurement is different from standard error of the mean. Standard error of measurement focuses more on the spread of errors, as they relate to a true score compared to an observed score. Standard error of the mean focuses on error in relation to the estimation of population mean. In total, standard error investigates how varying the sample statistic is when numerous samples are extracted from the same population.

The standard error of a sampling distribution is calculated using multiple samples from the population in question. However, multiple samples are not always available, especially when administering a single test administration. Obtaining an infinite number of test administrations would be difficult due to money and time constraints but would also yield unfavorable effects, such as testing fatigue. Therefore, the researcher must assume that each individual test score is the best estimate of that examinee’s true score. Similar to the estimation of population mean, sampling errors in the estimation of true scores additionally occur. Similarly, these sampling errors will also be normally distributed, with a occur. Similarly, these sampling errors will also be normally distributed, with a standard deviation called the standard error of measurement. The estimate for the standard error of measurement is calculated using the following formula:

$$SE_{\text{measurement}} = s\sqrt{1 - r_{xx}},$$

where $SE_{\text{measurement}}$ refers to the standard error of measurement, *s* refers to the standard deviation of the measure or test, and r_{xx} refers to the reliability of the measure (e.g., Cronbach’s α).

The reliability coefficient represents the test’s consistency. The aforementioned formula shows that the standard error of measurement increases as the standard deviation increases. In addition, the standard error of measurement increases as the test reliability decreases, showing an inverse relationship. If a test is perfectly reliable ($r = 1.0$), an examinee will attain the same score for every test administration. If the reliability is close to perfect, the standard error will be small, indicating the examinee’s observed score is very similar to the true score.

Confidence Intervals

Standard error of measurement can be most beneficial in the construction of confidence intervals. Standard error and standard deviation are similar in that they both explore estimates of true scores. Under the assumptions of the normal distribution, 68% of the time, examinees’ true scores lie within one standard error of measurement above or below the mean (±1). Next, 96% of the time, examinees’ true scores would lie within two standard errors of measurement above or below the mean (±2). Last, 99.7% of the time, examinees’ true scores would lie within three standard errors of measurement above or below the mean (±3).

The standard error is combined with the examinee’s observed score, just as standard deviation is, to determine the upper and lower bound of the confidence interval for that specific examinee. Standard error can only apply to an individual score when developing a confidence interval. Standard error is often associated with probability and the prediction of true scores, which should be applied to the distribution of scores as a whole.

An example of standard error of measurement with confidence intervals can be illustrated through intelligence testing, such as the IQ test. On one test administration, a student may earn a score of 108. Other students in the same testing administration may have received similar scores, and a standard error of 5 was calculated. Using each student’s score and the associated group standard error, an interval can be calculated to determine where each student is likely to score if they were given the test again. The standard error is added and subtracted to the original score (108 + 5 = 113, 108 − 5 = 103). This student is 68% likely to score between 103 and 113 if given the IQ test again. The interval can be increased by adding and subtracting the standard error again, thereby becoming more likely to predict the student’s score on the next administration. As the confidence level increases, precision will decrease, but Type I error rate will decrease.

Standard Error of Estimate

Using a regression analysis, the standard error of estimate approximates how spread the prediction errors are when using *X* values to predict *Y* values. These errors occur because of unreliable measurement in one of the variables or because of unsystematic differences between the values. The regression analysis provides a best estimate as to an examinee’s predicted score, but similar to other standard error measures, there will be sampling errors around the estimate. The standard error of estimate should not be used as an estimator of true scores when comparing to observed scores.

Standard error of measurement is used to express test reliability. Standard error of estimate is used to express test validity. A small standard error of estimate indicates a more valid test.

Assessment and Measurement

Assessments are more frequently being used as a method of describing individuals, even as an indicator of the examinee’s fate. The raw score gives little information as to an examinee’s ability and characterization. The interpretation of these scores is becoming more essential in determining students’ actual abilities. Standard error of measurement serves as an indicator of reliability that is independent of the variability in sample groups. Using the idea of confidence intervals, a student’s ability can be perceived as a range of scores rather than one specific score. This concept of bands of scores can apply not only to one individual on different tests but also to multiple students on the same test.

test.

Dorothy J. Musselwhite and Brian C. Wesolowski

See also Normal Distribution; Reliability; Standard Deviation; Standard Error of Measurement; Validity

Further Readings

Boyle, J. D., & Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York, NY: Schirmer Books.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. London, UK: Sage.

Gravetter, F. J., & Wallnau, L. B. (2011). Essentials of statistics for the behavioral sciences. Belmont, CA: Wadsworth Publishing.

Kubiszyn, T., & Borich, G. (2003). Educational testing and measurement: Classroom application and practice. New York, NY: Wiley.

Payne, D. A. (2003). Applied educational assessment. Toronto, Canada: Wadsworth Publishing.

Reid, H. M. (2014). Introduction to statistics: Fundamental concepts and procedures of data analysis. Thousand Oaks, CA: Sage.

Boaz Shulruf Boaz Shulruf Shulruf, Boaz

Standard Setting

Standard setting

1590

1595

Standard Setting

It is widely accepted that standard setting methods aim to distinguish between competent and incompetent examinees who sit for a test or an examination (these terms are used interchangeably). Hans Pant and colleagues describe standard setting as an umbrella term that incorporates consensual approaches of panels of experts to set discrete cut scores on continuous test performance scales.

This description regards standard setting as a *consensual decision-making process*, which translates the cumulative understanding of decision makers of what constitutes competence and incompetence to establish a discrete cut score on a continuous scale, aiming to separate the competent from the incompetent. However, others regard standard setting as a decision-making process aiming to establish a cut score that is the boundary between acceptable and nonacceptable performance; that is, they focus on acceptability rather than competence.

The standard setting process is required when the observed test score cannot identify whether the examinee has clearly passed or clearly failed the test (or between any other consecutive grades). Scores within that range can be regarded as borderline. For example, there is a need to determine the test score in the final examination of medical knowledge that would ensure the examinee has acquired enough knowledge to practice medicine. Similar decisions need to be made about drivers, pilots, and other students to determine whether they are eligible to progress in their program.

Although in practice the determination of cut scores is normally made by experts or other representatives of the professional communities, these panels are not the decision makers but rather standard recommenders. The standards are actually set by the authorized bodies (e.g., professional associations, academies, boards of education, state agencies) that consider the recommendations and make operative decisions. Setting standards in a systematic way is important, as it enhances the confidence of the public and other stakeholders that the decision is robust, fair, reliable, valid, and unbiased.

There is a plethora of standard setting methods. Many, but not all, methods are based on making a consensual decision. There are a few ways to classify standard setting methods; however, standard setting methods are typically based on either a panel’s judgment or on statistical techniques using test scores generated by examinees. Some of the most advanced methods utilize both. Methods can also be classified by the frame of reference used for the decision: norm-referenced or criterion-referenced methods. The norm-referenced method uses the examinees’ test scores as a reference for the decision making (normally determined proportion of passes and fails), whereas the criterion-referenced method utilizes an external reference and is independent of the examinee’s overall performance and does not generate a fixed proportion of passes or fails. Others classify standard setting methods by two categories: examinee-centered and test-centered methods. Test-centered methods determine cut scores based on the content of the test, whereas examinee-centered methods determine the cut score by perceived attributes of the examinees.

The categories of competence and incompetence may relate to any two subsequent categories (e.g., fail vs. pass; pass vs. distinction). Standard setting is based on the assumption that the test scores provide sufficient information that allows reasonable estimation whether an examinee is competent or not, in other words whether the examinee met the performance criteria to be classified within the higher category among the two consecutive categories considered.

Estimating the quality of standard setting is a major challenge. The common practice is applying measures of reliability. The reliability may be measured by agreement across judges (panelists who are involved in the decision making), by resampling or by estimating measurement errors, particularly but not limited to methods that are based on statistical procedures.

The following provides a brief description of selected standard setting methods, which are either commonly used or introduce original and interesting approaches. The strengths and weaknesses of each method are also presented. The entry concludes with some additional considerations associated with standard setting.

The Nedelsky Method

The Nedelsky method is applicable for multiple-choice questions. Each judge in a panel estimates the number of options a hypothetically borderline examinee would be able to rule out. Then that item receives a score, which is 1 divided by the remaining options. For example, if an item has four options, the judge estimates that a borderline would rule out one option; the Nedelsky value for the item is 1/(4 − 1) = 0.33. Then the mean of the Nedelsky values for each item across all judges is calculated, and the sum of the means (rounded up) is the minimum number of the correctly answered items for “pass” (i.e., the cut score).

Strengths of this method include the following: Judgment is based on cumulative judges’ perceptions of borderline performance; the cut score is independent to examinees’ performance, as the cut score could be determined prior to examinees sitting for the test; and the cut score can be presented to the examinees prior to the test. Weaknesses are that this method is applicable to multiple-choice questions only, and reliability is achieved only by a large number of judges.

The Ebel Method

In this method, judges are asked to make two decisions for each item; first, the difficulty of the item (easy, medium, and hard) and second, the relevance of the item (essential, important, acceptable, and questionable). Then each judge places each item in a matrix comprising the 12 cross categories (3 × 4). In the next stage, the judges make another decision: Estimating the number of items in each cell that the hypothetical minimally qualified or borderline examinee is expected to answer correctly. The sum of the number of items expected to be correctly answered by borderline examinees, across all cells and judges, is divided by the number of items multiplied by number of judges; this is the passing percentage or the cut score.

Strengths of this method are that it is a test-centered method that is not affected by a particular population; it considers relevance and is not limited to difficulty; it can be applied to different types of tests beyond multiple-choice questionnaires; and the cut score can be presented to the examinees prior to the test. It also has several weaknesses: Its lengthy process requires two stages of decision making by each judge. Reliability is achieved only by a large number of judges. Last, the use of a "questionable" category is problematic, as it may imply the item is not relevant and hence should be omitted.

The Angoff Method

Although introduced by William Angoff in 1971, this method was actually attributed by Angoff to Ledyard Tucker who was Angoff's colleague at the Educational Testing Service in Princeton. The Angoff method is widely used and a number of variations are applied. The fundamental method is based on judges' estimates of the probability that the minimally competent borderline examinees would give a correct answer to each of the items. In this process, probability is interchangeable with proportion and expressed as percentages. The mean percentage for all items is calculated for each judge and then the grand mean of the judges' means is calculated to yield the cut score as a percentage of items correctly answered for a particular test. The common modified Angoff methods include providing judges with additional information such as psychometric data from the test (e.g., item difficulty) and/or presenting the decisions made by other judges within the panel. Then a second round of Angoff method is undertaken, and each judge may change his or her decisions compared to the first round. The second round provides the final cut score.

Some of the strengths of this method are that it considers each item difficulty, and the modified Angoff method allows judges to correct their judgment based on additional information. In addition, the cut score can be presented to the examinees prior to the test. This method also has several weaknesses: The reliability of the Angoff method is affected by the number of items and number of judges (it has been suggested that at least 10 judges are required to reach acceptable reliability); the method is affected by judges' attributes such as leniency/stringency; and the concept of the probability of a hypothetical minimally competent borderline examinee giving a correct answer is complex and might be understood differently across judges.

Direct Consensus Method

This method is similar to the Angoff method but instead of judges' estimating probability of a correct answer to individual items, the test is divided into sections and the judges indicate the number of items in each section that the sections and the judges indicate the number of items in each section that the minimally competent examinee is expected to answer correctly. The rest of the calculation is similar to the modified Angoff method, whereby judges are informed on their peers' scores and then they are requested to reconsider their first judgment.

The direct consensus method is simpler than the Angoff method, particularly when the test includes a large number of items, which is considered a strength. Other strengths are that it provides information by sections, which may be useful in noncompensative types of examination, and the cut score can be presented to the examinees prior to the test. Its weaknesses are similar to those of the Angoff method: The impact of the judge's subjectivity requires a large number of judges to reach acceptable reliability.

Contrasting Groups Method and Borderline Group Method

The contrasting group method is based on a panel of judges who are familiar with the examinees' level of performance. Thus, it requires an additional form of assessment prior to the test. The judges use their prior knowledge of examinees' competency to classify the entire group of examinees into two subgroups: master and nonmaster. This is done without the judges being aware of the examinees' current test results. The distributions of the two subgroups are then plotted on the same chart, and the intersection between the distribution lines is the cut score.

The borderline group method is similar to the contrasting groups method but consists of two modifications. In the borderline group method, the judges classify the examinees into three categories: nonmaster, borderline, and master. Then only the borderline group is considered and a central measure, most commonly the median of the borderline test scores, is the determined cut score.

Both methods are simple and are examinee centered and both rely on familiarity of the judges with examinees' competence—all of which are considered strengths. As for weaknesses, in the contrasting group method, there is an inherent logical flaw. In these methods, judges may classify some examinees as master and then find them to be incompetent (i.e., their grade was classified below the cut score), or vice versa. This means that either the judges are wrong or the test does not provide accurate information regarding the examinees' proficiency. Another weakness is that the cut score cannot be presented to the examinees prior to the test.

The Bookmark Method

The bookmark method is now considered one of the most reliable standard setting methods. The first step in the bookmark method is to provide judges with a list of the items ordered by the probability of a correct response for each item based on an item response theory (IRT) model. Then the judges, working in small groups, place a “bookmark” at the point that discriminates two proficiency categories (e.g., “fail” and “pass”; “pass” and “distinction”).

The bookmark between fail and pass is the point at which the judges estimate that the minimally competent examinee has at least a 67% probability of giving a correct answer to that item. The point on which the group agrees as the boundary between fail and pass (i.e., probability response of 67%) is the test cut score. The bookmark method may take more than one round, with each round providing more information to judges to improve their decisions.

This method is statistically robust and enables judges making their decisions based on well-established measurement theories. For the judges, placing the bookmark is a relatively easy and intuitive process. However, this method is conceptually complex, particularly the use of IRT data. The reliance of IRT models means that data need to be suitable for IRT models (i.e., unidimensional and with local independence). Some examinations may not meet these requirements, which then prevents the implementation of this method. Also, the cut score cannot be presented to the examinees prior to the test.

The Hofstee Method

The Hofstee method is a compromised method that focuses on the practicality of standard setting. The method requires judges to provide four estimates: (1) the highest percentage correct cut score that would be acceptable, (2) the lowest percent correct cut score that would be acceptable, (3) the maximum acceptable failure rate, and (4) the minimum acceptable failure rate. Then, on a chart of the cumulative distribution (horizontal = percentage correct required; vertical = percentage failing) of the actual performance of a group of examinees, a line connecting points (1,4) and (2,3) is marked. The intersection between that line with the cumulative distribution curve is the cut score.

The Hofstee method addresses practical issues related to admissions or decision making when number of places is limited. It also considers all three main considerations: test difficulty, examinee population, and objective of the test. However, it is a compromised method; hence, none of the objectives is fully met. Moreover, the cut score cannot be presented to the examinees prior to the test.

The Beuk Method

This method has some similarities with the Hofstee method. The judges are asked to (a) estimate the minimum level of knowledge required to pass an examination and (b) estimate the expected passing rate for the same examination. Then, on a similar chart, the cumulative distribution (horizontal = percent correct required; vertical = percentage failing) of the actual performance of a group of examinees, a point of the Mean (1) and Mean (2) are marked. At the next stage, standard deviations of judges' estimates for both tasks are calculated and the ratio *SD*(2)/*SD*(1) is used (as the slope) to draw a line from point (2,1) down and backward to the cumulating distribution curve. The intersection of the curve and the line determines the cut score.

This method addresses practical issue related to admissions or decision making when number of places is limited. It also addresses four main considerations: test difficulty, examinee population, objective of the test, and agreement across judges. Although this method has not attracted any major critique, it is a compromised method; hence, none of the objectives is fully met. Using the standard deviations ratio is arbitrary, lacking any theoretical justification or proof of validity, and the cut score cannot be presented to the examinees prior to the test.

The Borderline Regression Method

This method is mainly used in the medical and health profession fields for clinical assessments that comprise examination scores (y) describing performance on particular tasks and an overall grade (x) which is the examiner's overall impression of the examinee. The overall grade scale is commonly comprised of at least four, deemed interval, categories (fail = 1, borderline = 2, pass = 3, distinction = 4). A linear regression model (y = ax + b) is then generated from all examination scores and overall scores of all examinees and the value of y when x = 2 (i.e., borderline) determines the cut score for the examination.

The strength of this method is that it is a simple method, which includes both checklist-type performance and overall examiner’s impressions of the examinees. There are some theoretical issues, particularly lack of theoretical justification for the interval nature of the overall score and the choice of linear model. Additional weaknesses are that application is limited to contexts when overall score is given in addition to the test score, and the cut score cannot be presented to the examinees prior to the test.

The Cohen Method

This method is mostly used in medical education but could be used more broadly. It is based on the assumption that the top 5% of the examinees have similar ability across cohorts, thus most of the variance in their mean score across tests would be attributed to the test rather than to the examinee population. When examinees sit for a test, the following formula is used to determine the cut score: Cut score = $C + 0.6 \times (P - C)$, where C is the expected percentage score due to guessing and P is the percentage score of the student at the 95th percentile. The value 0.6 is determined arbitrarily but reflects the institutional policy and ensures a minimum pass score.

The strength of this method is that it moderates the impact of variance in test difficulty and cohort ability, and its weaknesses are that it can be applied to multiple-choice tests only and the cut score cannot be presented to the examinees prior to the test.

Objective Borderline Method

This method assumes that there is a range of test scores (borderline) where it is uncertain whether an examinee within that range is competent or incompetent. To determine the cut score, a panel of judges agrees on three ranges of scores to be classified into three categories: pass (i.e., competent beyond doubt), fail (i.e., incompetent beyond doubt), and borderline (the remaining scores). The number of passes (P), borderlines (B), and fails (F) is placed in the following formula to establish the pass index: . Then the pass index determines the percentage of borderline scores that would be considered conceded pass. The cut score is determined by the lowest borderline score that was granted conceded pass.

Variations in this method have been introduced.

This method is a simple method whereby judges agree upon what is clear (clear pass and clear fail) rather than focusing on the elusive determination of the probability of a hypothetically minimally competent examinee giving a correct answer to an item. Judgment considers observed error, as the borderline range is the perceived test accuracy. Research on real and simulated data supports the utility and validity of the method. Although performing well compared to other methods, this method does have several weaknesses. A weakness is that no strong theoretical explanation has been found to explain the method. Other weaknesses are that the method is not applicable when test results do not include clear passes, and it is not effective when there are fewer than 50 examinees. In addition, the cut score cannot be presented to the examinees prior to the test, only the boundaries between fail and borderline and borderline and pass.

Additional Considerations

The literature describes many more standard setting methods; to date over 50 different methods and many more variations have been introduced. Standard setting is a technique or a process of decision making. Whenever decisions are made, human judgments must be involved.

Item writers and examiners may have perceptions of the desirable standards, but setting the actual cut scores is typically made by others, experts, or psychometricians. Standard setting is therefore a secondary decision-making process, which classifies test scores into different categories, most commonly pass and fail. Consequently, the following should be considered when standard setting process is performed: (a) the possibility that the standard setting process will overrule decisions already made by examiners, (b) the impact of judges' bias on the final cut score, (c) whether normative method or reference-based method was used for setting standards and determining cut scores, (d) whether the cut score can be presented to the examinees prior to the test, (e) the simplicity and perception of fairness and transparency across all stakeholders, (f) the logistics and feasibility of resources, and (g) the legal defensibility of the methods applied.

All methods have strengths and weaknesses, even beyond what is described here; and no method is superior to others. Ultimately, no standard setting can fully address all issues. Thus, when making the decision upon the preferable standard setting method, one needs to carefully weigh all strengths and weaknesses prior to deciding on a preferable method.

Boaz Shulruf

See also Achievement Tests; Angoff Method; Classification; Cut Scores; Ebel Method; Psychometrics; Tests

Further Readings

Angoff, W. (1971). Scales, norms, and equivalent scores. In R. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21(2), 147–152. doi:10.1111/j.1745–3984.1984.tb00226.x

Cizek, G. (2012). Setting performance standards: Foundations, methods, and innovations (2nd ed.). London, UK: Routledge.

Cizek, G., & Bunch, M. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. London, UK: Sage.

Cohen-Schotanus, J., & van der Vleuten, C. (2010). A standard setting method with the best performing students as point of reference: Practical and affordable. Medical Teacher, 32(2), 154–160. doi:doi:10.3109/01421590903196979

Hambleton, R., & Powell, S. (1983). A framework for viewing the process of standard setting. Evaluation & the Health Professions, 6(1), 3–24. doi:10.1177/016327878300600101

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14(1), 3–19.

doi:10.1177/001316445401400101

Nichols, P., Twing, J., Mueller, C., & O'Malley, K. (2010). Standard-setting methods as measurement processes. Educational Measurement: Issues and Practice, 29(1), 14–24. doi:10.1111/j.1745–3992.2009.00166.x

Pant, H., Rupp, A., Tiffin-Richards, S., & Köller, O. (2009). Validity issues in standard-setting studies. Studies in Educational Evaluation, 35(2–3), 95–101. Retrieved from http://dx.doi.org/10.1016/j.stueduc.2009.10.008

Shulruf, B., Poole, P., Jones, P., & Wilkinson, T. (2014). The objective borderline method (OBM): A probabilistic method for standard setting. Assessment and Evaluation in Higher Education. doi:10.1080/02602938.2014.918088

Kimberly Ethridge Kimberly Ethridge Ethridge, Kimberly

Anthony P. Odland Anthony P. Odland Odland, Anthony P.

Standardized Scores Standardized scores

1595

1597

Standardized Scores

Standardized scores are most often associated with statistics and statistical analyses. Generally, standardized scores refer to raw data being converted to standard or normalized scores in order to maintain uniformity in interpretation of statistical data. Typically, these interpretations are made based off of norm references, something that can be accomplished due to the standardization of scores. Although there are a number of examples of standardized scores used in the literature, it is typically the case that standardized scores are meant to represent standard scores, known as *Z* scores. After raw data are converted to *Z* scores, the scores can then be converted to a variety of other standardized units such as *T* scores.

In general, there are no units of measurement when discussing standardized scores. This is so that all scores are “standard” and uniform. Removal of units of measurement is done through the conversion or normalization process so that any units in the equation mathematically cancel each other out. The resulting standardized score is unitless and simply reflects a relationship in regard to other standardized scores. This entry explains *Z* scores and *T* scores and discusses how they are used.

Z Scores

Z scores are one of the most commonly used scores for data in statistics. They are also known as normal scores and standardized variables. The foundation for Z scores lies in the assumption that the scores in the population for a given variable is normally distributed. This means that if all scores were to be compiled, they would fall within a conventional bell curve layout when graphed. This assumption of normality allows for standardization of Z scores as well as standardization in the way in which the scores are interpreted.

Z scores are based off of population parameters, meaning that it is a representation of where a particular score falls in relationship to the entire population, not the sample of interest. A positive Z score means that a particular corresponding raw score fell *above* the population mean or average. A negative Z score represents a raw score that falls *below* the population mean. The numerical value of the Z score is actually the number of standard deviations above or below the mean, depending on the sign of the score. A Z score in the middle of the normal distribution has a mean of 0 and a standard deviation of 0, meaning that the score falls in the exact center of the normal distribution, at the 50% percentile.

The corresponding formula for the Z score is as follows: $Z = x - m\sigma$. The numerator contains the expression of the raw score x, being subtracted from the population mean, m. The denominator, σ, represents the population standard deviation. If this is unknown, it may be estimated from a random sampling of the population. All units mathematically cancel out, so that the Z score is unitless, as described earlier.

Z scores can be used to calculate many other statistical scores, such as T scores, to calculate prediction intervals, and in the z test. Of note, in order to use the z test for statistical analysis and hypothesis testing, the entirety of the population parameters must be known. This is unlikely and so while it is an option, z tests are rarely used in real-life statistical analysis.

T Scores

T scores are Z scores that have been shifted and converted to a different shaped distribution with a different mean and standard deviation. The mean for a T score is 50 and the standard deviation is 10. A T score can be used when the population mean and standard deviation are unknown, as this is commonly the case in research and other statistical analyses.

T scores can be calculated in a variety of different ways, depending on what information an individual has. If information about the sample is given, then the following equation may be used to calculate the T score, $t = x - mxsn$. The term

x represents the raw score, m is the sample mean, n is the sample size, and s is the standard deviation of the sample. Again, just as with *Z* scores, *T* scores are unitless as the units mathematically cancel out in the equations. A *T* score may also be calculated directly from a *Z* score with the following equation, $t = 10 \times z + 50$. Just as *Z* scores are used in the *z* test for hypothesis testing, *T* scores are used in hypothesis testing and in *t* tests. The *t* tests are more often used than *z* tests, as information about the population is not necessary to run them.

Although the sign of the *Z* score could tell the individual how many standard deviations above or below the mean a score fell, this is not necessarily the case for *T* scores. A *T* score below 50 would indicate that the score fell below the predicted population mean because the mean of the distribution of *T* scores is 50. On the contrary, a *T* score above 50 would indicate that the score fell somewhere above the hypothesized population mean. The *T* score itself represents how many standard deviations above the mean a score resides when plotted on the *t* distribution. This is identical to what the *Z* score represents on the normal curve or *z* distribution.

Clinical Implications

Standardized scores are clinically important and significant. Standardization of scores helps to keep score interpretation uniform across fields, disciplines, and even individuals. Just as standard metric measurements help to keep uniformity when used in literature and research, the same can be said for statistically uniform scores.

Additionally, almost all neuropsychological testing instruments as well as tests used in standardized settings in academia, such as the ACT and SAT, utilize standardized scores in order to interpret results. Standardization of academic test scores allows for individuals to be accurately compared to their peers who took the same measures, giving rise to events such as college admissions, job placements, and other situations in which ranking by test scores is involved. It also allows scores to be compared across time and cultures.

Use of standardized scores is important in multiple disciplines including psychology, education, the social sciences, and biology. Conversions between different standardized scores are straightforward and formulas are easily accessible for individuals. Furthermore, interpretations of scores, including positions on distribution curves and percentile rankings, are able to be used by different individuals and can be used in situations that rely on that information.

Kimberly Ethridge and Anthony P. Odland

See also Normal Curve Equivalent Score; Percentile Rank; Standardized Tests; *T* Scores; *t* Tests

Further Readings

Aron, A., Coups, E., & Aron, E. N. (2013). Statistics for the behavioral and social sciences: A brief course (5th ed.). Harlow, UK: Pearson Higher Ed.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284.

Dixon, W. J., & Massey, F. J. (1951). Introduction to statistical analysis (Vol. 146, p. 243). New York, NY: McGraw-Hill.

Gravetter, F. J., & Wallnau, L. B. (2016). Statistics for the behavioral sciences. Boston, MA: Cengage Learning.

Hunter, J. E., & Hamilton, M. A. (2002). The advantages of using standardized scores in causal analysis. Human Communication Research, 28(4), 552–561.

Kohn, A. (2000). The case against standardized testing: Raising the scores, ruining the schools. Portsmouth, NH: Heinemann.

Kristin M. Morrison Kristin M. Morrison Morrison, Kristin M.

Susan E. Embretson Susan E. Embretson Embretson, Susan E.

Standardized Tests

Standardized tests

1597

1601

Standardized Tests

Standardized tests are evaluative devices or procedures developed to ascertain a sample of behavior from an individual in a domain of interest, in which the test administration and scoring process is uniform across individuals, and both reliability and validity evidence exists such that inferences regarding the person’s trait can be made from the test score. These assessments are often used to compare individuals or groups on current and predicted performance in different domains. For example, a paper-and-pencil mathematics test may be administered to a group of individuals to obtain a sample of mathematics ability or knowledge. Each individual receives the same instructions, is given the same amount of time to complete the test, and all answers are scored the same way (e.g., one point awarded for a correct answer, no points awarded for an incorrect answer); this approach to administration and scoring has been standardized, such that it is equal across all examinees. Using this test score, inferences about the individual’s mathematics proficiency can be made.

Tests are developed to make claims about an individual and can be created for different fields, domains, and constructs. For example, standardized tests exist for educational achievement, attitudes, vocational interests, personality, cognitive functioning, and mental health. Each test in the field may be designed to examine one construct (e.g., mathematics ability in education, neuroticism in personality) or multiple constructs (e.g., mathematics and reading in education). The test may then be used to determine whether an individual possesses all necessary skills (e.g., master or nonmaster in the domain) or to compare the individual to others (e.g., how does the individual perform in relation to others).

individual to others (e.g., how does the individual perform in relation to others). A sample of behavior may be obtained using different methods, such as a paper-and-pencil test, a computer-based test, interviews, or observations. Typically, tests are composed of items, which may be multiple choice, constructed response, true/false, matching, or ratings. Lastly, each examinee receives the same set of conditions for test administration, and all responses are objectively scored and interpreted.

This entry begins with a brief history of standardized tests, discusses developing and interpreting standardized tests, and concludes with several examples.

Brief History

Testing has its roots in civil service, academic achievement, and individual differences. Examinations date back over 3,000 years with the inclusion of testing in China for civil service appointments. After appointments were made, routine testing was conducted to determine whether the civil service official should remain in office or be replaced. Standard practices, such as those still in existence today, were developed. Examinee identities were confidential, exams were reviewed by multiple independent sources before their use, and administration conditions were identical for all examinees. Universities and other schools developed methods for training test administrators to ensure objective implementation and scoring of both oral and written examinations. Subsequently, examination of individual differences, or how different individuals varied on a large scale, heralded a new age of testing.

Francis Galton pioneered the work in this area with his Anthropometric Laboratory; he gathered data and developed statistical techniques to examine it. From this interest in individual differences, researchers, such as Charles Spearman, Alfred Binet, and E. L. Thorndike, conducted work in the field of intelligence testing. With the development of intelligence and group testing, testing gained popularity and expanded into other professional fields.

Although testing has been around for millennia, World War I and the invention of group testing are often cited as the beginning of standardized testing. The U.S. Army created the first large-scale testing program to test all military recruits. With group testing, large numbers of recruits could be tested with uniformity, objectivity, and reliability. As testing became widespread in the military, its use in other domains, such as education and personnel selection, increased. Tests were used to determine whether a student should be admitted to a university or if were used to determine whether a student should be admitted to a university or if one job candidate was more ideal than another.

In 1950, the American Psychological Association convened a committee whose goal was to determine the aspects of a test that should be investigated before the test was published. In 1954, The *Technical Recommendations for Psychological Tests and Diagnostic Techniques* was published. A year later, the American Education Research Association and what is now known as the National Council on Measurement in Education released their *Technical Recommendation for Achievement Tests*. A joint committee comprising members of American Psychological Association, American Education Research Association, and National Council on Measurement in Education has since replaced these earlier documents with the *Standards for Educational and Psychological Testing*. In 2014, an updated version of these standards was published. These standards are applicable to all tests used to make inferences about examinees. The standards provide a framework that promotes the development of tests and the assessment of their reliability and validity to support the inferences made about examinees.

Test Development and Principles

A process of development is utilized to create standardized tests. First, the construct of interest is elucidated and understood, so that samples of behavior can be taken. This stage is followed by test specifications, such as a blueprint that explicates the representation of the various aspects of the target construct. Then, the development of the test begins. A large bank of items related to the construct are written. All items are examined for content relevance, clarity, vocabulary, and relevancy to the test specifications (i.e., blueprint). Any items that do not meet standards in these areas are revised or removed from the item bank. The remaining items are field tested by administering the items to a sample of people similar to the target population for the test. Field testing allows an empirical examination of the items and their characteristics. Empirical data allow for statistical information, such as item difficulty and item discrimination, to be obtained for each of the items. Previously, classical test theory was utilized to obtain this information, but recent trends indicate the use of item response theory (IRT), in conjunction with classical test theory, as a common approach to item examination, as well as other aspects of test development.

The empirical information gained through field testing is used to further pinpoint items that might be troublesome. For example, an item that differs in difficulty based on gender may not be appropriate for the test. Further analysis, such as based on gender may not be appropriate for the test. Further analysis, such as differential item functioning, may be conducted to determine whether the item-to-test-score relationship is the same across groups. Sources of differential item functioning may then be examined. For example, the wording of the item may be less appropriate for one of the groups and may need to be revised. Also, the items, and the test as a whole, may be examined for dimensionality issues through correlation with other measures. If the test aims to test mathematics ability, then high correlations should exist between the new test and older tests measuring the same domain. However, if high correlations exist between the new test and test in another domain (e.g., reading), the test might be multidimensional (i.e., measuring more than the one intended construct). Items contributing to inappropriate correlations may be rejected or revised and then field tested again.

Once items survive field testing, multiple activities occur. One activity is to create a standard method of administration. This standardization includes the mode of administration (e.g., paper and pencil, computer), the time given to complete the assessment, the instructions given, and any other conditions the test developer deems necessary to be similar across examinees. Another aspect of the test that needs to be established is whether examinees will receive the exact same assessment or parallel assessments. Parallel assessments are those that are composed of similar, but different, items. If examinees are given the exact same test, they all receive the same items and often receive them in an identical order. If parallel tests are given, examinees receive tests that have been equated. Equating is a process in which scores from the parallel tests have been placed on a common scale, so that the two scores can be compared. Traditionally, classical test theory was used to create parallel forms, but it has become common practice to also utilize IRT for its greater flexibility.

Although equating requires the use of statistical information (e.g., item difficulty) in the development process, content across the parallel test forms must also be comparable. If a test is designed to measure mathematics, but one version involves only algebra and another requires strictly geometry, these two tests are not comparable in terms of content. Thus, when equating takes place, specific conditions must be met. The tests must measure the same domain (e.g., algebra and geometry). The equating process must achieve equity; in other words, an examinee’s score should be similar regardless of which form of the assessment the examinee is given. Equating should be invariant across populations, regardless of which population was used during the equating process. Lastly, a given score on one test should correspond to a score on the other test, regardless of which test is equated to the other (i.e., Test 1 equated to Test 2 or vice versa).

Traditionally, tests have been administered via paper and pencil with a fixed length. Such tests are still heavily used, such as in personality measurement and within school systems for end-of-year testing. However, equating on the item level, via IRT, allows for the utilization of adaptive testing. Adaptive testing tailors the assessment to the individual; successive items are administered based on the individual's responses to previous items. For example, if an examinee answers an item correctly, a more difficult item may be administered next. However, if the examinee answered the item incorrectly, an easier item may be given instead. These types of tests are not readily administered in paper-and-pencil formats. However, the earliest adaptive test was Alfred Binet's intelligence test, which was administered to children by an administrative examiner. However, with the advent of computers, adaptive testing has become easier. Computerized adaptive testing, using IRT-based item calibrations, has now become a standard option for administering assessments to examinees. Items included on the test still go through all of the aforementioned standardization processes (i.e., field testing, content examination, equating) but allow for highly individualized assessments to be administered to increase measurement precision.

Remaining issues concern how the test items will be scored and how scores will be interpreted. For standardized tests, items are scored in a similar way for all examinees. Item scoring may be binary, where only two scores are possible. For example, a score of 1 may be given for a *correct answer* and a score of 0 for an *incorrect answer*. Often, this type of scoring is used for multiple-choice items, in which an examinee selects the correct answer from a set of possible responses. However, other scoring possibilities may exist. Polytomous scoring may be used when an item has more than two possible scores. An example of this is an essay question, where higher scores are given to better responses. Another example is a mathematics item in which the examinee is required to show all steps necessary to solve the item; points may be given for the number of steps successfully completed, thus resulting in partial credit. These scores are then utilized to make inferences and decisions regarding the individual. Two approaches exist to giving the score meaning and are discussed later in this entry.

Validity

A number of the procedures discussed in the test development section help ensure the reliability and validity of test scores. Reliability refers to the consistency of the estimate, or score, obtained from a test across conditions. For example, an examinee should receive similar scores on a test when given at different times as well as should receive similar scores on parallel forms. While reliability of measurement is necessary, it is not sufficient. The validity of the test for the intended purposes must also be established. Validity refers not just to the test or the test scores, but how these test scores are used to form interpretations. It requires evidence and theory to support these proposed uses to ensure that the interpretations are appropriate. Thus, reliability deals with consistency in the measure while validity relates to what the test measures.

Past views of validity specified distinct types of validity: content-related, criterion-related, and construct-related validity. However, the current validity concept (i.e., *testing standards*) is unitary; the single type of validity is construct validity. Validity evidence has several aspects, including test content, examinee response processes, internal structure, test consequences, and relationships to other variables.

Evidence for validity based on test content is achieved during development of items, review of items, and score development and interpretation. These processes ensure that the content represented is related to the construct of interest and to how the scores will be used. Response processes are examined, so that the relationship between the construct and the examinee’s process to a response is established. One way to establish this type of evidence is to require examinees to report on the process they use to arrive at their answer and then compare it to the intended approach to measure the construct. Internal structure evidence relates to the interrelationships of the items and dimensionality of the assessment. For evidence on relationships to other variables, test scores should have strong relationships to measures that examine the same construct and weak relationships to measures that examine different constructs. Furthermore, if test scores are used for prediction, appropriate correlations with criteria should be available. Lastly, score use and impact (i.e., test consequences) should be examined. Decisions based on tests should not have undesirable social consequences (i.e., adverse impact) for certain groups and should have equal predictive value for all examinees.

Norm-Versus Criterion-Referenced Assessments

Norm- Versus Criterion-Referenced Assessments

Development of standardized tests requires the test developer to make multiple decisions regarding the goals of the test. One decision that must be made is how an individual’s final score on the test will be used. In other words, what gives meaning to the score must be chosen. One approach to give meaning to this final score is to interpret the individual’s test score in reference to a representative group of peers. This approach is considered norm-referenced measurement, and individual differences are key. The final score is transformed from a raw score to a standardized scale score. Performance by a large, representative, reference (i.e., norm) group of individuals is used as a norm to which the scale score is compared. Therefore, this approach allows individuals to be ranked in terms of performance. Standardized tests using this approach to measurement are used in many settings. For example, the SAT norm-referenced scores are used to predict academic performance and select students for scholarships.

Another approach to attaching meaning to the final score on a test is criterion-referenced measurement, such as the General Educational Development assessment. This approach simply provides an absolute level of performance for the individual, which is then interpreted in relation to a predefined criterion. Tests using this approach are often used to classify individuals into various competency groups (e.g., master, nonmaster). This classification can be used to determine whether the individual needs remedial help or the individual can graduate, such as from high school. Often it is difficult to decide where the criterion should be located to determine mastery. A criterion of 100% mastery is infeasible, as doing something 100% every time is rare. Thus, lower criterions, such as 70% or 80% mastery, are often used as the standard for determining classification. Although criterion-referenced tests do not compare individuals, performance by a representative group of individuals might be used to help establish a realistic criterion.

Standardized Test Examples

Standardized tests can be of various formats, such as intelligence tests, college admissions tests, military tests, and state tests. A famous standardized intelligence test is the Stanford-Binet Intelligence Scales. The fifth edition of this test (SB5) examines four cognitive areas: verbal reasoning, quantitative reasoning, abstract/visual reasoning, and short-term memory. Another standardized intelligence test is the Wechsler Adult Intelligence Scale.

Personality and clinical tests are often standardized tests. Examples of these tests include the Minnesota Multiphasic Personality Inventory and the Myers-Briggs Type Indicator. The military requires all applicants to take the Armed Services Vocational Aptitude Battery to determine whether an applicant is qualified to enlist in the military and to place recruits into jobs.

Education utilizes standardized tests for many different focuses. High school exit exams, such as the California Achievement Test, are standardized tests to determine whether high school students have obtained the necessary knowledge to graduate from high school. The General Educational Development exam determines whether individuals who did not graduate from high school have the academic skills that a high school graduate possesses. Some tests, such as the GRE, the SAT, and the Graduate Management Admission Test, are used to determine examinees' eligibility for admission into different universities and programs. Standardized tests are used to determine language proficiency; an example is the Test of English as a Foreign Language.

Lastly, certification exams determine whether a candidate is qualified to perform a specific job or task. Lawyers must pass their state bar exam before they can practice. Accountants are required to pass their state's Certified Public Accountants exam. Teachers must pass certification exams, such as the PRAXIS, to determine teaching eligibility. Multiple certification exams are used in the medical field to ensure that doctors (e.g., medical licensing exam), nurses (e.g., NCLEX-RN for registered nurses), and pharmacists (e.g., NAPLEX) possess the required skills to obtain a license before practicing.

Kristin M. Morrison and Susan E. Embretson

See also Classical Test Theory; Computerized Adaptive Testing; Equating; Item Response Theory; Reliability; *Standards for Educational and Psychological Testing*; Validity

Further Readings

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Mason, OH: Cengage Learning.

DuBois, P. H. (1970). A history of psychological testing. Boston, MA: Allyn & Bacon.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Urbina, S. (2014). Essentials of psychological testing (2nd ed.). Hoboken, NJ: Wiley.

Ronli Diakow Ronli Diakow Diakow, Ronli

Standards for Educational and Psychological Testing Standards for educational and psychological testing

1601

1605

Standards for Educational and Psychological Testing

The *Standards for Educational and Psychological Testing* is a document that articulates a set of professional standards for the development and use of educational and psychological tests. These guidelines are intended to inform all aspects of testing and to provide a basis for evaluating the quality of tests. The *Standards* covers a wide range of issues related to all aspects of the testing process, including foundational concepts such as validity and fairness, details of test development and implementation, and applications in specific areas such as employment credentialing and educational accountability.

The *Standards* is a joint publication of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education and has been approved or endorsed by each of these organizations. The primary audience for the *Standards* is professionals working in the fields of educational and psychological measurement. Given the expanding reach of testing, the *Standards* is also increasingly relevant to a wider audience encompassing policy makers, classroom teachers, and employers. This entry first discusses the history, purpose, and use of the *Standards*, then looks at the content and organization of the 2014 edition of the *Standards*.

History, Purpose, and Use of the *Standards*

History of the Standards

In 1954, the American Psychological Association first prepared and published a set of guidelines for testing entitled *Technical Recommendations for Psychological Tests and Diagnostic Techniques*; in 1955, the American Educational Research Association and the National Council on Measurement Used in Education (this organization is now National Council on Measurement in Education) first prepared and published their own testing guidelines entitled *Technical Recommendations for Achievement Tests*. In 1966, the three organizations first published a joint set of standards entitled the *Standards for Educational and Psychological Tests and Manuals*. Joint committees with members representing all three organizations have since revised this document four times, with publication of subsequent editions of the *Standards* in 1974, 1985, 1999, and 2014. The name *Standards for Educational and Psychological Testing* was first used in 1985.

As reflected in the titles, the earliest documents focused on technical aspects of test development and on the documentation of the test development process. With the third edition in 1974, the document took a more expansive view of test development, use, and reporting. The 1985 revision was the first to discuss validity as a unitary concept, and the 1999 revision reframed validity around the interpretation of test scores for particular uses. The 1999 revision also broadened the definition of “test” to include an expanded set of instruments and emphasized the decision-making process of both the design and use of tests. Key changes in the 2014 revision include an increased emphasis on fairness and accessibility as fundamental issues, introduction of a new chapter on the use of tests for educational accountability and policy, and expansion of the discussion of the impact of new technologies throughout.

Purpose of the Standards

The introduction to the 2014 *Standards* articulates multiple purposes for the document. These include (1) promoting sound testing practices, (2) providing criteria for the development and evaluation of tests and testing practices, and (3) providing guidelines for assessing the validity of interpretations of test scores for the intended test uses. The *Standards* provides a general framework for ensuring that all of the relevant issues for testing are addressed when tests are being created and used. The *Standards* does not advocate for particular test development procedures or psychometric methodologies; rather, it is meant to be widely applicable across different testing situations in which different specific methods are appropriate.

The *Standards* is not a legal document and is not intended to prescribe legally binding requirements. The *Standards* itself also provides no mechanism for enforcement of their provisions. However, the *Standards* has been used by regulatory authorities and courts to define acceptable practices in testing. In education, the *Standards* is almost always referenced in requests for proposals to create high-stakes tests, and adherence with the *Standards* is even codified into state law concerning assessment in some states.

Use of the Standards

The use of the *Standards* is relevant across the testing process and across participants in the testing process, from the design of the test to the selection or purchase of the test to the implementation of the test to any reporting or decision making as a result of the test. Before a test is used, it is the responsibility of both test developers and test users to exercise due diligence regarding the evidence behind any statements made regarding adherence to the *Standards* for a specific test. When considering the specific standards, attention should be paid to the *Standards* as a whole. Individual standards are not intended to be read or applied in isolation but rather in conjunction with any related standards.

The *Standards* is not intended to be a checklist of criteria nor to supplant the need for professional judgment in designing and evaluating tests. Not all standards will apply to all testing situations, and not all standards will be appropriate or feasible in all situations. Standards that are most important will depend on the specific testing situation. Professional judgment is needed to interpret the standards in light of the particulars of a test or testing program. In addition, though the foundational concepts underlying the standards apply to all testing situations, there are often situations in which it would not be practical nor desirable to follow the standards, such as for teacher-developed classroom assessments. In general, critical adherence to the standards should increase as the stakes or potential consequences of testing increase.

Content of the 2014 *Standards*

Organization of the Standards

The *Standards* is organized into three parts, each of which contains a number of chapters. Part I: Foundations comprises three chapters containing standards for validity, reliability/precision, and fairness in testing. Part II: Operations comprises six chapters containing standards for test design and development; scores, scales, norms, score linking, and cut scores; test administration, scoring, reporting, and interpretation; supporting documentation for tests; the rights and responsibilities of test takers; and the rights and responsibilities of test users. Part III: Testing applications comprises four chapters containing standards for psychological testing and assessment; workplace testing and credentialing; educational testing and assessment; and uses of tests for program evaluation, policy studies, and accountability.

Each chapter begins with a section of background text that introduces the major themes and central concepts for applying or using the standards in that chapter. Then, each standard is presented along with accompanying comments to clarify, expand on, and/or give examples to aid in the interpretation and use of the standard. Within each chapter, standards are organized into clusters that comprise groups of standards with similar themes. In Parts I and II, the first standard in each chapter is an overarching standard that encompasses the central purpose of the standards in that chapter. Some standards, particularly in Part III, are repeated in multiple chapters with slight variations so that they are not missed by users who only refer to the chapters most relevant to their work. A glossary provides the definition of key terms assumed throughout the book.

Part I: Foundations

Standards for Validity: The *Standards* primarily defines validity in terms of the strength of an evidentiary argument behind test score interpretation for specific uses. The first cluster of standards addresses establishing the intended uses and interpretations of a test; these standards emphasize the definition of the construct, population, interpretation, and use and also discuss how to incorporate new or unanticipated interpretations or uses. The second cluster addresses the collection of data to evaluate validity, with a focus on the detailed description of the procedures, data, and analysis. The third cluster addresses a variety of specific kinds of validity evidence, including evidence based on test content, cognitive processes, internal structure, relationships with criteria and other constructs, and consequences. The ongoing and judgmental nature of a validity argument is emphasized.

Standards for Reliability/Precision and Errors of Measurement: The *Standards* takes a broad view of reliability by coupling it with the notion of precision and defining reliability as consistency of scores across repeated applications of the testing procedures. Two clusters of standards deal with defining the relevant set of (possibly theoretical rather than empirical) replications that underlie an evaluation of reliability/precision. A second set of clusters of standards deal with documenting and reporting reliability/precision and standard errors of measurement with an emphasis on minimizing misinterpretation or overgeneralization. Other clusters address reliability/precision in the context of individual decisions and group-level reporting. The implications for the validity argument are discussed in the background section.

Standards for Fairness in Testing: The *Standards* describes a fair test as one that neither advantages nor disadvantages any individuals because of characteristics irrelevant to the intended construct. The first cluster of standards reflects the principles of universal design and focuses on minimizing construct-irrelevant variance for a wide range of individuals and subgroups. The third cluster of standards addresses developing, providing, and documenting accommodations when they are required to remove construct-irrelevant barriers. The second and fourth clusters of standards focus on ensuring fairness of score uses and interpretations in light of potential barriers or noncomparability across individuals or subgroups. All of the standards related to fairness are framed as fundamental issues for validity.

Part II: Operations

The *Standards* contains guidelines for all aspects of test development and implementation. The main unifying thread in the six chapters of standards related to operations is the gathering of evidence, which relates back to the validity argument.

Standards for test design and development focus on ensuring detailed documentation of the process used and any evidence gathered to support that process, including detailed descriptions of test specifications, item development, and administration and scoring procedures. Standards for scores, scales, norms, score linking, and cut scores focus on the rationale for score selection and scoring procedures and on the interpretation of different types of scores. Standards for test administration, scoring, reporting, and interpretation focus on ensuring standardized procedures for administration and scoring and the integrity of reporting and interpretation. Standards for supporting documentation for tests focus on distilling the relevant information from the three preceding chapters on test design, development, administration, and scoring so that test users can make informed decisions about tests and use them appropriately.

Standards related to test takers' rights and responsibilities focus on fairness from the perspective of the test taker to guide test providers' policies around information for test takers, security of test data, and procedures for resolving irregularities (e.g., suspected cheating). Standards related to test users' rights and responsibilities focus on validity evidence and fairness from the perspective of professional test users (e.g., professionals who select or administer tests), with emphasis on how to consider issues of interpretation, reporting, and security for a specific population and use.

Part III: Testing Applications

The *Standards* contains guidelines relevant to specific applications relevant to the members of the three sponsoring organizations. The four chapters of standards related to testing applications contain guidelines that tailor general topics already addressed for the specific application as well as guidelines for issues that are particularly relevant to that application.

The standards for psychological testing and assessment include standards related to the use of test batteries and the use of tests for diagnosis. The standards for workplace testing and credentialing include standards related to the definition of content for selection decisions and evidence related to the prediction of criteria. The standards for educational testing and assessment include standards that address opportunity to learn, links to instruction, educational decisions (such as placement, promotion, and graduation), and mandated testing programs. The standards for uses of tests for program evaluation, policy studies, and accountability include standards that address using tests for multiple purposes, multiple sources of evidence, and negative or unintended consequences.

Ronli Diakow

See also *Guiding Principles for Evaluators*; Joint Committee on Standards for Educational Evaluation; Reliability; Testing, History of; Validity

Further Readings

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Educational Measurement: Issues and Practice, 33(4), 1–43.

Camara, W. J., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. Educational Measurement: Issues and Practice, 25(3), 35–41.

Eignor, D. R. (2001). Standards for the development and use of tests: The standards for educational and psychological testing. European Journal of Psychological Assessment, 17(3), 157.

Eignor, D. R. (2013). The standards for educational and psychological testing. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology (pp. 245–250). Washington, DC: American Psychological Association.

Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In Handbook of test development (pp. 27–38).

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? Educational Measurement: Issues and Practice, 33(4), 4–12.

Megan E. Welsh Megan E. Welsh Welsh, Megan E.

Standards-Based Assessment Standards-Based assessment

1605

1607

Standards-Based Assessment

The term *standards-based assessment* refers to any assessment used to gauge student mastery of state-adopted content standards, which outline things that students are expected to know and be able to do. Standards-based assessment is a key aspect of standards-based reform, which was implemented in the 1990s and is based on the notion that if states set standards, adopt high-stakes tests that measure them, and hold schools accountable for test performance, they will create an incentive for teachers to provide high-quality instruction on the standards. Classroom assessments are also often standards based in that they are intended to gauge student attainment of standards and provide information on student progress over the course of the school year. This entry first discusses standards-based assessment used on a statewide basis and in individual classrooms, then looks at issues with standards-based assessment.

Large-Scale Standards-Based Assessment

According to federal law, students must be tested in English/language arts (ELA) and in mathematics from Grades 3 through 8 and once in high school. Students are also tested in science at least once in elementary school, middle school, and high school. The results of these assessments are used to monitor student progress over time, to evaluate schools, and (in some states) to evaluate teachers or to determine which students may graduate from high school. As such, they are considered high-stakes tests with serious consequences. These uses are predicated on assumptions that the tests gauge the standards and that test scores reflect the extent to which teachers provided high-quality, standards-based instruction. However, there is strong evidence to suggest that standards-based assessments measure only a subset of state standards and also test skills omitted from standards altogether.

The match between standards and assessment is called alignment, and achieving strong alignment is difficult. Because of this, and because no single measure can provide an adequate picture of what people know or are able to do, several research organizations (the American Educational Research Association, the American Statistical Association, and the American Evaluation Association) recommend that accountability systems are based on more than just a single test score.

States determine which standards and assessments to implement, and there is variability in what is taught and assessed across states. To increase consistency in educational expectations nationwide, states joined forces to develop the Common Core State Standards (CCSS) in ELA and math, which were adopted by 46 states in 2010 and 2011 (one of the states, Minnesota, only adopted the ELA portion). By late 2016, nine states announced they would replace or significantly revise the CCSS, but critics in some of these states say the new standards ended up being nearly identical to the CCSS. The CCSS set expectations in ELA and in mathematics for each grade level and are purported to emphasize higher order thinking and focus on literacy with respect to both fiction and nonfiction texts. National standards have also been developed in science (the Next Generation Science Standards), and state departments of education and professional associations have developed standards in a wide array of other content areas.

Using federal funding, two consortia of states developed assessments that together were administered by 28 states in 2015–2016. Both the Smarter Balanced Assessment Consortium test and the Partnership for Assessment of Readiness for College and Careers test are computer administered, and the Smarter Balanced Assessment Consortium test is computer adaptive. Computer-adaptive tests tailor each test to examinees who are given items of increasing or decreasing difficulty based on their level of success with prior items. Both assessments are also designed as full-scale testing systems. They include a summative test, similar to prior standards-based large-scale tests, and also provide (a) optional interim assessments that teachers can administer throughout the school year to monitor student progress and (b) resources to help teachers conduct CCSS-aligned formative assessment.

States are independently developing Next Generation Science Standards–aligned assessments, which were expected to be administered beginning in 2018. Some states have also developed large-scale, standards-based assessments in other states have also developed large-scale, standards-based assessments in other content areas. For example, New York administers assessments in foreign languages and in social studies.

Standards-Based Classroom Assessment

Although many associate standards-based assessment with high-stakes, large-scale tests, standards-based classroom assessment is also widespread. Teachers are expected to teach to the standards and to monitor student progress, therefore virtually all classroom assessment should be standards based. Many school districts have adopted standards-based progress reports, which require teachers to report on attainment of key standards at each grade level instead of on overall content area performance and which use a grading scale consistent with large-scale test performance levels (e.g., meets standard, exceeds standard) in lieu of an A–F scale.

Those who interpret standards-based progress reports often expect consistency between end-of-year grades and large-scale test scores. However, there are several factors that might lead to differences. First, grades reflect an array of student performances and may therefore capture different aspects of the standard than are captured by a large-scale test. Second, teachers may differ from tests in their understanding of the achievement required to earn a certain performance level. Finally, each standard often includes multiple skills, knowledge, and abilities and multiple ways in which students are expected to demonstrate their knowledge. Therefore, both classroom assessments and large-scale tests can be aligned to standards but address different aspects of that standard.

Issues With Standards-Based Assessment

Standards-based assessment is often controversial because of the accountability systems that it informs and the amount of time schools devote to assessment. It is important to remember that standards, assessment, and accountability systems are distinct. However, evidence does suggest that schools have increased the amount of instructional time devoted to tested topics and are devoting more time to testing in response to test-based accountability.

To evaluate the quality of standards-based assessments, one must examine both the degree of alignment between standards and assessments and, if they are to be used as accountability measures, the extent to which standards-based test scores vary as a function of instruction. The research conducted thus far does not tend to identify tests that are strong according to either criterion. Classroom standards-based assessments have not been rigorously evaluated, and it would be impractical to do so across a wide array of classrooms. However, they do provide useful information about the ways in which teachers implement the standards and are very valuable to teachers, students, and parents. Ultimately, the power of standards-based assessment depends both upon the quality of the measures themselves and the ability of people to wisely use them.

Megan E. Welsh

See also Accountability; Common Core State Standards; Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium; Standardized Tests; State Standards

Further Readings

Council of Chief State School Officers. (2016). Common Core State Standards Initiative. Retrieved from http://www.corestandards.org/

Guskey, T. R., & Bailey, J. M. (2010). Developing standards based report cards. Thousand Oaks, CA: Corwin.

Herman, J. L., & Linn, R. L. (2014). New assessments new rigor. Educational Leadership, 71(6), 34–37.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? American Educational Research Journal, 48(4), 965–995. doi:10.3102/0002831211410684

Popham, W. J. (2007). Instructional sensitivity: Educational accountability's dire deficit. Phi Delta Kappan, 89(2), 149–155.

Tracy Paskiewicz Tracy Paskiewicz Paskiewicz, Tracy

Stanford-Binet Intelligence Scales

Stanford-Binet intelligence scales

1607

1609

Stanford-Binet Intelligence Scales

The *Stanford-Binet Intelligence Scales, Fifth Edition* (SB5) is an individually administered test of cognitive and intellectual abilities of preschoolers, school-age children, adolescents, and adults. It covers the age range of 2 through 85+ years. The SB5 is authored by Gale H. Roid and was published in 2003 by Riverside Publishing. This edition is a recent addition to a series of well-known instruments including the Binet-Simon Intelligence Scale (1908), the original Stanford Binet (1916), Stanford-Binet Forms L and M (1937), Stanford Binet-III (1960), Stanford Binet-IV (1986), and SB5. This instrument is widely used in the assessment of intellectual ability in clinical and education settings within the United States. This entry focuses on the structure and content of the instrument, available scores, uses in various settings, psychometric properties of the test, and strengths and weaknesses of the test.

Structure and Content of the Instrument

The SB5 measures five factors important to understanding intelligence across both verbal and nonverbal domains. The structure of the test allows practitioners to gather multiple forms of data relevant to understanding an individual’s cognitive abilities. The SB5 may be used to measure a single dimension of general intelligence, a two-factor verbal and nonverbal model, and a five-factor model in which both verbal and nonverbal scales are combined to measure a particular cognitive skill, such as fluid reasoning or working memory. The SB5 contains 10 subtests, five of which are considered verbal and five of which are considered nonverbal. The verbal tests include domains that are measured verbally—with a verbal response required by the examinee. The nonverbal tests tap into nonverbal cognitive skills; there is minimal or no verbal response required from the examinee. Each SB5 subtest is composed of a “testlet,” which is a brief mini-test at each level of difficulty. Examples include picture absurdities, matrices, vocabulary, memory for sentences, quantitative reasoning, and verbal absurdities. The instrument is aligned with several factors of the Cattell–Horn–Carroll model of intelligence and measures cognitive abilities in the following domains: fluid reasoning, visual/spatial processing, knowledge, working memory, and quantitative reasoning.

The 10 subtests comprise the Full-Scale IQ (FSIQ). The Abbreviated Battery IQ consists of two routing tests and can be used as a brief screener for IQ. The Stanford-Binet is unique in its use of routing subtests. Test administration is expedited through adaptive testing via the routing procedure. The first two subtests are routing tests and are used to determine the start points for the remaining verbal and nonverbal tests. By adapting the test to the examinee’s functional level, the SB5 routing procedure increases measurement precision (and minimizes testing time) by tailoring the difficulty of items to the examinee.

Available Scores

Composite scores available on the SB5 include the Verbal IQ, Nonverbal IQ, Full-Scale IQ, and Abbreviated Battery IQ. There are also five domain scores (fluid reasoning, visual/spatial processing, knowledge, working memory, and quantitative reasoning), made up of one verbal and one nonverbal subtest. Finally, the test offers a scaled score for each domain (e.g., verbal fluid reasoning and nonverbal fluid reasoning). The SB5 is easily scorable by hand; however, computer-based software is also available through the publisher. The composite scores are expressed as standard scores, with a mean of 100 and a standard deviation of 15. Subtest scores are expressed as scaled scores, with a mean of 10 and a standard deviation of 3.

Uses

The Stanford-Binet is a prominent measure of intelligence, used widely in schools, clinics, universities, hospitals, and research settings throughout the nation. The instrument is useful in clinical, neuropsychological, and psychoeducational assessment; early childhood assessment; research on intelligence; and adult social security evaluations. Typically, the SB5 would be used as part of a comprehensive evaluation to investigate or diagnose developmental disabilities, learning disabilities, and other exceptionalities.

Many practitioners find the SB5 attractive for younger or lower functioning individuals because the test kit offers numerous colorful manipulative toys, blocks, and everyday items. There is reduced emphasis on speeded performance, as processing speed (often measured on cognitive batteries) is not included in the SB5. There are many low-end items to better assess young children, low-functioning older children, and adults with intellectual disabilities. There are additional items that measure very high functioning, as in the assessment of intellectual giftedness.

Qualifications for individuals using the SB5 include having sufficient training to administer and score psychological tests accurately and reliably as well as training to report and interpret results of psychological tests. Typically, those using SB5 have college or graduate-level training on tests and measurement as well as statistical concepts essential for understanding test scores and have the ability to translate results into consumer-friendly language, so parents, educators, and adult clients can easily understand what test results mean. In addition, test users must be mindful of ethical behavior related to psychological measurement, including that test items and materials are kept secure at all times and abiding by copyright laws regarding the photocopying and use of the test materials.

Psychometric Properties

The SB5 was developed and standardized in the late 1990s, with a publication date of 2003. It was normed on a nationally representative sample of 4,800 individuals. The standardization sample was matched to demographics specified by the 2001 U.S. Census. Bias reviews were conducted on all test items for gender, ethnic, cultural, religious, regional, and socioeconomic issues.

Reliability Evidence

The technical manual of the SB5 reports strong evidence for reliability. Average internal consistency reliabilities are in the range of .91 (Abbreviated Battery) to .98 (Full-Scale). The reliabilities of the five factor indexes average .90 or higher. Reliabilities of the 10 subtests average .84 or higher. Both test–retest and split-half (i.e., internal consistency) estimates of reliability are discussed in the technical manual. Reliability data are appropriately high (i.e., test–retest average technical manual. Reliability data are appropriately high (i.e., test–retest average median reliability estimate = .86; split-half average median reliability estimate = .90), demonstrating that the instrument is both consistent across time and within form.

Validity Evidence

The relationships between the SB5 and other measures of cognitive ability and achievement are reported in the technical manual. Concurrent validity evidence is strong with a reported correlation of .90 between the SB5 Full-Scale IQ and the Stanford-Binet IV Composite. Convergent validity evidence between the SB5 and other established measures (e.g., academic achievement) is also provided in the technical manual. Extensive validity studies were conducted, including clinical group differences, age trends, factor structure, and consequential validity. The author relied on confirmatory factor analytical procedures to provide validation support for the structure of the instrument.

Strengths and Weaknesses

The SB5 has a number of useful applications to practitioners, primarily as part of evaluations to diagnose disabilities and exceptionalities in children, adolescents, and adults. Practitioners have cited engaging and fun activities for young children among the instrument’s strengths. The toys, manipulatives, and everyday items included in the test kit make the SB5 particularly useful for early childhood assessment, as children are engaged in play-like activities that do not seem like a test. However, the number of manipulatives and toys to handle can be initially daunting for a practitioner learning to use the instrument. For example, it takes a good deal of practice to know which test materials are needed for each item.

The nonverbal battery of the SB5 can be used to assess individuals with limited or questionable linguistic abilities, such as those who are deaf or hard of hearing and individuals with communication disorders, autism spectrum disorders, limited English-language background, or other conditions such as aphasia or stroke. The verbal battery of the SB5 is used for standard administrations as well as in special cases where subjects have limited vision or orthopedic impairment. Most verbal battery tasks involve listening to directions and making a verbal response to each item.

As with any test of intelligence, practitioners should exercise caution in interpreting results of individuals with limited English proficiency, if it is appropriate to interpret results at all. Furthermore, individuals with sensory disabilities (e.g., limited vision or hearing) may be at a disadvantage on traditional measures of intelligence, so practitioners must be aware of the best practices of assessment with these populations. Finally, individuals with diverse cultural or ethnic backgrounds may have experiential backgrounds that differ from that of the U.S. mainstream culture. As no intelligence test is completely free of cultural influences, practitioners should be aware of how to conduct assessment so as to minimize bias and maximize fairness for all individuals.

Tracy Paskiewicz

See also Ability Tests; Cattell–Horn–Carroll Theory of Intelligence; Confirmatory Factor Analysis; Developmental Disabilities; Ethical Issues in Testing; *g* Theory of Intelligence; Giftedness; Intellectual Disability and Postsecondary Education; Intelligence Quotient; Intelligence Tests; Norming; Norm-Referenced Interpretation; Percentile Rank; Psychometrics; Reliability; Standard Deviation; Standardized Scores; Standardized Tests; Test Battery; Test Security; Testlet Response Theory; Tests; Validity

Further Readings

Roid, G. (2003). Stanford-Binet Intelligence Scales: Fifth Edition. Chicago, IL: Riverside.

Roid, G. (2003). Stanford-Binet Intelligence Scales Technical Manual. Chicago, IL: Riverside.

Rebecca Jesson Rebecca Jesson Jesson, Rebecca

Stanines

Stanines

1609

1611

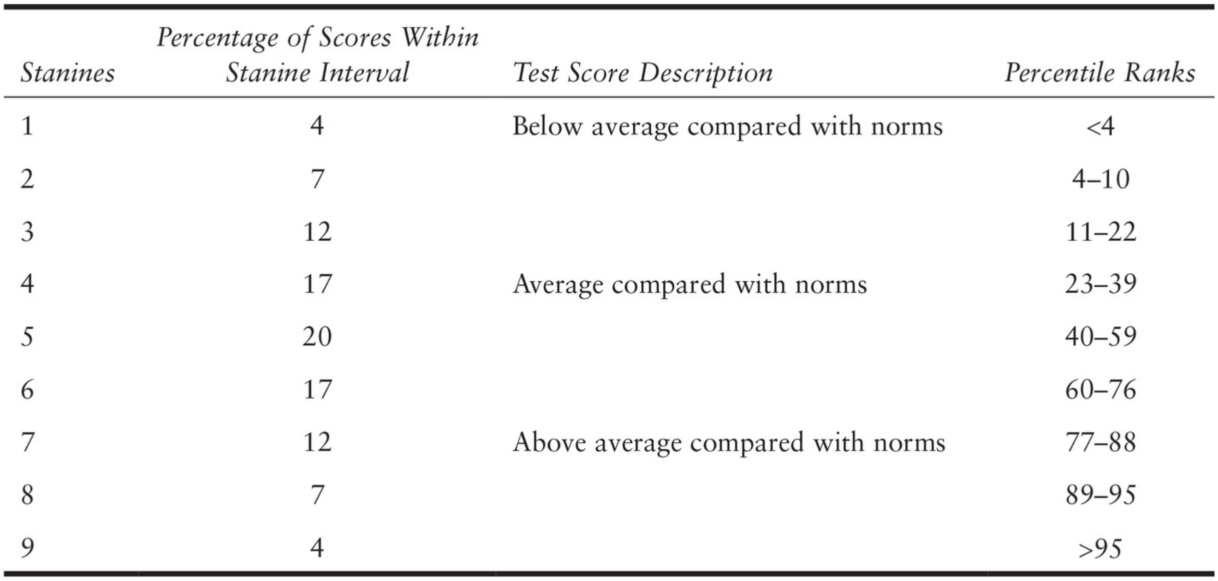
Stanines

A stanine is a type of standardized score, used to compare the position of a single score to a distribution of scores, on a scale of 1–9. Like other standardized scores, such as percentiles, *T* scores, and *z* scores, stanines are derived from a transformation of raw scores based on an assumption of normally distributed data. Stanine is an abbreviation of “standard nine” and is obtained by dividing a normal distribution into nine intervals, with a mean of five and a standard deviation of two. This scaling results in nine equal interval segments, each of which is half a standard deviation wide, except at each end of the distribution. The mean and median of the standard distribution is located at the center of stanine 5.

Stanines compare an individual test score with the comparison sample, or *norm*, which in education is often a state or nationwide sample of students at the same grade level, but might be an age equivalent or population sample. Stanines can be obtained by ranking all the scores in a distribution and assigning cut scores based on a normal distribution. Stanines 1–3 are commonly described as “below average,” 4–6 as “average,” and 7–9 as “above average” scores on the test. Table 1 indicates the percentage of scores at each table and their respective descriptions.

Stanines in Educational Settings

In education, stanines are commonly used to report to teachers and parents about a student’s relative standing compared with other students in the jurisdiction at the same grade level on a particular test. Because it gives a single digit score, which is always a whole number and is always positive, the general indication of relative standing is easily understood and is not likely to be confused with the child's actual score on the test, as can be the case with other types of standard score. As a score for an individual, stanines are also commonly used to select students for educational intervention or to assess a student's relative strengths across different tests. When tests are closely aligned with curriculum outcomes, stanines can be used to inform teachers and parents about school achievement.



Stanines	*Percentage of Scores Within Stanine Interval*	*Test Score Description*	*Percentile Ranks*
1	4	Below average compared with norms	<4
2	7		4–10
3	12		11–22
4	17	Average compared with norms	23–39
5	20		40–59
6	17		60–76
7	12	Above average compared with norms	77–88
8	7		89–95
9	4		>95

Limitations of Stanines

Like any system that divides scores into a limited number of equal intervals, stanines can be imprecise, in that the actual test scores or percentile ranks of two students with the same stanine may be more different from two students with different stanine scores. As an example, two students with a stanine level of 5 may be at the 40th and 59th percentiles, respectively, whereas a student with a stanine score of 4 may be only one percentile rank score different from a student at stanine 5. The differences between two students at either end of any given stanine band will be lost information. Similarly, over time, a change in only one raw score point may move a student from one stanine to another, which may give a false impression of relative improvement. For this reason, it is sometimes recommended that a shift of two stanines, the equivalent of one standard deviation, is needed to be assured of improvement. Stanines, like other standardized scores, are also unable to give information about learning needs, specific item knowledge, mastery, or learning progress in terms of curriculum outcomes.

Application of Stanines in Educational Research

In addition to their use for reporting individual scores, stanines have been used by educational researchers to report average levels of outcome variables of a population of interest. Treated as equal interval data, they have been used cross-sectionally, to identify the effect of specific predictor variables on the outcome of interest, reported in average stanine levels. They have also been used as standardized outcome variables to evaluate the success of an educational intervention for a specific group over time. Given that the stanine scale is assigned relative to the year group in a comparison or reference group, the stanine can be used across multiple grade levels and adjusts for normal progress. In such a case, the judgment of effectiveness for an intervention is made when a target group's average stanine levels are significantly increased, thus reflecting a shift in relative standing compared with jurisdiction equivalents. Stanines have also been used to compare proportions of groups of students within a distribution and to compare the distributions of achievement over time, for example by tracking changes in the percentages of students who achieve at different stanine bands in treatment and control groups before and after an educational intervention. These differences are commonly reported with relative caution using large data sets, bearing in mind the inherent imprecision of stanines as categories.

Rebecca Jesson

See also Norming; Norm-Referenced Interpretation; Percentile Rank; Quartile; Standardized Scores; Standardized Tests; Summative Assessment; Tests; Z Scores

Further Readings

Bjerkedal, T., Kristensen, P., Skjeret, G. A., & Brevik, J. I. (2007). Intelligence test scores and birth order among young Norwegian men (conscripts) analyzed within and between families. Intelligence, 35(5), 503–514. doi:10.1016/j.intell.2007.01.004

Burns, E. (1982). The use and interpretation of standardized grade equivalents. Journal of Learning Disabilities, 15(1), 17–18.

Durost, W. (1959). The use of local stanines in reporting test results in a large cosmopolitan school system. The Yearbook of the National Council on Measurements Used in Education, 16, 140–148.

Lai, M. K., McNaughton, S., Amituanai-Toloa, M., Turner, R., & Hsiao, S. (2009). Sustained acceleration of achievement in reading comprehension: The New Zealand experience. Reading Research Quarterly, 44(1), 30–56. doi:10.1598/RRQ.44.1.2

Mink, O. (1964). Using stanines to study IQ-achievement relationships. The School Counselor, 12(1), 43–45.

Morrison, D., Mantzicopoulos, P., & Stone, E. (1988). Screening for reading problems: The utility of SEARCH. Annals of Dyslexia, 38, 181–192. doi:10.1007/BF02648255

Lawrence C. Hamilton Lawrence C. Hamilton Hamilton, Lawrence C.

Stata

Stata

1611

1616

Stata

Stata is a full-featured statistical software package with capabilities ranging from basic through advanced data management, analytical graphics, tables, tests, and statistical modeling. A graphical user interface and consistent command syntax make Stata relatively easy to learn. Beyond the interactive documentation (with quick help files but also thousands of pdf pages of manuals providing formulas, explanations, examples, and references), users can consult a library of paper or ebooks for more information about particular topics such as categorical dependent variables, structural equation modeling (SEM), multilevel modeling, econometrics, graphics, survival analysis, or programming. Stata's built-in programming language, dedicated journal, and annual meetings support the development of original programs to accomplish specialized tasks or implement new statistical procedures. After providing the history of Stata, this entry reviews various features of this statistical software package.

History

Stata version 1.0, designed by Southern Californians William Gould and Sean Becketti, was released in 1985. Written in C for the first generation of MS-DOS computers, Stata 1.0 was principally a regression package with data management features. Basic graphing and programming features were introduced later that year with Version 1.3, and new statistics (analysis of variance, logit, and probit) with 1.5. Many versions later, data management, regression, graphing and programming continue to be core strengths of Stata, although now these are much expanded (e.g., including perhaps a hundred kinds of regression) and complemented by a full array of modern statistical tools. Each new version brought incremental extensions, punctuated by more radical jumps such as a graphical user interface of Version 8 in 2003. Functionality of programs written for earlier versions of Stata has been protected through a version control feature. Users who find a command line interface more efficient than point-and-click can accomplish most tasks using either or both.

Varied and sometimes personal accounts of Stata history, written from the viewpoints of participants, were published in a special issue of *The Stata Journal* for Stata’s 20th anniversary in 2005 and in a 30-year retrospective book edited by Enrique Pinzon in 2015. The 2005 articles include a detailed history and timeline of Stata development by Nicholas Cox and a conversation with William Gould, along with stories about the first out-of-house Stata book and launching *Stata Technical Bulletin*, precursor to *The Stata Journal*. The 2015 book includes reflections by Becketti and assessments by other authors on Stata’s contributions to epidemiology, biostatistics, public health, public policy, microeconomics, political science, and psychology.

Platforms

Stata is available for Windows, Macintosh, and Linux/Unix computers. Licenses are not platform-specific, so a user could, for instance, install copies under the same license on a Linux system at work, a Windows computer at home, and a Macintosh laptop for travel. Stata data sets, programs, and other data can be used interchangeably across these platforms.

Flavors, Versions, and Updates

Stata comes in a variety of flavors from Small Stata, inexpensive and meant for students, through progressively more capable IC, SE, and MP (multicore/multiprocessor) versions. Unlike modular statistical packages, all Stata flavors have the same complete set of features and documentation, so that even Small Stata can, for example, fit a mixed-effects generalized structural equation model. The main difference among flavors involves the size of programs and data sets that they can handle. Small Stata is limited to no more than 99 variables and 1,200 observations; Stata/MP can analyze 10–20 billion observations or more given current computers and should expand to 281 trillion as hardware advances.

Licenses could be time limited or perpetual, with a reduced charge if users choose to upgrade to the next major version. Minor upgrades within versions are free. Data saved under previous versions remain readable in newer versions. Old programs remain usable through version control statements that specify the appropriate version within any program.

Features

The full list of Stata features is quite long but can be explored on the Stata website. Beyond basic data management, tables, graphs, statistical tests, and other strengths worth mentioning include time series, generalized linear modeling, cluster analysis, factor analysis, power analysis, mixed-effects or multilevel modeling, survival analysis, multiple imputation of missing values, Bayesian statistics, survey analysis, and SEM, including the graphical SEM builder illustrated in Figure 1. Generalized SEM incorporates generalized linear and mixed-effects modeling within an SEM framework, so, for example, one can estimate structural models using multilevel data, including categorical or counted endogenous variables.

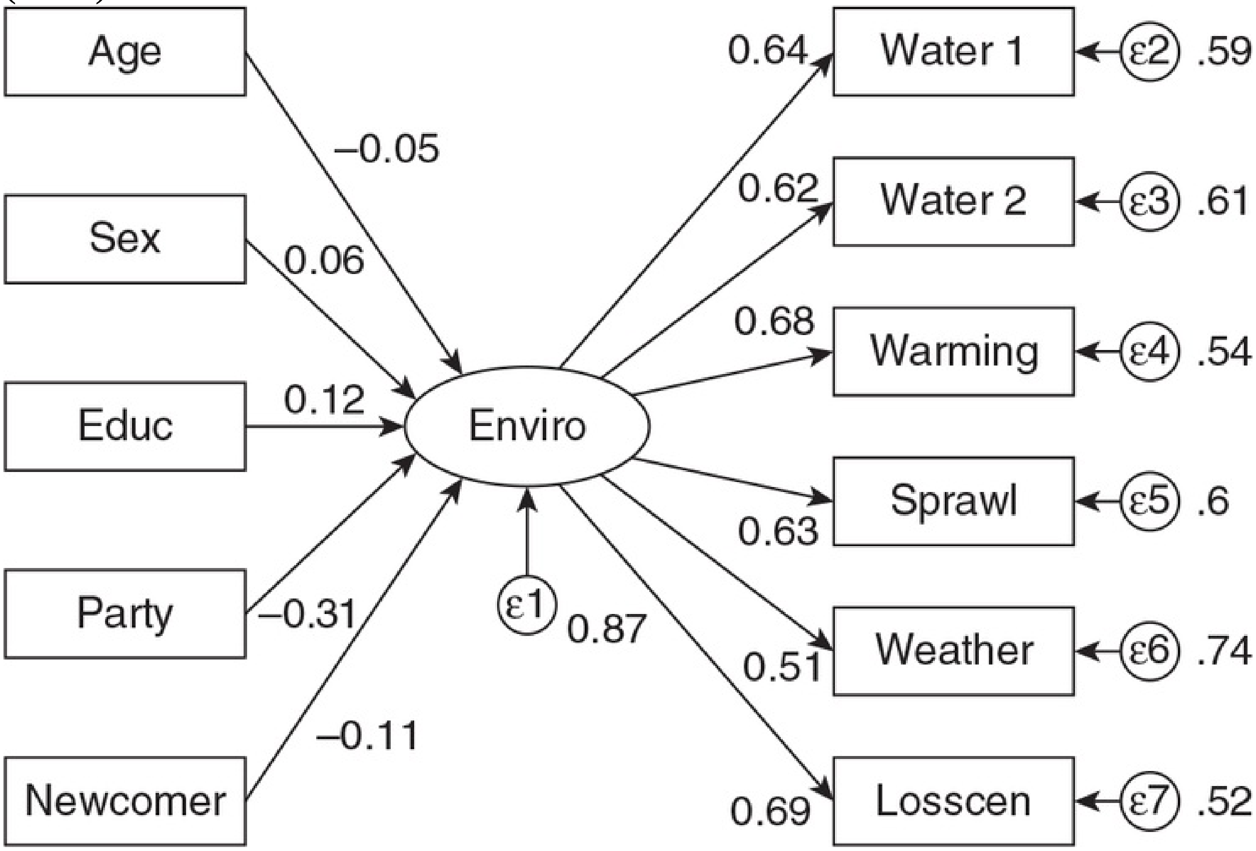
Specialized information about the data set structure supports some of these procedures. For example, data can be declared as time series, survey, panel, or survival-time structures. Declaration of data set structure makes other features available, such as lagged variables, smoothing, and modeling for times series; sampling weights for survey data analysis; or recognizing the parallel time series design of panel data.

Stata features are made easier to learn and use by broad consistencies in command syntax. For example, a basic regression command where variable *opinion* is regressed on four predictors could have the form:

```
regress opinion age gender education party
```

A robust, quantile, logit, multinomial probit, Poisson, or negative binomial regression command could look the same except with **rreg**, **qreg**, **logit**, **mprobit**, **poisson,** or **nbreg** in place of **regress** (and likewise with many other modeling procedures). Commands are easily restricted to subsets of the data by adding a qualifier such as:

Figure 1 Structural equation model in Stata's structural equation modeling (SEM) builder



```
regress opinion age gender education party if income < 60
```

A survey-weighted version could be specified with a **svy** prefix:

```
svy: regress opinion age gender education party
```

Interaction effects between *education* and *party*, both treated as continuous variables, can be specified without explicitly generating new variables by using Stata's factor variable notation:

```
svy: regress opinion age gender c.education##c.party
```

Alternatively, *education* and *party* might be viewed as indicator variables, and their interaction included without creating sets of {0,1} dichotomies and products:

```
svy: regress opinion age gender i.education##i.party
```

Factor variable notation, **svy** prefix, **if** qualifiers and many other options behave similarly with other procedures, where appropriate. Moreover, both the organization of tabular output and function of postestimation commands such as prediction and nonstandard hypothesis tests are similar using other modeling procedures, to the extent this is reasonable.

Although Stata covers methods that could be used in any field, its greatest popularity and development has been in social and biomedical fields: behavioral science, biostatistics, economics, education, epidemiology, finance and marketing, medicine, political science, public health, public policy, and sociology.

Resources and Support

All versions of Stata come with complete documentation, including interactive help files and extensive pdf manuals. Refereed *The Stata Journal* publishes new research and applications, providing a platform for adding new features to Stata that helps to maintain its state-of-the-art standing. The Stata Press publishes a library of Stata-specific books covering introductory through advanced reference needs (a complete list is available in the Bookstore at the Stata website). Other resources include the quarterly *Stata News*, Stata Blog, Netcourses, video tutorials, Statalist, and Stata conference and user groups meetings. Technical support is free to registered users.

Programmability

No programming is needed to use Stata, but built-in programming and matrix programming languages that make it highly extensible are available. Do-files, which are text files containing any sequence of Stata commands, streamline complex or repetitive tasks such as database management or graphing. Programming comes into play with automatic do-files, called ado-files, through

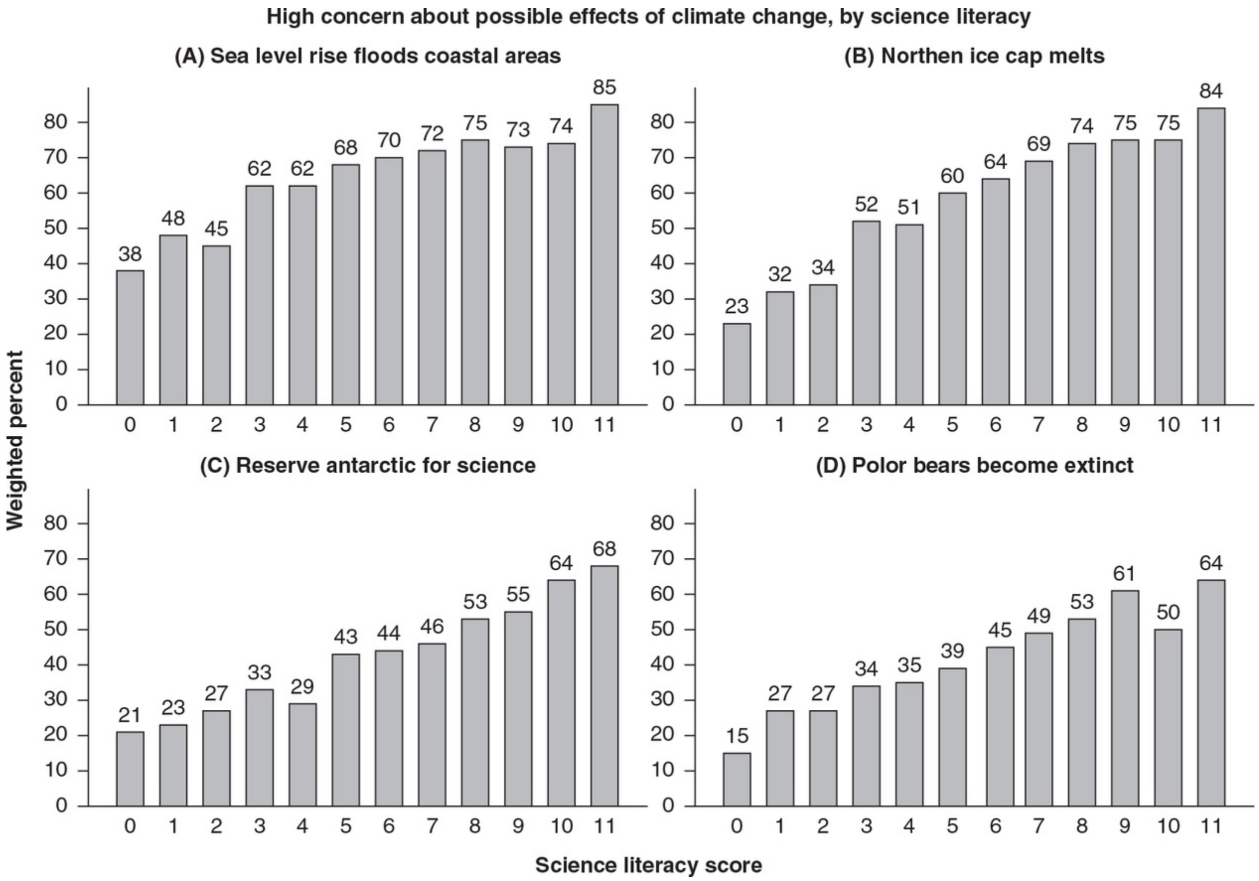
Programming comes into play with automatic do-files, called ado-files, through which users define new commands. Once an ado-file defines a new command that can be used transparently in the same manner as any other Stata command. Subroutines in ado-files might (optionally) take advantage of Mata, which itself is a full-blown matrix programming language.

Programmability has played a large part in keeping Stata up-to-date. New procedures recently described in journals, or desired by users, can be implemented as needed and disseminated through *The Stata Journal* and other sources. Much of Stata's code now resides in thousands of ado-files, which from a casual user's perspective appear seamlessly part of the package.

Graphics

Basic graph types include bar charts, histograms, box plots, line graphs, and scatterplots. There are dozens of other types, including specialized graphs for time series, survival analysis, and panel data. Beyond the defaults, each type offers dozens of options controlling such things as orientation, appearance, colors, and labeling for publication-quality images. Multiple graphs can be combined into one image if needed. For example, Figure 2 displays four bar charts, each graphing the weighted percentage of U.S. survey respondents who express high concern about a particular consequence of climate change. These percentages are broken down by respondent scores on a 12-point science literacy scale. Concern about each of the climate-change consequences rises with science literacy. Options control the labels to make this figure self-documenting.

Figure 2 Multiple bar graphs: concern about climate change effects, by science literacy score



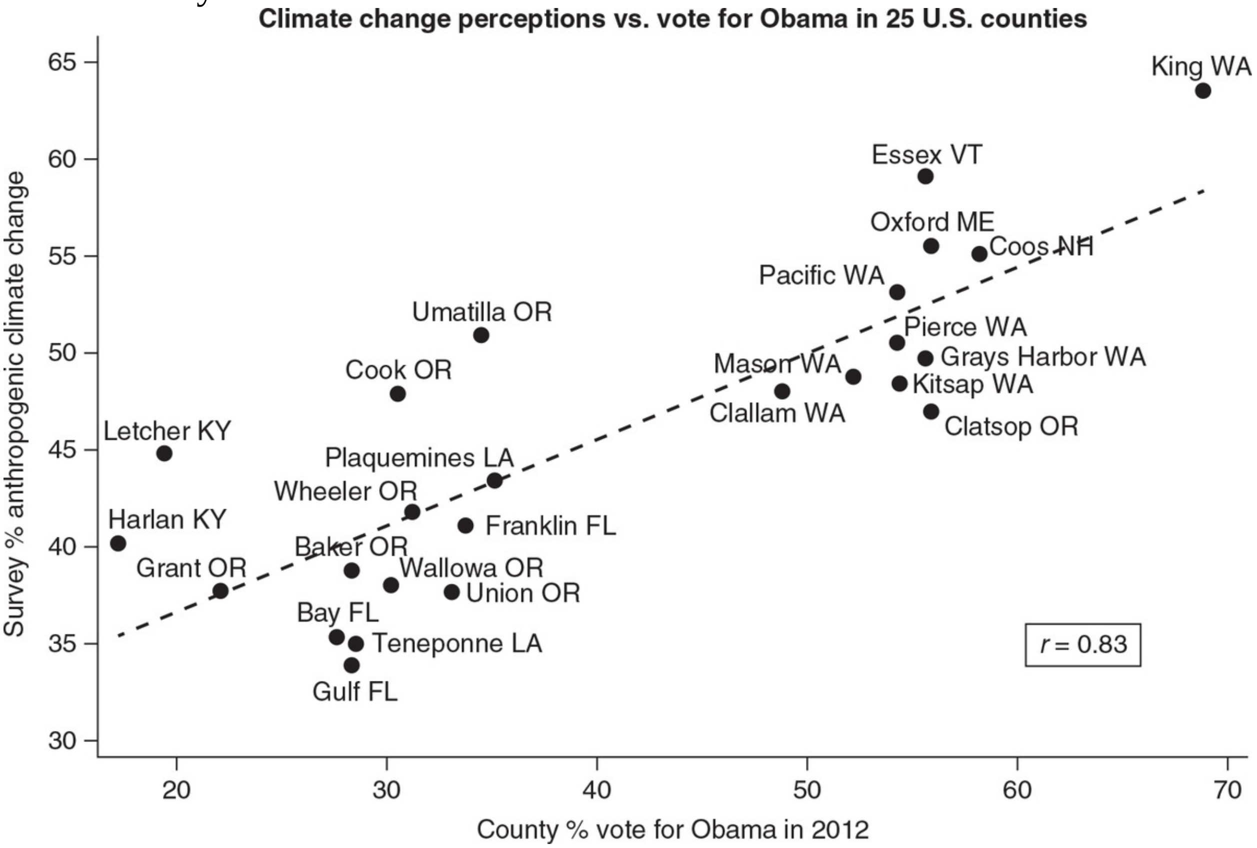
Source: Adapted from Hamilton, Cutler, & Schaefer (2012a).

Multiple scatterplots, line plots, and other two-variable graphs can be overlaid to see relationships. Figure 3, graphing survey-assessed, climate change beliefs against objective voting percentages in 25 U.S. counties, overlays two graphs: a scatterplot with labeled points and a regression line (from a study by Lawrence Hamilton and colleagues). A text option was used to place the correlation coefficient within this image as well.

One Stata graph type of particular relevance to social/behavioral and biomedical scientists is the adjusted margins plot, which allows visualizations showing interaction effects and nonlinear relationships (including their uncertainties) within complex models. Figure 4 depicts an interaction between education and political party affecting beliefs about the reality of anthropogenic climate change, from a logit regression model that includes age and gender as covariates.

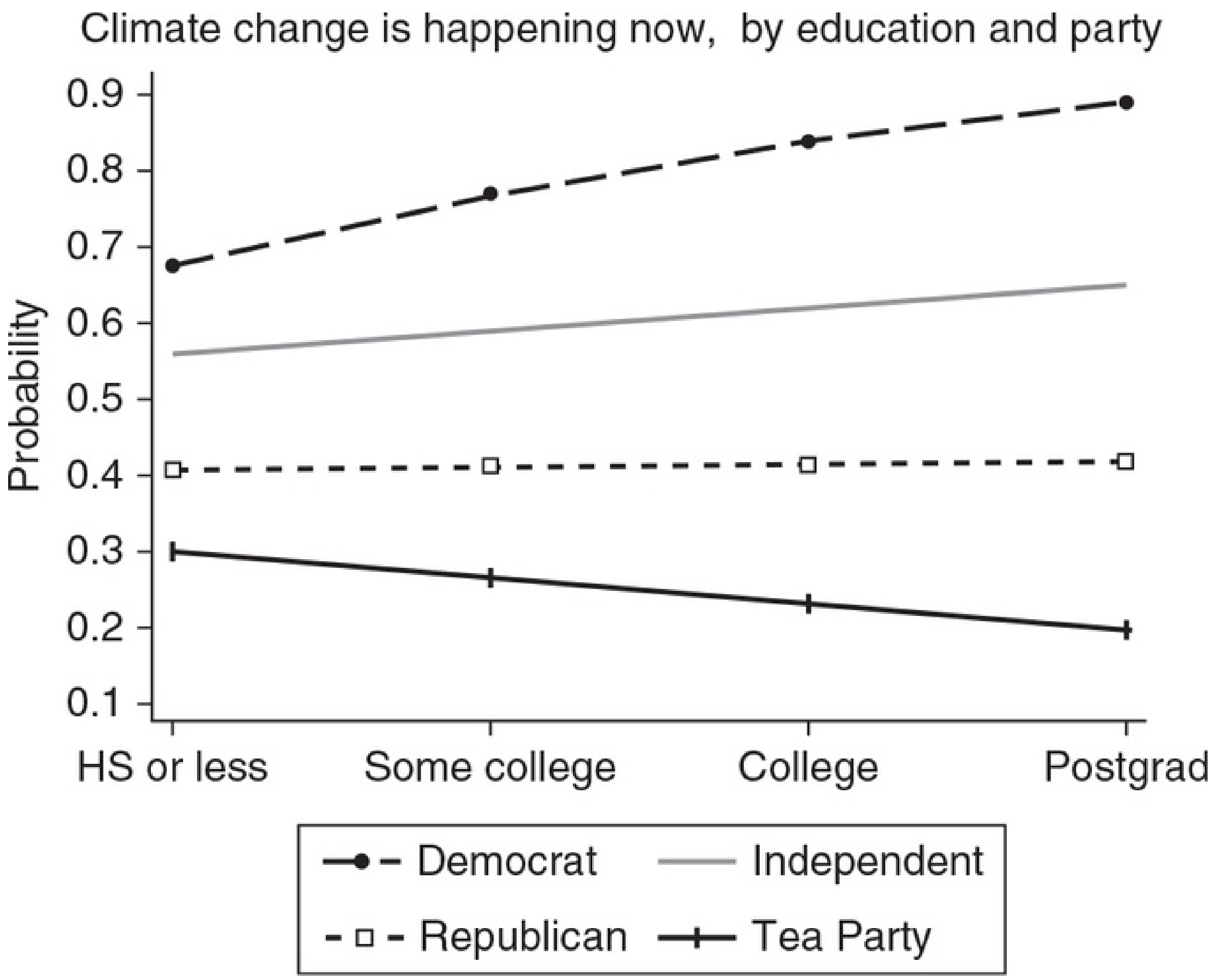
Figure 3 Overlaid scatterplot and regression line: climate change perceptions

versus county vote for Obama



Source: Hamilton, Wake, Hartter, Safford, & Puchlopek (2016).

Figure 4 Adjusted margins plot: climate change perception by education and party



Source: Hamilton & Saito (2015).

These illustrations represent only a small fraction of the available Stata graph types or the scope for creativity. Complex graphs typically are built in steps, elaborating on simple elements using do-files or the interactive Graph Editor. Saved graphs can be reedited later.

Lawrence C. Hamilton

See also Logistic Regression; R; SAS; SPSS

Further Readings

Cox, N. J. (2005). A brief history of Stata on its 20th anniversary. The Stata

Journal, 5(1), 2–18.

Hamilton, L. C. (2005). A short history of Statistics with Stata. The Stata Journal, 5(1), 34–36.

Hamilton, L. C. (2013). Statistics with Stata, Version 12. Belmont, CA: Cengage (seven previous editions, 1990–2009; Chinese translations of 2006 and 2009 editions by Zhigang Guo; Arabic translation of 2012 edition by Ramadan El-Faitouri).

Hamilton, L. C., Cutler, M. J., & Schaefer, A. (2012a). Public knowledge about polar regions increases while concerns remain unchanged. Durham, NH: Carsey Institute. Retrieved from http://scholars.unh.edu/carsey/157/

Hamilton, L. C., Cutler, M. J., & Schaefer, A. (2012b). Public knowledge and concern about polar-region warming. Polar Geography, 35(2), 155–168. doi:10.1080/1088937X.2012.684155

Hamilton, L. C., & Saito, K. (2015). A four-party view of U.S. environmental concern. Environmental Politics, 24(2), 212–227. doi:10.1080/09644016.2014.976485

Hamilton, L. C., Wake, C. P., Hartter, J., Safford, T. G., & Puchlopek, A. (2016). Flood realities, perceptions, and the depth of divisions on climate. Sociology, 50, 913–933. doi:10.1177/0038038516648547

Hilbe, J. M. (2005). The birth of the *Bulletin*. The Stata Journal, 5(1), 39–40.

Mitchell, M. N. (2012). A visual guide to Stata graphics (3rd ed.). College Station, TX: Stata Press.

Newton, H. J. (2005). A conversation with William Gould. The Stata Journal,

5(1), 19–31.

Pinzon, E. (Ed.). (2015). Thirty years with Stata: A retrospective. College Station, TX: Stata Press.

Websites

Stata: http://www.stata.com/

Roger Fischer Roger Fischer Fischer, Roger

State Standards

State standards

1616

1617

State Standards

Educational standards serve as a frame of reference for measuring academic achievement. State standards are the criteria used by individual states for measuring the quality of an educational skill or product. This entry discusses the origins of the standards movement in the United States and distinguishes between content standards and performance standards.

Origins of the Standards Movement

The origin of the standards movement can be traced to *A Nation at Risk*, a 1983 report compiled by the National Commission on Excellence in Education. This report painted a bleak picture of public education in the United States, indicating steadily decreasing scores on a variety of standardized tests, dropping matriculation rates in college, increasing enrollments in remedial mathematics courses, and an alarming rate of functional illiteracy.

In response to this report, numerous professional teachers’ associations compiled standards documents designed to target the reported deficiencies. In 1989, National Council of Teachers of Mathematics published the first such document titled *Curriculum and Evaluation Standards for School Mathematics*.

The federal government played a substantial role in shaping the standards movement as well. The National Education Goals Panel was formed in 1990 to monitor state progress toward attainment of educational goals. In 1994, the Goals 2000: Educate America Act was enacted with the goal of improving public education by promoting high achievement and equity for all students. The

No Child Left Behind Act, signed into law in 2002, tied federal funding for public schools to adequate yearly progress, a measurement of academic achievement indexed by performance on standardized tests. In 2010, the publication of the *Common Core State Standards* by the National Governors Association and the Council of Chief State School Officers brought increased attention to the need for clarity of standards and preparation of students for either postsecondary education or entry into the workforce.

Content Standards

Content standards are statements of desired student knowledge and ability. They incorporate detailed and explicit statements about particular content and are typically organized by either grade level or by course. Content standards not only describe specifically what students at a particular grade level or in a particular course should know and be able to do but also provide a framework for connecting skills, procedures, and concepts across grade levels and courses.

To make the connections between distinct grade levels or courses more explicit, educators rely on learning progressions. These narrative statements describe stages through which learners will likely progress as they master particular content standards. Additionally, learning progressions highlight the ways in which student understanding changes across grade levels or courses.

Performance Standards

Although content standards specify what students should know and be able to do, performance standards specify a level of mastery to be achieved in relation to the content standards. Each state is responsible for setting its own performance standards and determines what level a student must achieve to be deemed proficient.

Roger Fischer

See also Common Core State Standards; Curriculum; Formative Assessment; Goals and Objectives; High-Stakes Tests; Policy Evaluation; Problem Solving; Race to the Top

Further Readings

Jennings, J. F. (1998). Why national standards and tests? Politics and the quest for better schools. Thousand Oaks, CA: Sage.

Kosar, K. R. (2005). Failing grades: The federal politics of education standards. Boulder, CO: L. Rienner.

Richard D. Harvey Richard D. Harvey Harvey, Richard D.

Ana H. Kent Ana H. Kent Kent, Ana H.

Static Group Design Static group design

1617

Static Group Design

In 1963, Donald Campbell and Julian Stanley initially designated the static group design, one of the three preexperimental designs. Preexperimental designs are best thought of as premature designs in that they have serious flaws and therefore should be avoided. The static group design is also called the *posttest-only nonequivalent groups design*. Consistent with this name, participants are nonrandomly assigned into two groups (experimental vs. comparison). Participants then take a posttest after receiving treatment. Importantly, there is no pretest, so it is impossible to ascertain whether any group differences are due to the experimental manipulation/treatment or preexisting differences.

Preexperimental designs such as the static group design are typically used to explore a relationship prior to a true experiment, although they are occasionally used in applied research after more rigorous experimental research has established a causal relationship. Preexperimental designs are the simplest type of research design, and they use existing groups (i.e., no random assignment). In contrast to the other two preexperimental designs, the static group design introduces a comparison group. However, participants are not randomly assigned to the control and treatment groups. This flaw, along with the lack of a pretest, makes it difficult if not impossible to establish a causal relationship.

Among the list of factors that could jeopardize the internal validity of an experiment, the static group design appears to be vulnerable to three: selection, mortality, and maturation. With regard to selection, it is possible that there might be systemic differences between the groups prior to treatment. Mortality refers to the notion that the posttest might reflect differences in the dropout rate between the experimental and control groups rather than the treatment. Somewhat similar the experimental and control groups rather than the treatment. Somewhat similar to mortality, maturation points to the likelihood that changes in the internal states of the participants might account for differences in the posttest rather than the treatment. These threats to validity are the static group design’s biggest disadvantages and the reason this design is primarily used for exploratory purposes.

The static group design is sometimes used out of necessity rather than exploratory or negligent science. Occasionally, ethical considerations would prevent researchers from imposing a treatment or variable of interest upon participants. Consider, for example, a researcher who would like to study the effects of in utero maternal drug use on newborn infants. The researcher cannot ethically randomly assign mothers to drug use and nondrug use conditions nor can the researcher give a pretest to the infants. The static group design becomes the flawed, but only option. Ideally, the comparison group should be matched as closely as possible to the experimental group.

Although the static group design has some serious limitations (i.e., threats to internal validity), it can be a cost-effective, ethical, and/or exploratory way to obtain prima facie evidence of a treatment effect. Notwithstanding, any conclusions drawn from a study that has utilized this design must be tentative and interpreted with caution. For this reason, it is best to follow up any study that has used this design with a replication study that utilizes a more rigorous true experimental design.

Richard D. Harvey and Ana H. Kent

See also Nonexperimental Designs; Posttest-Only Control Group Design; Preexperimental Designs; Pretest–Posttest Designs; Threats to Research Validity; Validity

Further Readings

Campbell, D., & Stanley, J. (1963). Experimental and quasiexperimental designs for research. Washington, DC: American Educational Research Association.

Craighead, W. E., & Nemeroff, C. B. (2002). The Corsini encyclopedia of psychology and behavioral science (Vol. 4). Wiley.

Thyer, B. A. (2012). Quasiexperimental research designs. Oxford University Press.

Kerry Lee Kerry Lee Lee, Kerry

Bruce Granshaw Bruce Granshaw Granshaw, Bruce

STEM Education

STEM education

1618

1620

STEM Education

The acronym *STEM* refers to science, technology, engineering, and mathematics. Although each individual subject has its own extensive history, the notion of STEM education is relatively new. The commonalities and overlap of subject matter has meant that delineation of each subject area is very difficult. This has led to a combining of subjects such as: S&T (science and technology), STS (science, technology, and society), SMET (science, mathematics, engineering, and technology), TAS (technology as applied science), SET (science, engineering, and technology), MST (mathematics, science, and technology), and STEAM (science, technology, engineering, art, and mathematics). This entry further defines the term *STEM* and discusses the value of STEM subjects to society, interest levels and gender disparities in STEM, and research on STEM education and the choice to pursue STEM careers.

Although many countries utilize the STEM acronym, there is little consensus about its meaning. When people refer to the *multidisciplinary* nature of STEM, they are generally focusing on the four different subject disciplines working independently. However, the *interdisciplinary* nature of STEM refers to the integration of knowledge and modes of thinking drawn from these four disciplines.

Science, mathematics, and engineering are not new subjects, but technology education is. There has been, and continues to be, disagreement and confusion over what technology education actually entails. This range of views embraces concepts such as design and innovation, product design, intervention by design, and the development of technological literacy. However, it is also understood by some to mean the study of technical and vocational knowledge and skills, and to others, the study and utilization of computers, computerized equipment, and a wide range of digital tools.

To eliminate confusion, this entry uses the broader and more holistic interpretation of the term *technology education* where the focus is not on gaining or applying technical skills and computing literacy, but rather on thinking creatively to solve design problems that may or may not require these technical skills. The term *STEM education* will refer to teaching in an interdisciplinary holistic manner, whereas the term *STEM subjects* will refer to the individual subjects of science, technology, engineering, and mathematics.

Value of STEM

Quality STEM education has the potential to enhance success for students in the 21st century. It can prepare students for jobs that have yet to be conceptualized by providing life skills such as teamwork, problem solving, lateral thinking, creativity, resilience, and critical thinking. For example, a STEM course design might require students to develop and create a solution to a technological problem drawing upon and integrating knowledge from all four STEM subjects.

Advisory bodies internationally, including those in Australia and the United States, have highlighted the need to explicitly teach generic skills such as problem solving and critical thinking, skills that characterize STEM education programs. The U.S. federal government has led the way with this thinking and has made STEM education a funding priority. The United Kingdom, which historically has a strong record of its scientists winning Nobel prizes, has recognized a perceived gap between scientific advancement and the development of technological innovations and products. Substantial economic advantage may be gained by placing a focus on STEM subjects and STEM education as a means of closing this gap. STEM education also can equip students for a rapidly changing technological world.

Interest Levels in STEM Subjects

Employment projections for the United States highlight an increasing need for additional employees in occupations that draw upon STEM education. However, students in the United States, United Kingdom, and in some Asian countries often do not regard STEM subjects as attractive options when selecting courses. Even though mastery of STEM subjects has been associated with college success and retention, many high school students do not opt for these subjects. This has an obvious impact on their career path.

Although some students avoid STEM subjects, others are precluded from them for a variety of reasons. For example, the socioeconomic status of a student has been shown to be a significant predictor of whether the student will undertake advanced mathematics and science courses at high school and then pursue a STEM major at university. There is also evidence that students' progress is inhibited by the slowness of educational bodies in advancing curriculum and pedagogical practices that support integrative teaching and learning in STEM areas.

Gender Disparities in STEM Education

Historically in Western education, girls have tended to outperform boys in verbal language–based curricula, and boys have outperformed girls in many of the STEM-based curriculum areas. In the optional areas of the school curriculum, boys are more likely to engage with certain STEM subjects, and this predictably follows through to tertiary study. In the United States, women earn about half of the bachelor's degrees awarded in science and engineering, but the percentage varies greatly by discipline. Women outnumber men in bachelor's degrees awarded in psychology and the biological sciences, whereas men outnumber women in bachelor's degrees awarded in computer science and engineering.

The gender gap in STEM has been an issue in many education studies and government reports. Some have argued that males are innately superior in spatial and numerical abilities (i.e., a cognitive difference exists) and thus are likely better suited to STEM fields than women, although researchers have found only small overall differences in math performance at the elementary and secondary school levels. Yet, the issue of the gender gap in STEM fields remains hotly debated, with researchers attributing the gap to a variety of biological, psychological, sociocultural, and contextual influences that impact on women's interests, self-efficacy, and other motivation-related beliefs.

A concerted effort from schools, universities, and the community can play an important role in preparing women for careers in STEM fields and in providing a means to support women toward gender parity. Some specific strategies to achieve this may include inspiring girls to take higher level STEM subjects at school and encouraging them to pursue a STEM career; supporting the development of spatial skills in girls; exposing girls to successful female, as well as male, role models in STEM careers and achievements; and promoting attitudes that encourage confidence in addressing negative stereotyping about female competence and success. A further strategy is to ensure that pay equity is achieved between men and women in occupations that draw upon STEM-related qualifications and experience.

Research on STEM Education and Careers

Given the importance placed upon STEM subjects, education, and career pathways in the United States and other countries, it comes as no surprise that the number of research publications focusing on STEM has increased exponentially in recent years. This is particularly so in the case of technology, where publications include those focusing on technology education and on educational technology, or the digital tools used for learning.

Teaching and learning in the STEM subjects and information and communication technology are leading research areas. Research into gender issues and learning strategies has also been popular, but there has been less emphasis on research on innovations and problem solving, which reflect creativity and higher order thinking. Research in the field of STEM education—how best to encourage and promote disciplinary knowledge and the generic skills that enable integrative thinking and practice—is essential for preparing students for successful participation in employment and society more widely.

In order to build a common understanding of STEM and factors that influence individual educational and career choices, the United States government has made a significant investment in STEM education. This includes the promotion of STEM curricula, teaching and learning, research into gender disparities in the uptake of STEM education, and the collection of longitudinal data on the impact of STEM curricula in schools on later vocational participation and success.

How and why women and men behave as they do regarding STEM-related subjects or careers cannot be simply explained. There is no straightforward way to form an adequate conception of a particular group and the dynamics of its behavior. However, education emerges as one key intervention for gender inequality.

In relation to future research, studies need to draw together both quantitative indicators of participation and achievement of policy targets and qualitative data drawn from in-depth investigations such as that found in case study research. Qualitative research has the capacity to enrich understanding both of students' experience of STEM education and the factors that promote participation and achievement.

Kerry Lee and Bruce Granshaw

See also Common Core State Standards; Creativity; Curriculum; Gender and Testing; Organisation for Economic Co-operation and Development

Further Readings

Amirshokoohi, A. (2016). Impact of STS issue oriented instruction on pre-service elementary teachers' views and perceptions of science, technology, and society. International Journal of Environmental and Science Education, 11, 359–387. doi:10.12973/ijese.2016.324a

Asunda, P. A. (2012). Standards for technological literacy and STEM education delivery through career and technical education programs. Journal of Technology Education, 23, 44–60.

Breiner, J., Harkness, S., Johnson, C., & Koehler, C. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. School Science and Mathematics, 112, 3–11. doi:10.1111/j.1949–8594.2011.00109.x

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2016). Why are some STEM fields more gender balanced than others? Psychological Bulletin. doi:10.1037/bul0000052

Constantinou, C., Hadjilouca, R., & Papadouris, N. (2010). Students' epistemological awareness concerning the distinction between science and technology. International Journal of Science Education, 32, 143–172.

doi:10.1080/09500690903229296

DeJarnette, N. (2012). America’s children: Providing early exposure to STEM (science, technology, engineering and math) initiatives. Education, 133, 77–84.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. Psychological Science in the Public Interest, 8, 1–51.

Hill, C., Corbett, C., & St Rose, A. (2010). Why so few? Women in science, technology, engineering, and mathematics. Washington, DC: AAUW.

Jayarajah, K., Saat, R. M., & Rauf, R. A. A. (2014). A review of science, technology, engineering & mathematics (STEM) education research from 1999–2013: A Malaysian perspective. Eurasia Journal of Mathematics, Science & Technology Education, 10, 155–163. doi:10.12973/eurasia.2014.1072a

Knezek, G., Christensen, R., & Tyler-Wood, T. (2011). Contrasting perceptions of STEM content and careers. Contemporary Issues in Technology and Teacher Education, 11, 92–117.

National Science Foundation, National Center for Science and Engineering Statistics (2017). Women, Minorities, and Persons With Disabilities in Science and Engineering: 2017. Special Report NSF 17–310. Arlington, VA. Retrieved from www.nsf.gov/statistics/wmpd/.

Sanders, M. (2008). STEM, STEM education, STEMmania. The Technology Teacher (December/January), 20–26.

West, M. (2012). STEM education and the workplace. Office of the Chief Scientist, Australian Government. Retrieved from

http://www.chiefscientist.gov.au/wp-content/uploads/OPS4-STEMEducationAndTheWorkplace-web.pdf

Dale E. Berger Dale E. Berger Berger, Dale E.

Stepwise Regression

Stepwise regression

1620

1623

Stepwise Regression

Stepwise regression is an automatic computational procedure that attempts to find the “best” multiple regression model using only statistically significant predictors from a larger set of potential predictive variables. The regression model describes the relationship between a dependent (outcome) variable (*Y*) and two or more independent (predictor or explanatory) variables (X_j), using a model where . Computer programs are used to find the B_j weight for each X_j variable so as to minimize the sum of the squared error (*e*) for cases used to generate the model. The unique contribution of each X_j variable can be tested with a *t* test and associated *p* value, using the null hypothesis that $B_j = 0$ in the population. Stepwise regression attempts to find the best regression model by adding or deleting variables one at a time based solely on the *p* values for the individual predictors at each step. Although stepwise regression can be useful if it is applied and interpreted appropriately, it has been heavily criticized because it is often misused and misinterpreted. This entry presents the stepwise estimation procedures, reviews the common criticisms, and provides a recommended alternative.

Stepwise Estimation Procedures

There are two main stepwise approaches:

1. *Forward selection* begins by selecting from a pool of potential predictors the single predictor variable with the smallest statistically significant *p* value (if any), and then on each successive step selecting the individual

variable with the smallest statistically significant *p* value for its added contribution to the model, until no remaining potential predictor would make a statistically significant contribution.

2. *Backward elimination begins* with all potential predictors in the model and deletes variables one by one beginning with the variable with the largest nonstatistically significant *p* value, until every variable remaining in the model (if any) is statistically significant.

Stepwise regression combines the two approaches by testing every variable at each step for both inclusion and exclusion, with criteria set to allow variables already in the model to be eliminated more easily than new variables to be added. For example, one might require $p < .05$ for a variable to be added, while a variable already in the model would be eliminated if $p > .10$ at any step. As an option, specific variables may be forced into the model or be given priority consideration before other variables are considered for stepwise inclusion. A related method is *all possible subsets regression*, whereby a computer program tests all possible combinations of potential predictors and identifies the best model based on some criterion.

All of these methods of allowing a computer program to select the best model according to some statistical criteria have come under severe criticism from statisticians. Common criticisms of these procedures are that the R^2 values are biased to be too large, incorrect tests of statistical significance are often used, models are unstable when potential predictors are correlated, and final models may not be the most useful for practical or theoretical purposes.

Criticisms of Stepwise Regression

Criticism 1: The Observed Multiple R Is Inflated

Selecting the best predictors from a larger set of potential predictors capitalizes on chance variation in the observed data set, and the model is unlikely to do as well when applied to new data sets. Statistical programs offer “adjusted” or “shrunken” R^2 as a better estimate of the population R^2. Consider a situation with $N = 40$ cases, where 18 potential predictors were considered, and the final stepwise regression model has three predictors and $R^2 = .500$. Adjusted R^2 would be reported as .458 based on 40 cases with three predictors. But this adjustment does not take into account that 18 variables were considered. If 18 predictors were used to generate an R^2 value of .500, the adjusted R^2 would be only .071! Thus, even the reported adjusted R^2 is inflated. Stepwise regression capitalizes on random variations in the sample data, so the model is unlikely to fit a new sample as well as it fits the sample that was used to generate the model. Inflation of R^2 is greater with smaller samples and more variables.

Criticism 2: The Tests of Statistical Significance Are Incorrect

Common tests of statistical significance for a stepwise model ignore the fact that the variables in the model were selected from a larger set of potential predictors. A regression model with three predictors is commonly tested with an *F* test that has $df = 3$ in the numerator. This would be appropriate if only three predictors were considered and used. However, as noted in the first criticism, selecting the best predictors from a larger set of potential predictors capitalizes on chance variation in the sample. A valid test of significance must consider the number of potential predictors as well as the number of variables that were included in the model. A conservative approach would be to test the observed R^2 value as if it was the result of a model that included all potential predictors. For the example with 40 cases and 18 potential predictors with $R^2 = .500$, this conservative test gives $F(18, 21) = 1.17$, $p = .364$. However, this test is too conservative; if all 18 variables had been included in the model, the R^2 value likely would have been greater than .500, resulting in a larger *F* value and smaller *p* value.

Criticism 3: The Model Is Unstable, Especially If Potential Predictors Are Highly Correlated

In practice, predictor variables are expected to be related, perhaps even with large correlations. An implication is that in a new sample, different variables may emerge as the best predictors. When one member of a highly correlated pair of variables is entered into the model, the added contribution of its pair is greatly reduced. Naive users may erroneously conclude that the predictors in the model are much more important than variables not in the model. Especially with large samples, stepwise regression may lead to over fitting because even trivial effects may attain statistical significance. Larger samples tend to produce stepwise models with more variables.

models with more variables.

Criticism 4: The Model Produced by Stepwise Regression May Not Be the Most Practical

The automatic stepwise procedure does not consider practical issues inherent in the context of the study, such as cost or convenience of collecting various measures. For example, if two potential predictors are nearly equivalent in their contribution to the model, the practitioner may prefer to use a less expensive variable or a more readily available variable that works almost as well.

Criticism 5: Stepwise Regression Is Not Appropriate for Testing Theories

Theoretical considerations commonly dictate a logical order for entering variables into a model. For example, one may wish to control for background variables such as age, previous experience, or base rates of performance prior to assessing the added impact of an intervention. The researcher can use knowledge about the context and meaning of the variables to determine a meaningful order, while stepwise procedures ignore the meaning of variables.

Recommendations

Many statisticians caution strongly against using stepwise regression, especially for developing a theoretical model. A recommended alternative to stepwise regression is hierarchical regression. With hierarchical modeling, the order of entry of predictor variables is determined by the researcher prior to the analysis. Variables are entered into the model in an order that is meaningful either practically or theoretically. Because only a limited set of specific *a priori* tests are considered, the tests of statistical significance are correct.

With regression, as with other statistical procedures, it is important to examine data carefully to assure that assumptions are met (e.g., appropriate sampling, linearity, reasonably normal distributions, and homoscedasticity of errors). Turning the analysis over to a stepwise computer program does not avoid potentially serious problems caused by violations of assumptions.

It is important for the researcher to keep in mind the distinction betw

It is important for the researcher to keep in mind the distinction between hypothesis generating and hypothesis testing. Stepwise regression can be used as a hypothesis generating tool, giving an indication of how many variables may be useful, and identifying variables that are strong candidates for prediction models. However, it is essential to establish generalizability of findings as with replication and cross validation with a different sample.

Dale E. Berger

See also Effect Size; Multiple Linear Regression; Replication; *t* Tests

Further Readings

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45, 265–282. doi:10.1111/j.2044–8317.1992.tb00992.x

Henderson, H. V., & Velleman, P. F. (1981). Building regression models interactively. Biometrics, 37, 391–411.

Pedhazur, E. J. (1997). Multiple regression in behavioral research (3rd ed.). Orlando, FL: Harcourt Brace.

Pituch, K. A., & Stevens, J. P. (2016). Applied multivariate statistics for the social sciences (6th ed.). New York, NY: Routledge.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson, (Ed.), Advances in social science methodology (Vol. 1, pp. 99–105). Greenwich, CT: JAI Press.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement, 55, 525–534. Retrieved from https://doi.org/10.1177/0013164495055004001

Wilkinson, L. (1979). Tests of significance in stepwise regression. Psychological Bulletin, 86, 168–174.

Maria DeYoreo Maria DeYoreo DeYoreo, Maria

Stratified Random Sampling Stratified random sampling

Stratified Random Sampling

Stratified random sampling is a method for sampling from a population whereby the population is divided into subgroups and units are randomly selected from the subgroups. Stratification of target populations is extremely common in survey sampling. Stratified sampling techniques are often used when designing business, government, and social science surveys; therefore, it is important for researchers to understand how to design and analyze stratified samples. To obtain a stratified sample, members of a population are first divided into nonoverlapping subgroups of units called *strata*. The strata must be mutually exclusive and exhaustive, and there is an assumption of homogeneity within the strata. Following stratification, a sample is selected from each stratum, often through simple random sampling.

Determining the Strata and Sample Sizes

Although target populations are almost always heterogeneous, strata are assumed to be internally homogeneous. Survey practitioners should define strata such that the survey variables or measurements of interest have small variation compared to the variation across the population as a whole. In addition, the subpopulations defining the strata may be of interest in themselves. For example, states or regions are often considered important output categories in household surveys. Common stratification variables for surveys of individuals include age, gender, socioeconomic status, and educational attainment.

Once the researcher has divided the population into strata, the researcher must select units from each stratum. Samples are often selected through simple random sampling, a method for sampling in which each unit has the same probability of being chosen, and every possible subset of *k* units has the same probability of being selected.

There are a couple of ways to determine the strata sample sizes. Assume the population is of size *N*, the size of stratum *h* is N_h, for $h = 1, \ldots, H$, and the desired sample size is *n*. The sample sizes may be allocated proportionally, such that the fraction of units sampled from each stratum is proportional to the size of the stratum in the population. For instance, if the strata are defined by states and the population consists of all individuals in the United States, then the fraction of units sampled from each stratum is proportional to the population of each state relative to the entire U.S. population. Under proportional allocation, , where is the sample size of stratum *h*.

Alternatively, one can use optimal allocation, in which the size of each stratum is proportional to the standard deviation of the variable of interest. Larger samples are taken in the strata with the greatest variability to generate the smallest possible sampling variance. Under optimal allocation, for $h = 1, \ldots, H$. There are also variations on this strategy that take into account the cost of sampling from each stratum.

Stratified sampling ensures that at least one observation is picked from each stratum, even if the proportion of population units in a particular stratum is close to 0. The statistical properties of the population may not be preserved if there are strata with very few observations; hence, this should be avoided. Generally, it is recommended to use 5 to 10 strata; however, the main factor that limits the number of strata is the size of the population and strata. If there are 50 states each containing more than 100,000 units, viewing the states as the strata and sampling 5% of the units from each stratum is appropriate.

Advantages and Limitations

One reason to use a stratified sample is simply that parameters of each stratum may be of interest. For instance, the average number of children belonging to married couples by state may be a quantity of interest; therefore, it makes sense to stratify on state.

Relative to simple random sampling, stratified procedures can be viewed as superior because they improve the potential for the units to be more evenly spread over the population. In political surveys, it is important to include respondents that reflect the diversity of population. Such surveys seek to include participants of all races, religions, and from all regions or states. If a simple random sample is used, it is possible that very few units of a particular race or religion will be selected. The stratified random sample improves the representation of particular strata within the population, while ensuring that no strata are **overrepresented**.

Stratified sampling produces estimators that are more efficient than those from simple random sampling because of the homogeneity of the strata relative to the overall population. When the standard deviation of some variable is smaller within strata than when based on the entire population, stratification gives smaller error in estimation. Because a stratified random sample can have more precision than a simple random sample, it may require a smaller sample, saving time and money.

To perform a stratified sample, it must be possible to divide the population up into strata and to be able to list all members of the population. The former requires partitioning the population into disjoint and exhaustive subgroups, which may not be possible.

Poststratification

Sometimes it is not possible to place units into strata until the units have been sampled. For instance, in a telephone interview, respondents cannot be placed into gender or age strata until they have been contacted. Poststratification involves forming strata after selecting a sample and is often used when one has obtained a simple random sample that is not representative of the population. Assume one is interested in differences in average income in some city and conducts a telephone survey of 1,000 people, reaching 700 males and 300 females. Because income is likely to differ by gender, the estimate of mean income is likely skewed toward that of males. The poststratification mean can be obtained by weighting the average incomes for males and females by the proportion of males and females in the population.

Maria DeYoreo

See also Sample Size; Simple Random Sampling; Standard Deviation; Survey Methods; Surveys

Further Readings

Chambers, R., & Clark, R. (2012). An introduction to model-based survey sampling with applications. New York, NY: Oxford University Press.

Cochran, W.G. (1977). Sampling techniques (3rd ed.). New York, NY: Wiley.

Lohr, S. L. (2010). Sampling: Design and analysis (2nd ed.). Boston, MA: Cengage Learning.

Sardal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. New York, NY: Springer.

Thompson, S. K. (2012). Sampling (3rd ed.). New Jersey, NY: Wiley.

Cecilia Ma Cecilia Ma Ma, Cecilia

Daniel Tan-lei Shek Daniel Tan-lei Shek Shek, Daniel Tan-lei

Structural Equation Modeling Structural equation modeling

1624

1629

Structural Equation Modeling

Structural equation modeling (SEM) has received growing attention by researchers in social sciences and education. There are several distinct features of SEM. First, the SEM can estimate the complex relationships between variables. Second, it allows researchers to test hypothesized models based on theory and prior empirical findings. Third, unlike traditional multivariate statistical methods, such as multiple regression analyses, multivariate analysis of variance, and correlation, SEM takes measurement error into account, thereby giving unbiased parameter estimates. Last, it provides multiple fit indices of model fit and suggests how a model can be modified.

Given the increased popularity of SEM, many software packages, including LISREL, EQS, Amos, SAS, and Mplus, are available to conduct the related analyses. All are equation based, except Amos, which is commonly used in the graphical interface-based mode. These programs conduct SEM analyses differently; for example, they vary in their methods for handling missing and screening data, generating the program’s syntax and diagram, and fit indices. Beginners are recommended to read the software manuals to assist them in selecting the SEM program that best meets their research needs.

This entry begins by presenting the basic concepts of SEM. Then, the entry details the steps in conducting SEM. Last, the entry discusses common SEM models used in educational research.

Basic Concepts of SEM

In SEM, latent variables (also known as constructs or unobserved variables) refer to variables that cannot be directly measured, such as personality, motivation, or self-esteem. Observed variables (also known as measured or manifest variables) serve as indicators of the underlying latent variables. In addition, exogenous and endogenous variables are handled by SEM. Although exogenous variables (similar to independent variables) are not influenced by other variables, endogenous variables (similar to dependent variables) are predicted by other variables in the model.

Steps in SEM

Model Specification

Based on a theory and/or prior research, researchers specify the parameters and relationships among variables in a hypothesized model. In SEM, researchers hypothesize the relationships between the latent variable and the observed variables. For example, as is shown in Figure 1, parental control as a latent variable (represented as oval shape) is composed of three observed variables (i.e., "parental control too harsh," "parents force children to do things," and "parents scold and beat children"). Similarly, parental concern is measured by three observed variables (i.e., "parents love their children," "parents take care of their children," and "parents do not care about their children"). The six observed variables were loaded on two latent variables, which were correlated with each other (i.e., covariance).

Figure 1 A hypothesized two-factor model of confirmatory factor analysis

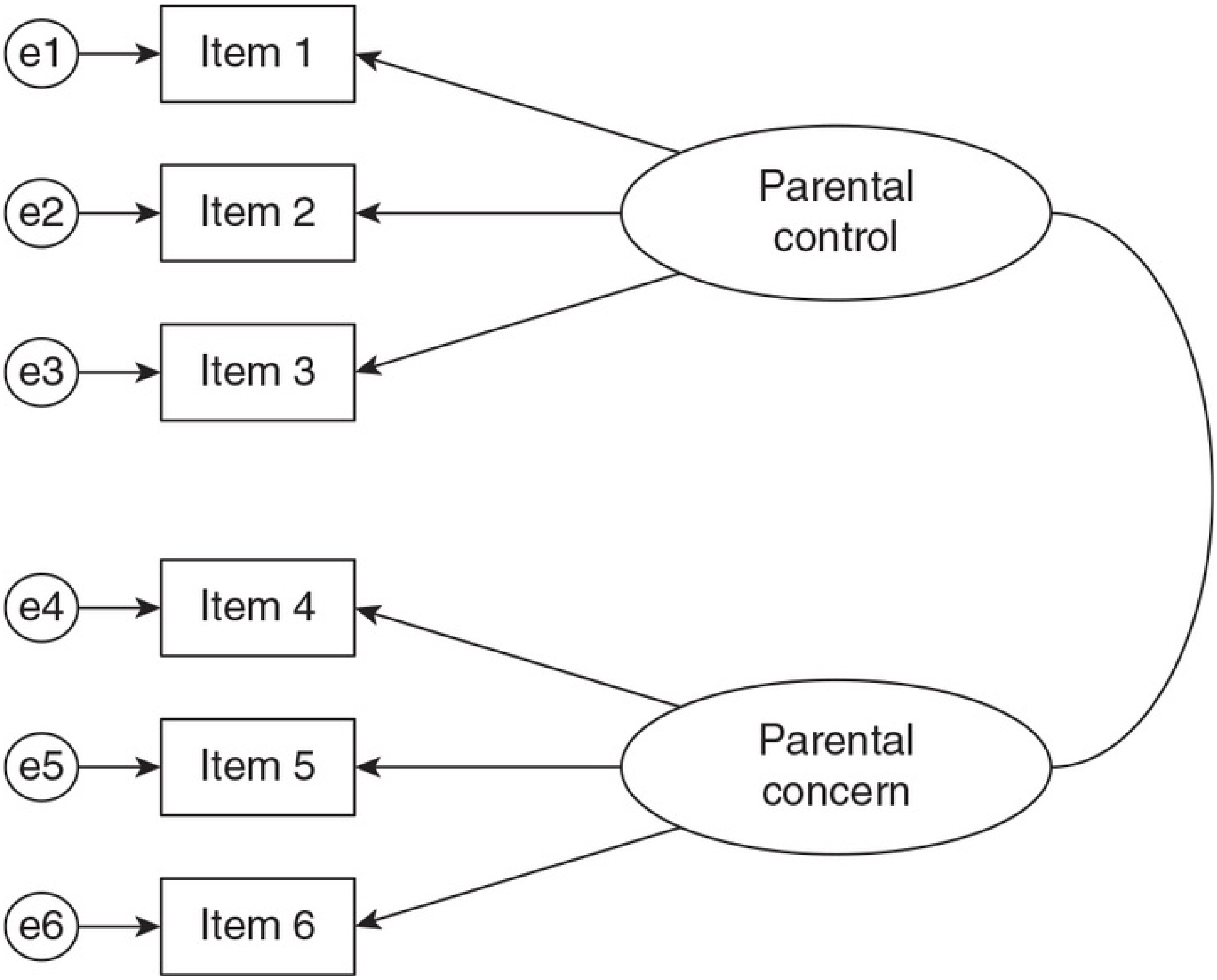


Figure 1 depicts this hypothesized model. Parental control (presented as an oval) has a direct effect (presented as a single-headed arrow) on observed variables from A1 to A3 (presented as rectangles). Similar to "parental concern" with three observed variables (i.e., A4 to A6) were loaded on this latent variable. The relationship between parental concern and parental control was represented by a two-headed arrow. These arrows only indicate the directionality, but no causal relationship is implied.

For each observed variable, a single-headed arrow representing an error is presented. In this model, there are (a) six factor loadings between the latent variable and six observed variables; (b) one covariance between the two latent variables; and (c) six errors are associated with six observed variables, thereby suggesting a total of 13 parameters being estimated.

Model Identification

In SEM, potential parameters can be *fixed* (either 0 or 1), *free* (needs to be estimated), or *constrained* (equal to one). This step is to assess the number of degrees of freedom (*df*) by testing the differences between the number of parameters to be estimated (unknown) and the information available (known) in the variance/covariance matrix. A model cannot be identified if the number of parameters to be estimated is larger than the information available in the variance/covariance matrix. Randall E. Schumacker and Richard G. Lomax noted three types of models: (1) a *just-identified* model refers to the number of parameters to be estimated that is equal to the information available in the variance/covariance matrix ($df = 0$), (2) an *underidentified* model refers to the number of parameters to be estimated that is larger than the information available in the variance/covariance matrix (*df* = negative), (3) an *overidentified* model refers to the number of parameters to be estimated is smaller than the information available in the variance/covariance matrix (*df* = positive).

Researchers generally use the formula $(p[p + 1])/2$ (where p refers to the number of observed variables) to assess whether the model is over-, just-, or underidentified. For example, in Figure 1, there are six observed variables with 15 parameters that need to be estimated (i.e., six factor loadings, six measurement error variances, two factor variances, and 1 covariance). The available information in the variance/covariance matrix is 21, $(6[6 + 1])/2$. This model is identifiable as the degrees of freedom are positive ($21 - 15 = 6$).

Model Estimation

Different estimation methods, including maximum likelihood, generalized least squares, weighted least squares estimation, Satorra–Bentler (S-B scaled chi-square) scaling method, and asymptotically distribution-free estimation, can be used to assess parameter estimates, standard errors, and fit indices. To select an appropriate estimation method, several factors (e.g., assumption of normality, sample size, and the number of categories in the observed variables) need to be considered. Maximum likelihood is widely used and available in most SEM software. However, it assumes the data are continuous and multivariate normally distributed. Both weighted least squares and asymptotically distribution-free estimation methods do not assume multivariate normality of the observed variables, but they require a large sample ($N \geq 500$). Recently, Satorra–Bentler (S-B) scaling method is suggested when handling nonnormally distributed data.

Model Fit

Model Fit

Once a model is estimated, several fit indices are used to evaluate how well the model fits the data. A nonsignificant chi-square indicates a good fit of the model. Yet, the value chi-square depends on sample size (inflated chi-square when the sample size is large). Therefore, other fit indices are suggested. Comparative fit indices (also known as incremental fit indices, such as the comparative fit index, the Tucker–Lewis Index, and the nonnormed fit index) may be used to assess the relationships among the variables. Absolute fit indices evaluate how well the hypothesized model fit the data. Examples are goodness-of-fit index (GFI) and adjusted GFI. Residual fit indices assess the difference between the observed data and the proposed model. These types of indices include standardized root mean square residual and root mean square error of approximation. Lastly, predictive fit indices assess how well a model fits the alternative model, which has the similar sized samples from the same population. Examples are the Akaike information criterion and the expected cross-validation index. SEM researchers recommend the following criteria for an excellent fit of the model: comparative fit index ≥ .95, GFI ≥ .95, Tucker–Lewis Index ≥ .95, nonnormed fit index ≥ .95, adjusted GFI ≥ .95, root mean square error of approximation ≤ .06, and standardized root mean square residual ≤ .08. Small values of the Akaike information criterion and expected cross-validation index indicate better fit of the model.

Model Modification

Once the model is selected, modification index (also known as the Wald and Lagrange Multiplier test) is inspected to improve the model. Large modification index suggests that the value of chi-square drops when a certain parameter is freely estimated. SEM researchers should modify a model based on the theoretical framework and empirical findings rather than be statistically driven. In the latter case, the findings may capitalize on chance.

Common SEM Models in Educational Research

Path Analysis

Unlike the traditional methods, such as analysis of variance and multiple regression analysis, this method allows researchers to test the direct, indirect, and total effects of variables simultaneously. Researchers hypothesize a model to test the causal relationships among variables. For example, in a study by Cecilia M. S. Ma and Daniel T. L. Shek, both family functioning and positive youth development qualities (independent variables) were found to influence consumption of pornographic material (dependent variable). In Figure 2, two direct effects are depicted by arrows from family functioning and positive youth development qualities to consumption of pornography. To test the mediating role of positive youth development qualities (i.e., the effect of family functioning on consumption of pornography has been intervened through positive youth development qualities), a proposed model was suggested by adding an indirect path from family functioning to pornography consumption through positive youth development qualities (see Figure 3).

Confirmatory Factor Analysis

The goal of confirmatory factor analysis is to test a hypothesized dimensional structure of a scale by assessing the relationships between the latent variable and its observed variables. Unlike EFA, a hypothesized model is needed to confirm the validity of the theoretical model in confirmatory factor analysis. Based on theoretical knowledge and/or empirical evidence, researchers test the underlying structure of the measurement scale by hypothesizing the relationships between observed variables and latent variables (see Figure 1) and comparing it with another competing model. For example, the six observed variables are accounted for parental control and parental concern, which are loaded on a general parental factor. These hierarchical relationships are shown in Figure 4.

Figure 2 A hypothesized direct effect path model

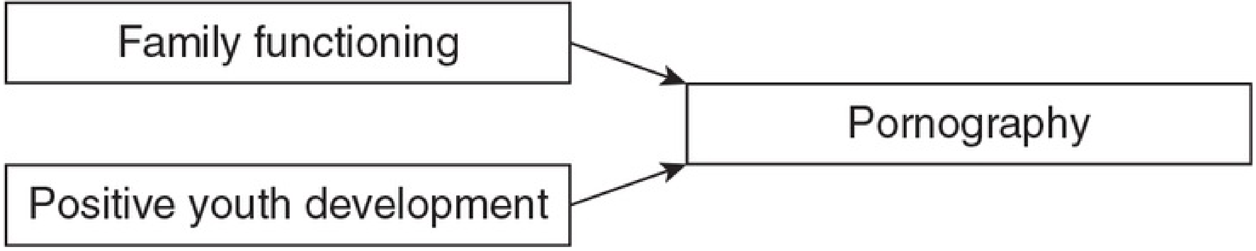


Figure 3 A hypothesized mediation path model

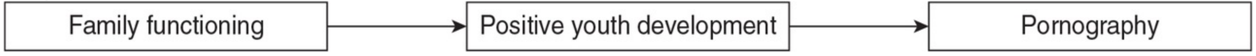
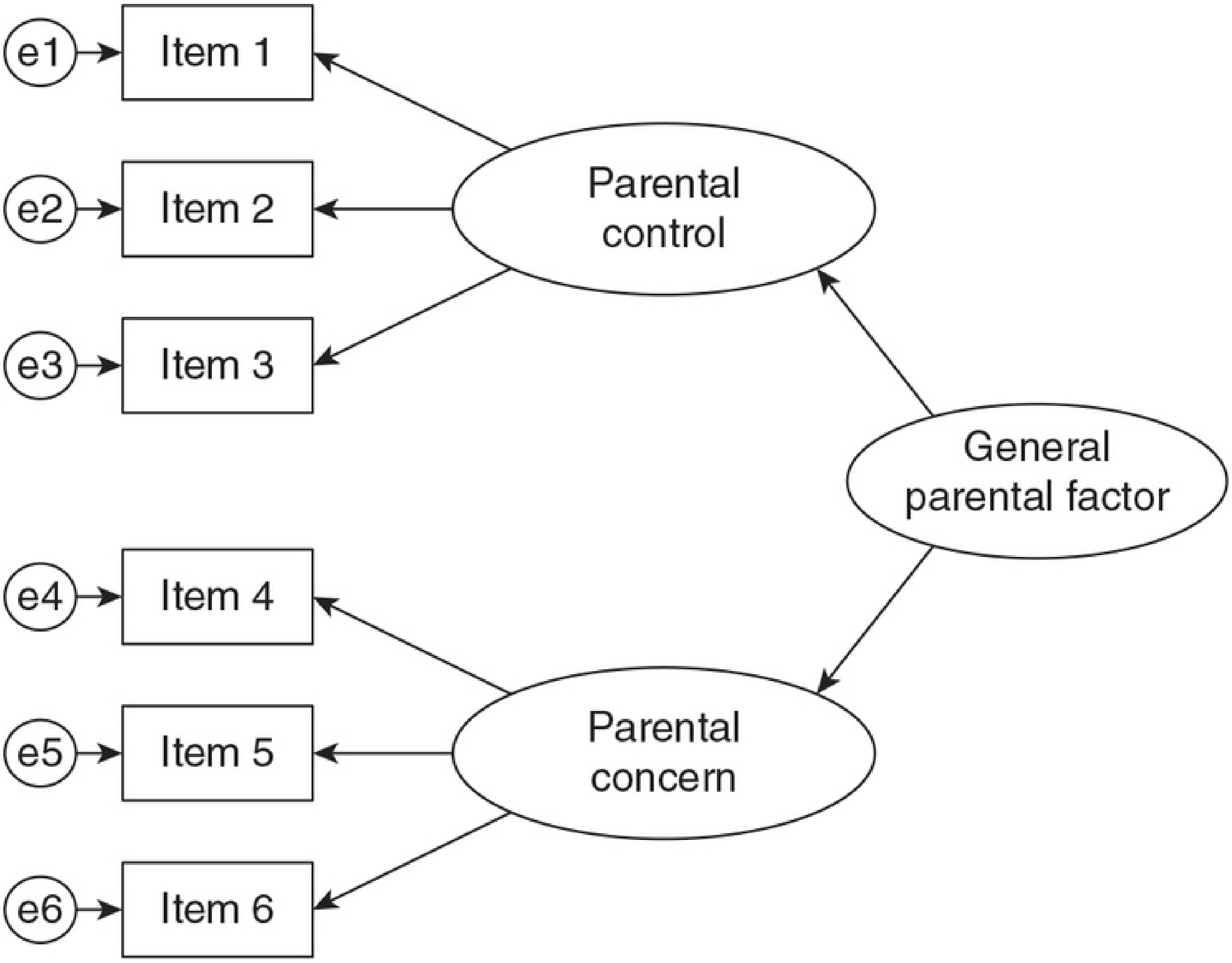


Figure 4 A hierarchical confirmatory factor analysis model



Latent Growth Model

The latent growth model is useful for longitudinal studies, as it assesses changes in an individual across time. This method describes individuals' behavior at an initial status (intercept) and observes their developmental trajectories (e.g., linear and quadratic) and tests how other variables contribute or affect the initial status and growth trajectories (e.g., age and socioeconomic status). For example, a theoretical model of adolescent consumption of pornography at three time points over three evenly spaced time intervals is presented in Figure 5. Intercept represents the initial state of consumption of pornography at the beginning of the study (Time 1). Slope indicates the rate of change from Time 1 to Time 3. To test a linear rate of change in consumption of pornography, the loading of the slope are fixed to be 0, 1, and 2 across the three time measurement points; whereas all loadings from the intercept are fixed to be 1. For example, researchers are interested in how boys and girls consumed pornography differently at Time 1 (the initial status) and/or changes over time. Therefore, a predictor is added in the model to test the effects of gender on adolescents' consumption of pornography at Time 1 (initial status) and such behavior changes over time (slope; See Figure 6).

Figure 5 A latent growth model with intercept and slope factors

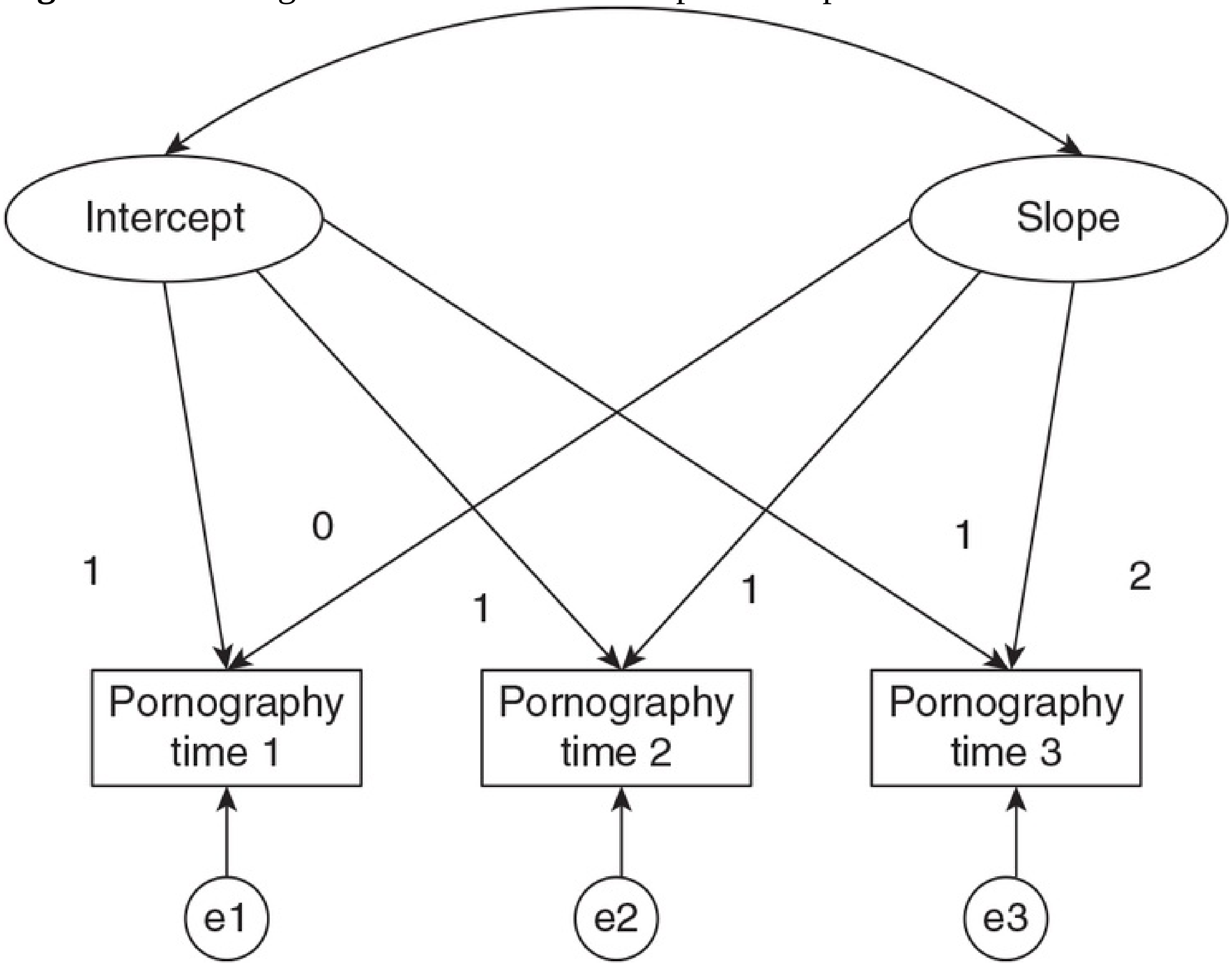
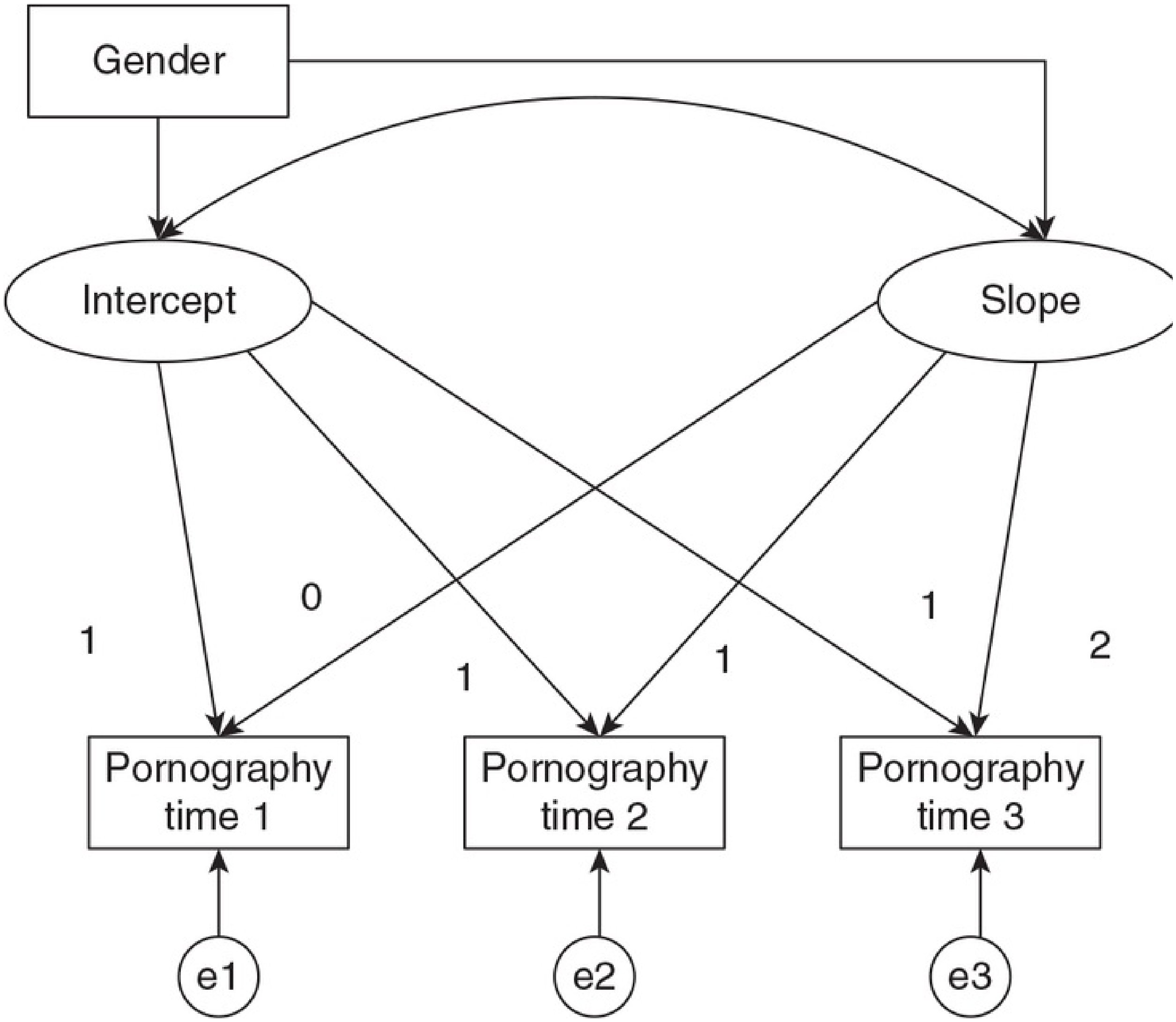


Figure 6 A latent growth model with predictor of the intercept and slope factors



Cecilia Ma and Daniel Tan-lei Shek

See also Confirmatory Factor Analysis; Mediation Analysis; Path Analysis; Validity

Further Readings

Arbuckle, J. L. (2012). Amos user's guide. Chicago, IL: Small Waters.

Bentler, P. M. (2006). EQS structural equations program manual. Encino, CA: Multivariate Software.

Chou, C., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 37–55). Thousand Oaks, CA: Sage.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), Structural equation modeling: A second course (2nd ed., pp. 439–493). Greenwich, CT: Information Age Publishing.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55. Retrieved from http://dx.doi.org/10.1080/10705519909540118

Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.80: Structural equation modeling with the SIMPLIS command language. Chicago, IL: Scientific Software International.

Kline, R. B. (2005). Principles and practice of structural equation modeling (2nd ed.). New York, NY: Guilford.

Muthén, L. K., & Muthén, B. O. (1998–2015). Mplus user’s guide (7th ed.). Los Angeles, CA: Author.

SAS Institute. (2012). SAS/ETS software: Changes and enhancements (Release 12.1). Cary, NC: Author.

Schumacker, R. E., & Lomax, R. G. (2004). A beginner’s guide to structural equation modeling (2nd ed.). Mahwah, NJ: Erlbaum.

Shek, D. T. L., & Ma, C. M. S. (2010). The Chinese family assessment instrument (C-FAI): Hierarchical confirmatory factor analyses and factorial invariance. Research on Social Work Practice, 20(1), 112–123. Retrieved from https://doi.org/10.1177/1049731509355145

Shek, D. T. L., & Ma, C. M. S. (2013). Subjective outcome evaluation of the project P.A.T.H.S. in different cohorts of students. International Journal of Child Health and Human Development, 6(1), 37–47.

Shek, D. T. L., & Ma, C. M. S. (2014). The use of confirmatory factor analyses in adolescent research: Project P.A.T.H.S. in Hong Kong. International Journal on Disability and Human Development, 13(2), 217–226.

Shek, D. T. L., & Ma, C. M. S. (2014). Using structural equation modeling to examine consumption of pornographic materials in Chinese adolescents in Hong Kong. International Journal on Disability and Human Development, 13(2), 239–246.

Yuko Goto Butler Yuko Goto Butler Butler, Yuko Goto

Student Self-Assessment Student self-assessment

1629

1632

Student Self-Assessment

Student self-assessment (SA) is a type of assessment in which learners evaluate their own performance or knowledge based on some criteria. Unlike other types of assessment, which are usually conducted by external agents such as teachers, administrators, or assessment developers, SA, being based on learners' judgments, has been considered a self-directed activity. SA has gained popularity because it is aligned well with modern educational approaches such as learner-centered, self-regulated, and autonomous learning. SA has been used for both formative and summative purposes and has taken a variety of forms. In practice, SA can be embedded in other assessments, such as portfolios, and can even be used as a replacement for externally measured assessments. Numerous can-do descriptors, which are a kind of SA, have been developed and implemented across different disciplines and contexts. Compared with many other types of assessment, administering SA is less constrained by large class sizes or time limitations, making it an attractive assessment option for practitioners.

This entry discusses SA in educational contexts from two assessment orientations—*assessment of learning* and *assessment for learning*. These two orientations come from different theoretical and empirical traditions and thus conceptualize the role of SA in education quite differently.

SA From the Assessment of Learning Point of View

In the traditional approach to assessment, or the assessment of learning orientation, the purpose of assessment is to gain information to make accurate and consistent inferences of students' true abilities or the level of their acquired knowledge and skills. Thus, students are the object being measured and are external to the inferences being made. Based on this orientation to assessment, the self-directed and subjective nature of SA can be a threat to validity and reliability. Concerns have been addressed regarding the extent to which SA accurately and consistently captures students' mastery of the target skills or knowledge.

The published research investigating the relationship between students' SA results and their externally measured skill or knowledge levels (i.e., levels measured through objective test scores, teachers' evaluations, and course grades) has shown mixed results. Correlations vary greatly. Although a number of studies report that students, at least among adult learners such as college students and trainees in professional courses, are capable of accurately self-assessing their own performance and knowledge, other studies do not support such claims.

Researchers have identified a number of factors that seem to contribute to the variability in SA accuracy (i.e., correlations between SA results and externally measured skill or knowledge levels). First, the accuracy of SA responses varies depending on the target knowledge or skills being assessed. It is known that students can more accurately self-assess lower order cognitive skills than higher order cognitive skills. Interestingly, students are also better at self-assessing their current skill or knowledge levels (i.e., absolute levels) than self-assessing the improvement of such skill or knowledge. Perhaps this is because using SA to assess gains in skill and knowledge requires judgments based not only on external criteria but also on learners' self-reflected criteria, meaning that learners need to notice and capture changes in skill and knowledge levels by comparing them at different points in time.

Second, the wording and construction of SA items influence students' responses. Students respond differently when items are negatively formed (e.g., "It is hard for me to do XX" or "I cannot do …") versus when they are positively formed (e.g., "I can do XX"), although the degrees of inconsistency in response to the differently worded items can vary depending on the items. Not too surprisingly, in second-or foreign-language programs, students can more accurately self-assess their proficiency in the target language when they have a chance to respond to the SA items in their first language rather than in their target language.

Third, experiential and environmental factors that help students develop or better understand criteria increase the accuracy of SA. Having greater experience with receiving feedback, especially feedback on accuracy of their self-assessed receiving feedback, especially feedback on accuracy of their self-assessed results, helps students improve their SA accuracy. Environments in which students have more opportunities to observe other people’s performance and knowledge attainment enhance their abilities to self-assess their own performance and knowledge. For example, students who learn in physical classroom settings or in courses that teach interpersonal skills are more likely to self-assess their learning outcomes more accurately than students who learn solely via web-based programs or courses with fewer opportunities to interact with others. These findings suggest that the process of responding to SA is in fact a social activity as well as an individual activity, contrary to the general beliefs about SA being a purely introspective activity.

Additional individual factors that influence students’ SA responses include their skill/knowledge mastery levels, personality, affective states, and age. Advanced learners tend to underestimate their skills and knowledge, whereas beginners or less advanced learners tend to overestimate their skills and knowledge. Students’ self-esteem, motivation, emotion, and confidence can affect their responses to SA items. Indeed, there is some evidence that adult learners’ SA responses have stronger correlations with their motivation and course satisfaction than with their externally measured knowledge and skill levels.

Age has been suggested as an influential factor for SA accuracy, but our understanding of how children age 12 years and younger respond to SA items remains quite limited. Research on child development has consistently shown that younger children (children younger than 7 years of age) tend to have high self-appraisal regardless of their actual skill or knowledge levels. Influenced by Jean Piaget’s theory of cognitive development, psychologists used to consider that this tendency was largely due to young children’s lack of mental maturity to self-assess their abilities. More recently, however, psychologists have suggested that children’s high self-appraisal is largely due to age-related factors (e.g., lack of experiences and contexts) rather than their cognitive immaturity per se. It turns out that children are capable of accurately self-assessing their performance on familiar tasks. When children have greater experience with SA and receive appropriate scaffolding when conducting SA, they also make more accurate and more stable SA responses. Children with extensive experience interacting with others, compared with children without such experience, tend to use more normative information (information based on social comparison) and to be less egocentric in their SA responses.

SA From the Assessment for Learning Point of View

SA From the Assessment for Learning Point of View

As seen so far, from the assessment of learning perspective, researchers and educators are mainly concerned with how SA can best capture students' skills and knowledge accurately and consistently. If SA is reasonably valid and reliable, then it can be used for summative purposes or as a replacement for existing external measures. From the assessment for learning perspective, however, the role of SA can be conceptualized quite differently. According to this perspective, which is greatly influenced by constructivist theories, the primary goal of assessment is to obtain information about the process of students' learning in order to inform and assist their ongoing learning. Thus, major validity concerns include the extent to which the content and methods of assessment are matched with instruction or students' actual learning experiences. Students are no longer merely objects being measured but active agents making inferences about their abilities and performance and, along with their teachers, taking actions based on those inferences. When it comes to SA, therefore, validity concerns include the extent to which learners can self-reflect through assessment and can benefit from using SA as a means to enhance their learning. From the view of assessment for learning, SA is promising in that, by having students engage in self-reflective activities, it can help them enhance their autonomy, motivation, and learning. For example, learners accustomed to reflecting on their own writing performance through SAs might be better aware of the strengths and weaknesses of their essay, which makes it easier for them to set a goal for the next writing assignment to overcome their weaknesses.

Among adult learners, some evidence indicates that SA has a positive influence on both students' perceived effectiveness and actual learning outcomes as captured by objective measurements such as external tests, grades, and evaluations by instructors. To benefit from SA, students have to (a) have a clear understanding of the criteria or goals of the assessment tasks, (b) self-reflect on their current level of learning or understanding and identify the gap between the current level and the goal, and (c) take appropriate actions to achieve the goal.

To facilitate the processes just described, researchers have made a number of pedagogical suggestions. For example, to help students better understand the criteria or goal, teachers are encouraged to incorporate concrete examples along with descriptors (e.g., showing writing examples with the writing rubrics). Another suggestion is for teachers to discuss the criteria or goals with their students, either in person or in groups. Some researchers have suggested that peer assessment (i.e., assessment of one's performance or abilities undertaken by one’s peers) should be employed before SA because peer assessment helps students understand the criteria. Indeed, empirical studies have shown that peer assessment has higher internal consistency psychometrically and higher correlations with external measures than SA, although the evidence also shows that SA can assist student learning more directly than peer assessment. To help students develop self-reflection abilities and take appropriate actions to achieve their goals, sufficient SA experience and feedback are indispensable. Research has shown that the effectiveness of feedback varies substantially across studies, suggesting that both the quality and timing of feedback matter. For example, social cultural theory, a constructivist view, emphasizes the significance of identifying the gap between the level at which a learner can solve problems independently and the level at which the learner can solve problems with external mediations through social interaction with capable others. According to the theory, this gap provides teachers and learners with optimized instructional and learning opportunities.

For young learners, research on SA from the assessment for learning perspective remains relatively limited. Developmental psychologists have found that children gradually develop self-regulatory abilities during their preschool and primary school years. However, self-assessing one’s current level of understanding appears to be harder in some domains than in others. As seen already, self-assessing one’s *progress* of skills and knowledge is even harder than self-assessing one’s *current* level. This is particularly evident among children. Although we have limited understanding of how children arrive at their judgments, namely their processes and rationales for responding to SA items, some evidence indicates that they rely on various sources to make self-evaluations, including the amount of effort they exert to complete the task in question.

As with adults, it is important for young learners to clearly understand the reasons for doing SA and the criteria for evaluation, although in the assessment for learning orientation, the criteria can be flexible in order to meet individual student’s needs. Compared with adults, however, young learners usually need much greater and individualized assistance from teachers or capable others. Research also indicates that the effect of SA on young learners’ learning is influenced by both their learning environment and their teachers’ attitudes toward assessment. When teachers have a deep understanding of assessment for learning and foster collaborative learning environments, SA for learning is more effective.

Yuko Goto Butler

See also Constructivist Approach; Formative Assessment

Further Readings

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy and Practice, 5(1), 7–74. Retrieved from http://dx.doi.org/10.1080/0969595980050102

Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning. In M. Nikolov (Ed.), Assessing young learners of English: Global and local perspectives (pp. 291–315). New York, NY: Springer.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning for affective measure? Academy of Management Learning & Education, 9(2), 169–191. Retrieved from http://dx.doi.org/10.5465/AMLE.2010.51428542

Janice A. Hogle Janice A. Hogle Hogle, Janice A.

Success Case Method

Success case method

1632

1633

Success Case Method

The success case method (SCM) is a type of qualitative research used by evaluators to achieve an in-depth understanding of a phenomenon—a program, person, place, or other entity. The modifier “success” means that the entity has met objectives or attained an appropriate accomplishment. The in-depth nature of case studies provides detailed textual description that can inspire or motivate readers in ways that more “simple” quantitative indicators cannot do, by exploring the particularity and complexity of a single instance of success. A case might also be described as a vignette, a story, or an example and can vary in length from a paragraph to hundreds of pages. There are many different ways to conduct case study evaluation, but of critical importance is the need to accurately document evidence presented in a way that is honest, credible, and confirmable.

The SCM as used in the field of educational program evaluation highlights the critical role that qualitative methods can play in contributing to in-depth understanding and explanation of complex variables used for assessment and measurement. When requirements for evaluation are limited by time constraints, budget restrictions, data quality, and political challenges, the SCM, used appropriately, can greatly enhance the utility of evaluation results for decision making by key stakeholders.

Success cases can be particularly useful for programs challenged with meeting long-term objectives for which attribution is difficult or questionable. A benefit of the SCM is the focus on examples of programs that are “working” including the how and why of working. The SCM cannot definitively establish causality of outcomes, but as part of a mixed-methods (both qualitative and quantitative data) outcomes, but as part of a mixed methods (both qualitative and quantitative data) evaluation, the technique can substantially contribute to better understanding of program processes and resource accountability. In addition, successful cases often suggest approaches useful for continuous program improvement based on the details contained in the descriptions. Some evaluators prefer to include cases that are unsuccessful as well as successful in the belief that the contrast between the two can provide additional insight into program improvement. This entry notes the methods commonly incorporated in SCM, discusses how evaluators use SCM, and reviews the limitations of this approach.

Qualitative Methods

The SCM might incorporate a variety of qualitative methods including ethnography, participant observation, individual or key-informant (in-depth) interviews, or focus groups. These qualitative data collection techniques generally require a face-to-face interaction between two people (individual interviews) or within a group (focus groups). The interaction can involve a semistructured question guide and is characterized by conversation consisting of questions and answers. The interviewer is trained in how to pose questions and how to respond to answers with additional probing questions designed to better understand the responses of the interviewee.

How Evaluators Use SCM

Within the context of a program evaluation, an evaluator would first clarify key goals, objectives, and expectations of the program. Priority objectives then suggest evaluation questions, the answers to which might be provided by indicators of achievement. Quantitative data often pose additional questions that are better answered using qualitative approaches. The contribution of successful cases (qualitative descriptions) to understanding program achievements can then complement quantitative indicator data. It is possible to collect case data relatively quickly and simply; however, the training and experience of interviewers is critical to data quality, interpretation, and utility to program managers.

Initial conversations with program participants at multiple levels can identify cases that might be ideal for more in-depth investigation. Researchers then decide on the following: which and how many cases to include, the size of the data collection team, time available for data collection and analysis, content of the question guides, analytic process, desired report format, most appropriate dissemination strategy, and ways in which findings can be used to improve future performance.

Limitations

As with any qualitative data, findings are not generalizable. SCM focuses in depth on details and descriptions specific to successful examples in an effort to elucidate explanations and causal suggestions. Program impacts are often long term, while funding cycles are relatively short with intense pressure to “show impact” in order to justify funding levels. While SCM can showcase examples of success, the textual descriptions may not convince readers that a program is actually meeting its objectives to the desired degree.

Understanding what one is hearing from a respondent is, of course, critical to interpretation and explanation of the details of the case study. Thus, if the interviewer is unfamiliar with the subject area of the topic being discussed, the interviewer may not know how to probe in response to complex explanations provided by the respondent, and then the resulting analysis will be less useful.

While interview data can be transcribed relatively quickly by paid transcribers, the analysis and interpretation is performed by those who conducted the interviews, as they are the most familiar with the cases. Analysis of textual data is often tedious and time consuming, so if there are many cases, or even a large amount of information on one case, then the analysis can take a prohibitively long time.

Case study examples are snapshots in time of a certain phenomenon and thus rapidly become out of date. If a set of cases are to be studied, the time frame involved has to be fairly compact, so that all the cases are characteristic of a particular time period. This requirement then poses challenges in terms of accurately planning for the number of cases needed, number of interviewers, and necessary analysis time. If data collection occurs over months or years, then the findings relating to each case may no longer be relevant. If it is important to follow cases over a longer time period, then additional resources are needed.

Anonymity and confidentiality are important issues in developing success case examples. Usually, the total number of cases is small and textual description is detailed and specific, so that the identity of respondents is often impossible to protect. These concerns need to be discussed up front with program stakeholders and the evaluation team, so that appropriate precautions can be taken to protect respondents.

Finally, although it is possible to use *only* qualitative success cases to evaluate a program or some other entity, generally, an evaluation is most effective when incorporating mixed methods to assess achievement of objectives.

Janice A. Hogle

See also: Ethnography; Focus Groups; Interviews; Mixed Methods Research; Participant Observation; Qualitative Research Methods

Further Readings

Brinkerhoff, R. O. (2002). The success case method. San Francisco, CA: Berrett-Koehler.

Patton, M. Q. (2011). Developmental evaluation: Applying complexity concepts to enhance innovation and use. New York, NY: Guilford Press.

Stake, R. E. (1995). The art of case study research. Thousand Oaks, CA: Sage.

Yin, R. K. (2003). Applications of case study research. Thousand Oaks, CA: Sage.

Joseph R. Nichols, Jr Joseph R. Nichols, Jr Nichols, Jr, Joseph R.

Summative Assessment

Summative assessment

1633

1635

Summative Assessment

Summative assessment refers to the assessment of students that occurs at the end of a period of instruction. The purpose of summative assessment is to provide teachers and others with a summary view of learning and accomplishments. In this sense, summative assessment provides a holistic measurement of an individual’s knowledge, skills, or dispositions. Summative assessment is best understood in comparison to formative assessment. Summative assessment, on one hand, describes the traditional use of classroom (and most standardized achievement) testing; it comes at the end, it is used to assign a grade, and the information is meant primarily for teachers, parents, and administrators, though it is shared with students. For students, it provides the answer to the question: How did I do? Formative assessment, on the other hand, tells students and teachers: How am I doing? It occurs during instruction and provides feedback to both learners and teachers. It rarely is counted into a grade. This entry discusses two uses of summative assessment. The first section outlines how summative assessment is used to measure an individual’s learning. The second section outlines the role summative assessment plays in measuring the teacher quality.

Measuring Learning

Summative assessment helps teachers gauge how much their students have learned—especially over a given period of time. Summative assessment consists of a variety of different formats from multiple-choice exams to research papers to portfolios of student work.

Assigning grades at the end of a period of instruction has been the common purpose of assessment in schools for hundreds of years. It is almost as common today, especially in standards-based education. Standards define the knowledge, skills, or dispositions a student should learn in a given period of time or through a specific set of experiences. Summative assessments are designed to evaluate a student's obtainment of the knowledge, skills, or dispositions outlined in the totality of the standard the summative assessment is measuring.

A key focus of standards-based summative assessment is attention to the verbs articulated in the standard. For example, if the learning standard in question states "students will evaluate historical arguments using primary sources," the summative assessment will ask students to judge and determine the quality of historical arguments by using primary source evidence. As such, summative assessments focus on the criterion that defines what students are asked to do. Most often, instructors use this type of summative assessment to measure how much students learn in their individual classroom. Ask the "man on the street" to picture classroom assessment, and images of summative assessments like pop quizzes, unit tests, and finals come to mind. Until recently, almost all assessment in classrooms was summative. Its purpose was to assign a grade, differentiate students from each other, and separate good students from bad students. Although formative assessment has grown in popularity since the 1990s, summative assessment remains the predominate approach to evaluating student learning in classrooms. Summative assessment, however, is also used across classrooms to measure the overall quality of academic experiences. This form of summative assessment focuses on program quality and is taken up in the next section.

Measuring Teachers

Summative assessments are used in the evaluation of teachers in training and after training. Summative assessment in this context is often used as an accreditation and program approval tool. This practice is exemplified in teacher and school leader preparation. For example, content-based state licensure exams are summative assessments that measure whether teacher or leader candidates from specific programs meet the minimum knowledge criteria necessary to teach in or lead a public school. Performance-based assessments such as the edTPA and PPAT provide summative reviews of skills graduates have when they exit a teacher or leader preparation program. These assessments measure how well candidates exiting a particular program can perform the daily tasks they are expected to complete while on the job. Summative assessment is used not only to judge preservice teachers but also to assess them while in practice. It is becoming more common for states and districts to use class-wide performance on state-wide tests as part of formulas for evaluating teachers. These classic summative assessments, state achievement tests, are not designed for evaluating instruction or teachers, and there are many validity problems with this approach. It is consistent, though, with the traditional view that a summative exam is the best way to judge classroom success.

Summative assessment—because it is a measure *of* learning rather than a measure *for* learning—is high stakes. Whether summative assessment is used to measure an individual's learning or to gauge teacher quality, the nature of the assessment focuses on the knowledge, skills, or dispositions a student was supposed to master or a teacher was supposed to teach. As such, summative assessment draws attention to where individuals and schools excel as well as where they fall short. Summative assessment can, theoretically, be used like formative assessment to change teaching and program delivery. Although the feedback to teachers is slower than with frequent formative assessment, over time, teachers can improve instruction so that individual students learn more and academic programs produce better outcomes. In this sense, summative assessment can be a powerful tool in the teaching and learning process.

Joseph R. Nichols, Jr

See also Accreditation; Formative Assessment; Formative Evaluation; Program Evaluation; Summative Evaluation

Further Readings

Angelo, T. A., & Cross, K. P. (1993). Classroom assessment techniques: A handbook for college teachers. San Francisco, CA: Jossey-Bass.

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. Theory Into Practice, 55(2), 153–159. Retrieved from http://dx.doi.org/10.1080/00405841.2016.1148989

Harlen, W. (2007). Teachers' summative practices and assessment for learning—tensions and synergies. The Curriculum Journal, 16(2), 207–223. Retrieved from http://dx.doi.org/10.1080/09585170500136093

Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve student performance. San Francisco, CA: Jossey-Bass.

William, D., & Black, P. (2006). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? British Educational Research Journal, 22(5), 537–548. doi:10.1080/0141192960220502

Anthony Jason Plotner Anthony Jason Plotner Plotner, Anthony Jason

Summative Evaluation Summative evaluation

1635

1637

Summative Evaluation

An evaluation is a systematic and purposeful collection and analysis of data used to document the effectiveness of programs or interventions. Rigorous evaluation can determine if programs or interventions should be maintained, improved, or eliminated. The term *summative evaluation* (sometimes referred to as *ex-post evaluation* or *outcome evaluation*) was first introduced in the mid-1960s by Lee Cronbach and Michael Scriven and refers to a process of evaluating a program's or intervention's impact or efficacy through careful examination of program design and management. It is often used to assess the accountability of a program or intervention. As such, summative evaluation is outcome focused more than process focused and most often undertaken at the end of the project, when the program or intervention is stable and/or when program services are implemented with consistency (otherwise known as fidelity). Furthermore, there are some types of summative evaluation that require the collection of baseline data in order to provide a before and after understanding; thus, it is important to factor this into the evaluation. Summative evaluation is undertaken to determine whether the program or intervention achieved its goals, objectives, or outcomes; how the program's impact compares to different programs; and to better understand the process of change, what works, what doesn't, and why.

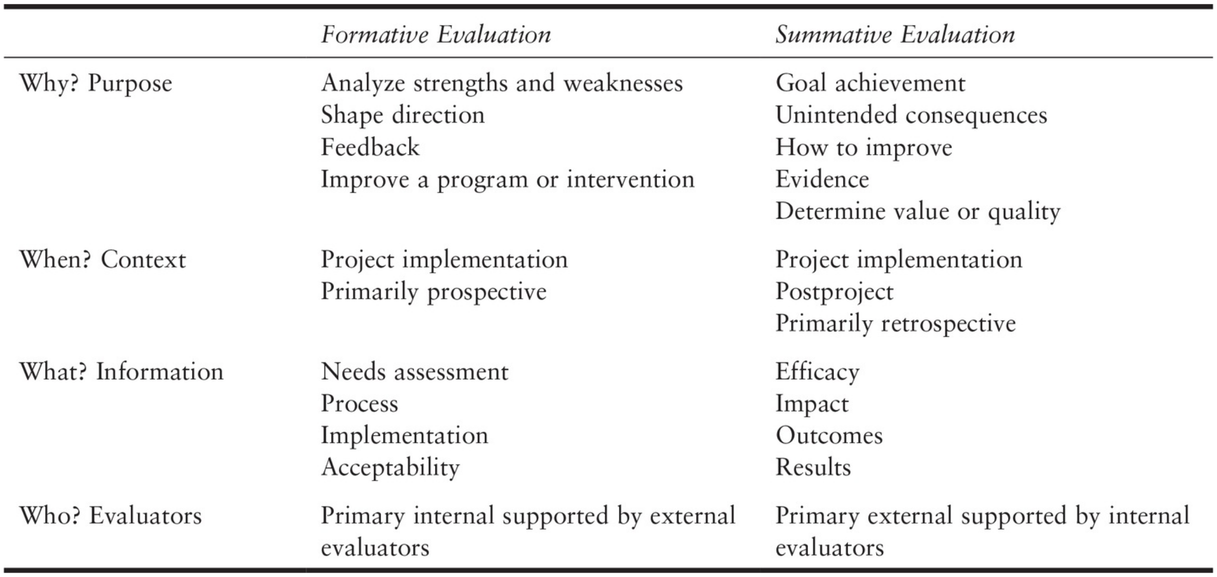
Understanding Summative Evaluation

Summative evaluation is also often conducted or undertaken by people considered independent or external of the responsible project. The methods used to gather the data used in a summative evaluation should incorporate a detailed step-by-step procedure that is carefully designed and executed to ensure the data are accurate and valid. A balance of both quantitative and qualitative methods can help researchers obtain a better understanding of project achievements and information that led to these achievements. The various instruments or tools used to collect data when conducting a summative evaluation include interviews, questionnaires, surveys, observations, and testing.

Summative evaluations are conducted to determine the value of a program or intervention—its merit or worth, often in comparison with other programs or interventions. Summative evaluation can enable stakeholders to make decisions regarding specific services and the future direction of the program that cannot be made during the beginning or middle of program or intervention implementation. By contrast, formative evaluation (also known as process or implementation evaluation) is designed to form or improve the program or intervention being evaluated by examining aspects of an ongoing program in order to make improvements as the program is being implemented. Most evaluations can be summative (i.e., have the potential to serve a summative function), but only some have the additional capability to serving formative functions. One way to truly understand summative evaluation is to differentiate between formative and summative evaluation. It is considered good evaluation practice to include both formative and summative evaluation. Table 1 shows some fundamental differences between formative and summative evaluation.

Common Types of Summative Evaluation

There are a variety of types of summative evaluations. Some of these types include cost-benefit/cost-effectiveness analysis, goal-based evaluation, outcome evaluation, secondary analysis, meta-analysis, and impact evaluations. Cost-effectiveness and cost-benefit analysis address questions of efficiency by standardizing outcomes in terms of their dollar costs and values. Goal-based evaluation determines if the intended goals of a program or intervention were achieved. Outcome evaluation investigates whether the program caused demonstrable effects on specifically defined target outcomes. Secondary analysis examines existing data to address new questions or use methods not previously employed. Meta-analysis integrates the outcome estimates from multiple studies to arrive at an overall or summary judgment on an evaluation question. Impact evaluation is broader and assesses the overall or net effects—intended or unintended—of the program or intervention.



	Formative Evaluation	*Summative Evaluation*
Why? Purpose	Analyze strengths and weaknesses Shape direction Feedback Improve a program or intervention	Goal achievement Unintended consequences How to improve Evidence Determine value or quality
When? Context	Project implementation Primarily prospective	Project implementation Postproject Primarily retrospective
What? Information	Needs assessment Process Implementation Acceptability	Efficacy Impact Outcomes Results
Who? Evaluators	Primary internal supported by external evaluators	Primary external supported by internal evaluators

Anthony Jason Plotner

See also Evaluation; Formative Evaluation; Program Evaluation; Summative Assessment

Further Readings

Coryn, C. L. S., & Scriven, M. (Eds.). (2008). Reforming the Evaluation of Research: New Directions for Evaluation, Number 118. San Francisco, CA: Jossey-Bass.

Coryn, C. L. S., & Westine, C. D. (Eds.). (2015). Contemporary trends in evaluation research. Sage Benchmarks in Social Research Methods (Vols. 1–4). London, UK: Sage.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), Perspectives of curriculum evaluation (pp. 39–83). Chicago, IL: Rand McNally.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. D. Phillips (Eds.), Evaluation and education: At quarter century (pp. 19–64). Chicago, IL: University of Chicago Press.

Wholey, J. S. (1994). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), Handbook of practical program evaluation (pp. 15–39). San Francisco, CA: Jossey-Bass.

Website

Web Center for Social Research Methods: www.socialresearchmethods.net

Carrie La Voy Carrie La Voy Voy, Carrie La

Supply Items

Supply items

1637

1638

Supply Items

Supply items are a name given to certain types of assessment questions. Supply items are named this way because students are asked to supply the answer to a question, rather than selecting or choosing an answer to a question. Sometimes assessments contain supply item questions that have one correct answer. An example of this type of supply item is a fill-in-the-blank question. Sometimes assessments contain supply item questions that ask students to construct a more detailed and original response. These types of supply items are referred to as constructed-response items. Examples of constructed-response items include short answer or essay questions. When supply item assessment questions ask students to construct and supply an answer, the result could be in the form of a performance. These types of items are further categorized as performance-based assessments.

On most tests, the assessment items can be categorized as either objective or subjective. Each type of item serves a purpose. Most subjective assessment items are supply items, and some objective assessment items are supply items. The term *objective* describes test items that are more factual and have short, unambiguous, right or wrong answers. Examples include true–false, matching, multiple-choice, and completion questions. Some objective test items ask students to choose the correct response to a question. This often takes the form of a multiple-choice question, where students select the correct response from a list of possible choices. Other objective test items are supply items that ask students to supply a word or phrase to answer a question. A specific example is a fill-in-the-blank question.

The term *subjective* describes test items that are not based on one right or wrong answer. Subjective items have students create their own original written response to a question. Because students supply their own response, subjective questions are categorized as supply items. There is usually more than one correct way to answer these questions. In fact, often there are usually several possible correct answers, and students can earn full or partial credit. There is more flexibility in scoring answers to these types of questions. Examples of subjective assessment items include short answer essays, extended response essays, and performance-based items.

Performance-based assessment items ask students to supply a response, so these types of items are also categorized as supply items. Performance-based items measure students' ability to apply their knowledge and skills and then demonstrate this ability as a performance of some kind. Examples of performance-based assessments include giving an oral presentation or speech or even participating in a debate. Some performance-based assessments ask students to produce or construct a product. Regardless of the format, the intent of performance-based assessments is for students to provide evidence of their knowledge and demonstrate application of skills. Options for products for performance-based assessments vary greatly.

Carrie La Voy

See also Alternate Assessments; Authentic Assessment; Constructed-Response Items; Fill-in-the-Blank Items; Matching Items; Objectivity; Performance-Based Assessment; Tests

Further Readings

Frey, B.B. (2014). Modern classroom assessment. Thousand Oaks, CA: Sage.

Center for Innovation in Teaching and Learning. (2015). Improving your test questions. Champaign: University of Illinois. Retrieved from http://cte.illinois.edu/testing/exam/test_ques.html

Jason T. Siegel Jason T. Siegel Siegel, Jason T.

Natalie D. Jones Natalie D. Jones Jones, Natalie D.

Survey Methods

Survey methods

1638

1642

Survey Methods

Although definitions and colloquialisms vary, *survey research* generally refers to the systematic collection of self-report data from a sample of a larger population. When *survey methodology* is employed, a central goal is to obtain valid data that accurately represent a predetermined population. Tasks associated with survey research can include but are not limited to the following: chronicling demographics, assessing attitudes and beliefs, and documenting the frequencies of specific behaviors and intentions to engage in such behaviors. The main thing that inhibits this goal of obtaining valid data is *error*, which in this context refers to data that fail to capture the true physical or psychological characteristics of the population. Thus, one overarching goal of survey research is to reduce error so that the data collected are an accurate representation of the population.

There are myriad decisions at every step of the survey process that can either limit or amplify the amount of error and, therefore, the accuracy of the data collected. Furthermore, there is a constant need to consider the benefits of each approach in contrast to the costs. For example, a longer survey will provide more information but also leads to more participant fatigue—and potentially more error. Incentives can increase response rates but must be balanced with the reduction in sample size that will occur due to a restriction of resources.

Survey methodologists make decisions in regard to the people that will constitute the population of interest (e.g., teachers who have been teaching 10 or more years in urban schools) as well as the sampling approach to be employed (e.g., multistage cluster sampling). Researchers must make decisions regarding the sample size, the mode of the data collection, the research design, and the questions they will ask. Judgments are also needed in regard to the length of the survey, the appearance of the survey, and how missing data will be handled and the statistical assessments are to be used. Illustrative of the breadth of goals and challenges associated with survey research, entire encyclopedias have been dedicated to the topic of survey research. The current entry provides a snapshot of the complexities of the survey research process, with a focus on some of the errors that are most likely to occur at each stage.

Determining the Population: Coverage Error

When conducting a survey, one of the first steps is to define the target population (i.e., the population of interest). For example, a school board might be interested in the opinions of first year teachers from urban areas, or there might be interest in the opinions of students from schools built in the past 3 years. Once this population of interest is determined, a sampling frame (i.e., a list of the entire population of interest) must be acquired or created—a task that could be relatively easy or challenging. A list of the population to be surveyed might be readily available (e.g., teachers of a specific school district) or a list may not exist (e.g., homeless people in Los Angeles). The goal at this stage of the process is to be sure that the population is properly represented. If the population is not properly represented, the data obtained will be unlikely to accurately represent the population of interest, as segments of the population could have been left out. If a specific aspect of a population is left out of the survey process, the survey data could lead to misguided decision making.

Coverage refers to the extent to which the sampling frame, that is intended to include the population of interest, achieves its goal. Relatedly, *coverage error* occurs when there is a disconnect between the population of interest and the people from whom data were actually collected. Coverage error can occur during two central phases of the survey research process—in creating the sampling frame (e.g., the entire population is not included) or during data collection (e.g., a certain part of the population is unable to be reached). In regard to the creation of the sampling frame, there are two forms of coverage errors: undercoverage and overcoverage. For instance, if a researcher wants to conduct a survey using all schools in the county as the sampling frame, undercoverage would occur if the list were outdated and did not contain schools built in the past 3 years in the sampling frame. Conversely, overcoverage would occur when schools from three counties were included in a sampling frame that was only supposed to include one county. Once the population of interest is determined and a sampling frame created, decisions must be made regarding which participants of the total population will be given the opportunity to take part in the survey.

Sampling the Population: Sampling Error

Researchers and evaluators that utilize surveys are tasked with determining the best means through which a percentage of the predetermined population can be sampled, while ensuring that the sample is representative of the broader population and minimizing error. The primary goal of the sampling plan is to determine an approach that will minimize survey error; a secondary goal is to do so with the least resources possible. Numerous variables, such as cost per participant, difficulty in accessing participants, and the percentage of people who are expected to respond, influence the sampling plan.

There are numerous sampling approaches that researchers commonly utilize; one way to categorize these approaches is to determine whether they utilize *random sampling*. In a random sampling approach, researchers use some form of randomization mechanism (e.g., random number generator) to select participants, thus allowing all participants within the sampling frame an equal chance of being selected. For example, cluster sampling is one approach that falls under the category of random sampling. When cluster sampling is implemented, there is a random sampling of geographic locations (e.g., schools within a district) and all members from the cluster are included in the sample. Other types of random sampling techniques include simple, systematic, and spatial sampling. The benefit of random sampling is that it reduces the likelihood that there will be error introduced as a result of who is selected to take the survey. The downside of random sampling is the cost that is sometimes involved in creating, buying, or some other way of obtaining a list of the population of interest.

An alternative to random sampling is *nonrandom sampling*, which includes sampling techniques that do not give all participants within a sample an equal chance of being selected. For example, in one nonrandom sampling technique, snowball sampling, participants are surveyed and then asked to recruit other people to participate. Snowball sampling is particularly useful when trying to obtain data from a hard to reach population. Other examples of nonrandom sampling include quota sampling, purposive sampling, accidental sampling, and convenience sampling. Nonrandom sampling is typically less expensive than random sampling, but the extent to which the results are representative of the population of interest may be minimized as well.

Regardless of the sampling plan utilized, researchers only survey a sample of the population, thus making it unlikely that the data collected will perfectly represent the attitudes, behaviors, or beliefs of a population. *Sampling error* refers to the extent to which data collected from the sample fail to accurately represent the population. If the sample is not representative of the population of interest, the behaviors reported, the attitudes expressed, and the beliefs documented might represent the perceptions of the respondents—but they will not be representative of the population of interest.

Creating the Cover Letter and Questionnaire: Measurement Error

Measurement error refers to the errors in reporting that are associated with issues in the survey instrument itself. In particular, researchers may encounter problems with question wording, implementation issues, problematic interviewing, and behaviors on the part of the respondents (e.g., the provision of socially desirable responses). The questionnaire itself (i.e., the vehicle through which researchers collect survey data) is perhaps the most influential component in measurement error. Even if a perfect sampling plan is implemented, it will be for naught if the questionnaire is flawed. Five aspects of the questionnaire can influence measurement error: the respondents' motivational survey state, the survey demand, the writing of the questions, the ordering of the questions, and the appearance of the questionnaire itself.

Motivational Survey State

The respondents' motivational survey state refers to whether the respondents feel that participating in a survey is a good investment of their resources (e.g., mental energy and time). The influence of participants' motivational state on their responses is best understood through the concept of the motivational life-bar. The motivational life-bar refers to the level of motivation that the respondent feels in regard to filling out the survey in a thoughtful and accurate manner and can be thought of as a finite amount of energy that each respondent has when completing a survey. The more motivated the respondents are to complete a survey, the higher their motivational life-bar. A major influence on the motivational life-bar is whether the respondents perceive it to be in their best self-interest to fill out a survey and to do so accurately. For example, survey takers would be more vested in accurately responding to a survey about a local restaurant than an eatery in a distant city that they will no longer visit. In one situation, the survey has possible benefits for the survey takers (i.e., a better local restaurant); in the other, its self-benefits are relatively limited (i.e., a better restaurant they will never visit). When a participant's life-bar is high, the participant is less likely to skip items and is more likely to provide effortful, accurate responses. When the life-bar is low, respondents are less likely to reread questions they do not understand, are more likely to guess or skip questions, and will be less likely to exert effort to ensure accurate responses are put forth. Survey takers' motivational life-bar can be influenced, positively or negatively, based on the communication that accompanies the survey (e.g., a motivating cover letter that explains how the survey impacts the survey taker), reminders of the importance of the survey placed throughout the questionnaire, the level of appreciation showed to the respondent throughout the survey (e.g., a statement of appreciation placed at the start of instructions), the appearance of the survey, and many of other facets of questionnaire design that will be discussed shortly.

Survey Demand

A second crucial component in regard to measurement error is survey demand, which refers to the level of difficulty associated with filling out the survey. A questionnaire that is lengthy and filled with complicated questions requesting extensive recalling of events will have a greater survey demand than a 3-item multiple-choice questionnaire. Survey demand can directly impact survey quality—if respondents become cognitively exhausted as a result of completing a questionnaire, they will have less effortful thought available for the remaining questions. Also, if the survey is highly demanding, respondents may choose to terminate participation. Other factors impacting survey demand include the difficulty of the questions, the cognitive capabilities of the respondents, and other questionnaire components (e.g., the writing of the items and appearance of the survey). A well-designed survey can decrease survey demand, whereas a poorly designed survey will increase it due to respondents exerting additional effort to understand the questions and interpreting the response sets. In addition, respondents' motivation to complete the survey will influence the extent to respondents' motivation to complete the survey will influence the extent to which survey demand affects survey quality. For instance, if respondents' motivational life-bar is high, they might accurately complete even the most demanding survey; however, if survey motivation to complete the survey is limited, even the slightest difficulty in responding could lead to survey termination.

Writing of the Questions

The writing of the survey questions is a third critical component when it comes to measurement error; it is imperative that the respondents' perception of the question matches the question as the survey writer intended to ask it. There are many errors that can be made in the construction of the survey items. For example, a survey writer must be sure to use simple rather than more complex words. If respondents encounter a word that they do not understand, they may just guess—thus reducing the validity of the data. Also, questions like “Do you like ice cream and cake?” can be problematic because they ask two questions simultaneously. Respondents will be confused as to whether the question is asking about their affection for ice cream and cake together or whether having a love for both the two desserts on an individual basis is in line with the spirit of the question. Survey writers must also take care not to ask respondents questions that are beyond their ability to answer (e.g., how many emails have you ever received?). Such confusion can frustrate respondents, reduce the motivational life-bar, and potentially negatively influence data quality.

Ordering of the Questions

The ordering of the questions is an additional element that can negatively impair survey data. A central consideration is that answers to one question can influence how respondents answer subsequent items in the questionnaire. The question context refers to the context in which respondents will answer the questions. Parts of the question context that influence respondents' answers to questions include the emotions that they are currently feeling and the information that is top of mind. For example, if respondents are asked about five things that cause frustration, it is likely they will be infused with at least some negative affect. If that occurs, and then the participant is asked about an educational policy, the respondent will likely be more negative toward the policy than if the survey taker had just written about five things that lead to happiness.

A complementary concern is how the ordering of questions can impact survey completion rate. For example, if a questionnaire begins with a very sensitive question, respondents may assume that all questions will be equally sensitive and difficult, thus leading them to terminate the survey. However, if the sensitive question is placed toward the end, after respondents have already exerted effort into the survey and the survey writer has built a rapport between the survey and the respondent, the respondent will be less likely to quit—and will certainly not assume that the entire survey is filled with similar highly sensitive questions.

Survey Appearance

By having a visually appealing, easy-to-navigate survey, the perceived credibility of the researcher will increase—thus increasing the respondent's motivation to answer accurately. The appearance of the survey can also have a positive or negative impact on the data in numerous ways. A well-designed survey can attract attention to the most important elements of a question by properly using shading and bolding or increasing font sizes. Likewise, the survey demand placed on the respondent can be minimized by taking actions such as making skip patterns easy to follow (e.g., skipping a follow-up question when it is not applicable), ensuring all scales are displayed in the same direction, and providing a clear visual path for respondents. Conversely, there are missteps that can negatively impact survey data. For example, if the survey is riddled with typographical errors, respondents may question the credibility of the research being conducted, thus reducing their motivation to provide thoughtful answers.

Data Collection: Nonresponse Error

Even if the population is properly selected, the sampling plan is well designed, and the questionnaire induces inspiration and thoughtfulness, nonresponse error can impair the validity of survey data. Nonresponse error occurs when people invited to complete the survey reject the invitation in a way that impairs the representativeness of the study. Nonresponse can occur if certain members of the population of interest were unable to fill out the survey, if the potential participants felt incapable of filling out the survey, or if there was a lack of willingness to complete the survey. If a segment of the population skips a particular question, the data will be an invalid representation of the population. As such, nonresponse introduces systematic bias into the data. A partial means of reducing the likelihood of errors due to nonresponse is to take steps to ensure that as many people from the sample complete the survey as possible. By taking steps such as personalizing invitations, proving incentives, shortening surveys, and sending reminders through mixed-mode approaches (e.g., calling potential participants for a mail survey), researchers can increase response rates.

A different form of nonresponse occurs when participants respond but then return the survey incomplete. Missing data refer to data that are missing from questionnaires of respondents who otherwise completed the survey. Data can be missing due to a respondent skipping a question (either intentionally or unintentionally), a response being provided but not recorded (i.e., instrumentation error), or the response being recorded is misplaced (e.g., data files being lost). A key concern when considering nonresponse error is the extent to which there are differences between those respondents with complete and incomplete data. Part of the decision-making process for handling missing data involves determining if data are missing completely at random (i.e., no relationship between the missing data and any other values, observed and unobserved), missing at random (i.e., systematic differences likely exist between the missing and observed values, but the missing data can be explained by the observed variables), or missing not at random (i.e., there is a relationship between data that are missing and unobserved data). A common fix for incomplete data is to remove participants' responses from the data set in their entirety; however, there are drawbacks to this approach, such as estimates of the population becoming less accurate. If everyone who supported a particular educational policy skipped a question due to finding it offensive, removing all those participants would likely cause the data to be nonrepresentative of the population of interest. Another means of dealing with nonresponse is imputation, which is the process of using statistical data to replace the data that are missing due to nonresponse. This is done by filling in the missing values with replacement values and then treating the data set as complete.

Costs Versus Benefits

Finally, it is essential to take into consideration the trade-offs that occur as a result of each decision made throughout the survey research process. For example, higher incentives can allow for a reduction in nonresponse error, but then it might not be possible to minimize sampling error. An effective survey researcher will weigh the benefits of any given approach with the costs in relation to all other opportunities for gains and losses.

Jason T. Siegel and Natalie D. Jones

See also Data; Nonresponse Bias; Order Effects; Quantitative Research Methods; Quota Sampling; Representativeness; Response Rate; Selection Bias; Simple Random Sampling; Snowball Sampling; Surveys; Weighting

Further Readings

Crano, W. D., Brewer, M. B., & Lac, A. (2014). Principles and methods of social research. Routledge.

Dillman, D. A. (2011). Mail and Internet surveys: The tailored design method—2007 Update with new Internet, visual, and mixed-mode guide. Wiley.

Lavrakas, P. J. (2008). Encyclopedia of survey research methods. Thousand Oaks, CA: Sage.

Cara N. Tan Cara N. Tan Tan, Cara N.

Jason T. Siegel Jason T. Siegel Siegel, Jason T.

Surveys

Surveys

1642

1645

Surveys

Although terminology and usage varies within and across domains, a *survey* is generally understood to be a form of data collection that relies on self-reported responses to a previously prepared set of questions. Self-report requires respondents to introspect (i.e., look inside themselves to gain an understanding of their own thoughts and feelings) and then provide a response to the survey question. Questions can cover a variety of topics, including past or present behavior (e.g., number of hours watching television in the past week), intentions for the future (e.g., intention to watch television in the next week), self-perceptions and other thoughts (e.g., attitudes toward watching television), and perceptions of others (e.g., how many hours friends spend watching television each week). Even though some surveys are as short as one or two questions, others are hundreds of questions long and can take hours to complete. Surveys can be conducted with paper and pencil, over the phone, or with electronics (e.g., computer programs or webpages). This entry reviews key characteristics as well as the advantages and disadvantages of surveys; it also discusses the use of surveys as part of a mixed methods research approach.

Comparison of Key Characteristics

Two key characteristics set surveys apart from other forms of data collection: self-reported responses and predetermined (i.e., structured) questions. These characteristics provide unique advantages, and some disadvantages, in comparison to other common forms of data collection, such as unstructured self-report, the observations of researchers, and physiological testing. As stated previously, self-report requires participants to think about and then report an answer to a question posed by researchers. This process is most noticeably different from physiological tests, which avoid asking for participants' introspections. For example, a survey measuring illicit drug use might directly ask participants to report the type and number of times they had used illicit drugs, while a physiological test might involve measuring the amount of illicit drugs in samples of participants' blood.

The second key characteristic of surveys is predetermined questions. Researchers using a survey must decide what topics should be measured before starting data collection and choose or create the best questions to assess those topics. For example, researchers doing a survey on mental health might first decide to measure symptoms of depression and then decide whether to use a preexisting set of questions (e.g., the Beck Depression Inventory) or develop new questions. This characteristic can be best understood when contrasted with unstructured self-report. Like surveys, unstructured self-report methods (e.g., unstructured interviews or focus groups) directly ask participants to respond to questions but do not follow a predetermined set of questions. Instead, researchers follow the flow of a participant's thoughts and spontaneously develop new questions to probe these responses in directions that might not have otherwise been anticipated.

Advantages of Surveys

One advantage of surveys is that they allow researchers to assess participants' thoughts more directly than observational methods or physiological tests, both of which require researchers to infer participants' thoughts. Although some physiological tests may be able to approximate a person's thoughts (e.g., live images of brain activity using functional magnetic resonance imaging), researchers are currently unable to know exactly what a person is thinking based purely on these measures. When using observation, researchers usually avoid directly asking questions to the participants and must instead infer participants' thoughts based on their behavior. In contrast, survey participants are directly asked to introspect and then report their thoughts, which are arguably more likely to be accurate than the researchers' inferences. Thus, surveys are particularly useful when the goal of a study is to measure participants' perceptions, attitudes, emotional state, or other thoughts.

Another advantage of using surveys to collect data is that researchers can collect a large number of responses more quickly than with other methods. There are two primary reasons surveys are less time-consuming. First, researchers can often collect information from a single participant more quickly using a survey than with observational, physiological, or unstructured methods. For example, a survey asking parents how much time they spent reading to their children in the past week would take only a few minutes, while observing the amount of time parents spent reading to their children over a week would take hours each day. Second, because surveys can be administered without the direct involvement of a researcher, potentially thousands of participants can be surveyed at once. For instance, a researcher could give copies of a pencil-and-paper survey to every teacher at an elementary school to collect from students during class; this would be impossible with other methods, which require the researcher to be present during data collection (e.g., researchers actively listening to participants' responses and developing new questions to further probe these responses for unstructured self-report).

A third advantage that surveys have over many other forms of data collection is that they are less costly to administer. Survey administration typically does not require intensive staff training (if any staff is necessary), whereas unstructured self-report, observation, and physiological tests all require researchers to carefully train anyone involved in data collection. A related benefit is that surveys are typically less invasive than other methods, making it easier and less expensive to recruit participants, and also decreasing the likelihood of causing harm to participants. For example, researchers would not need to compensate participants with as much money to incentivize taking a survey about diet and exercise than taking a blood test and treadmill stress test. In addition, surveys require relatively inexpensive resources (e.g., printing out pencil-and-paper copies or hosting a survey online), whereas physiological tests often require specialized equipment that can be very expensive.

Using surveys to collect data confers several additional advantages over other forms of data collection. For example, surveys have an advantage over unstructured self-report when researchers are interested in replicating or comparing their results with other studies, as the questions are standardized across all participants. Likewise, it is also much easier to aggregate and create statistical summaries of participants' data because they have all responded to the same questions. Compared to observational methods, surveys are more useful when researchers are interested in participants' past thoughts or behaviors, as observation of the past is only possible when recordings are available. In a similar vein, surveys are more useful than physiological tests when researchers are interested in participants’ past physical state, as researchers can only conduct physiological tests on participants in the present moment. Although participants’ self-reports of past physical states would merely be perceptions of reality, any physiological data from the past would have to have been recorded and made available for researchers to use.

Disadvantages of Surveys

Although the self-report nature of surveys provides advantages in directly assessing participants’ thoughts, researchers have also debated whether participants can accurately recount their own thoughts and experiences (i.e., introspect). Richard Nisbett and Timothy Wilson state that participants can be unaware of how the current context might affect their cognition, leading to biased self-reports. These scholars suggest that participants do not actually know why they think or behave in a certain way, but instead use a priori (i.e., previously developed) or plausible theories (e.g., following social norms) to come up with explanations. This does not mean that all introspections are inaccurate. Nisbett and Wilson suggest that researchers can increase the accuracy of self-report by reminding participants of the context of the specific behavior or thought of interest and by avoiding plausible explanations that might not have been part of the thought or behavior to be recalled.

Another disadvantage of surveys is that participants will sometimes respond in socially desirable ways, resulting in data that reflect what participants want researchers to believe rather than reality. For example, a researcher might be concerned about the accuracy of responses to a survey on adolescent truancy, as adolescents might be hesitant to admit whether and how often they skipped class. The threat of the error introduced by socially desirable responses can be reduced through several strategies. One means of doing so is to write the questions in such a way that the socially desirable response is unclear (e.g., “Some people love cigarettes, some people hate cigarettes, where do you fall on the spectrum?”). Another way of doing so is to ensure anonymity. This is particularly important in organization and educational settings where respondents might fear retribution. As will be discussed shortly, combining survey methods with other approaches, such as observational research, can often be the most useful way to reduce error.

A third disadvantage of surveys is that they are not as flexible as other methods of data collection. Survey methods are less flexible than observational methods, as researchers can more easily note any contextual information that might have influenced the behavior being observed. Surveys are also less flexible than unstructured self-report because questions are predetermined before responses are gathered. Unstructured self-report allows researchers to ask for clarification when responses are unclear. For example, if a respondent used the name of an individual, the researcher could ask how they were related (e.g., sister, mother, friend, or coworker). Researchers using unstructured self-report can also immediately delve deeper into participants' thoughts when unanticipated topics are mentioned. Although it is possible to recontact and follow up with survey participants, this process can take much longer than it would with unstructured self-report.

Surveys have several other disadvantages that must be considered. Although participants sometimes take a survey in the same room as one another, interaction between participants is typically discouraged, thus preventing group dynamics that could otherwise facilitate responses. For instance, interaction among participants can result in responses that can *snowball*, creating a chain of ideas or thoughts that might not have been triggered without the group. Surveys are also unlikely to provide data with pinpoint accuracy when researchers are interested in collecting the real-time information, as it is cumbersome to have participants constantly engage in self-report. Another disadvantage of surveys is that it is difficult to create questions that will lead to accurate responses. For example, questions can sometimes be interpreted by participants in different ways if not carefully worded. If a question asked participants about "how often" they studied on weekends, participants might not interpret the word "often" in the same way.

Using Surveys as Part of a Mixed-Methods Approach

Researchers are increasingly starting to use other forms of data collection to complement survey data, thus allowing them to capitalize on the advantages of surveys while overcoming some of the disadvantages. Using multiple methods to collect data, often called a mixed-methods approach, has several benefits. First, researchers can take advantage of a survey's efficiency while at the same time obtaining rich information through observation or unstructured self-report. Second, researchers can clarify complicated or unclear survey responses by following up with another method of data collection. Finally, researchers can validate their findings by looking for similar patterns of results (i.e., triangulation) in both survey data and data collected through other methods.

Cara N. Tan and Jason T. Siegel

See also Attitude Scaling; Content Analysis; Essay Items; Internal Validity; Interviews; Item Analysis; Item Development; Item Response Theory; Mixed Methods Research; Nonresponse Bias; Paper-and-Pencil Assessment; Reliability; Response Rate; Scales; Self-Report Inventories; Survey Methods

Further Readings

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). Asking questions: The definitive guide to questionnaire design—For market research, political polls, and social and health questionnaires. San Francisco, CA: Jossey-Bass.

Crano, W. D., Brewer, M. B., & Lac, A. (2014). Principles and methods of social research (3rd ed.). New York, NY: Routledge.

Dillman, D. A. (2007). Mail and internet surveys: The tailored design method (2nd ed.). New York, NY: Wiley.

Lavrakas, P. J. (2008). Encyclopedia of survey research methods. Thousand Oaks: Sage.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231–259.

Paulhus, D. L., & Vazire, S. (2009). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 224–239). New York, NY: Guilford.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. Annual Review of Psychology, 55, 493–518.

R. Shane Hutton R. Shane Hutton Hutton, R. Shane

Survival Analysis Survival analysis

1645

1649

Survival Analysis

Educational researchers are often interested in studying longitudinal processes such as college completion, student dropout, or teacher promotion. Survival analysis is a method of statistical modeling that allows researchers to analyze longitudinal data where the outcome is the time to an event of interest (e.g., time to graduation, time to dropout, and time to promotion). Two distinguishing features of survival analysis that separate it from traditional logistic regression analysis are the ability to naturally incorporate time into the model and the ability to handle incomplete data (i.e., censored data). Survival models can accommodate time measured continuously (continuous-time survival analysis) or time measured discretely (discrete-time survival analysis); however, in educational research, time is typically measured discretely (i.e., per semester, per academic year). For that reason, discrete-time survival analysis is the focus here.

Censoring

The most common form of censoring, right censoring, occurs when the outcome of an event of interest is unknown for an individual. That is, the event did not occur during the time frame under consideration or the individual leaves prior to the end of the study. Censoring results in missing data—that is, there is incomplete knowledge about occurrence or nonoccurrence of the event of interest for that individual. Omitting censored individuals from analysis will potentially bias the results. Survival analysis naturally allows for the incorporation of censored data.

Censoring is classified into two types: informative and noninformative. An important assumption of the survival analysis model is noninformative censoring (i.e., censoring that occurs for random and not systematic reasons). For example, treating dropouts as censored when studying college completion is most likely a violation of this assumption. Specifically, students who leave the college are likely to have systematically different characteristics than students who did not graduate during the time frame studied.

Hazard and Survivor Functions

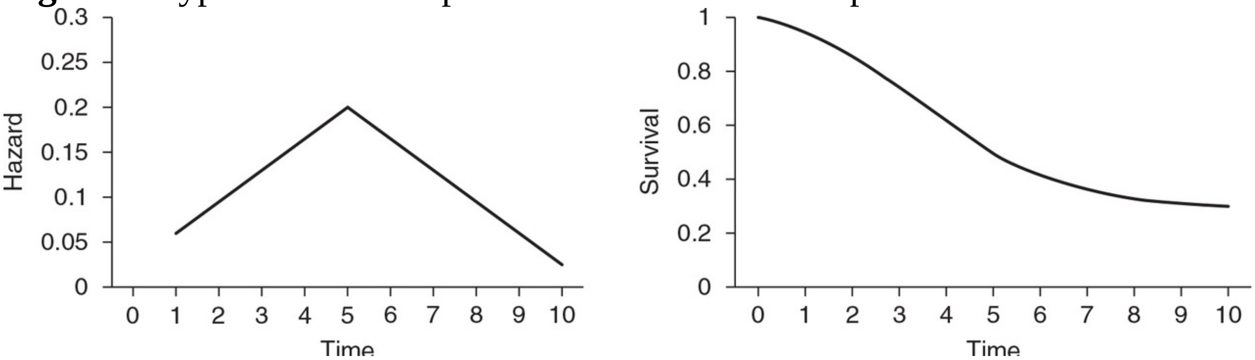
The hazard function consists of conditional probabilities that an individual will experience the event by a particular time period, given that the individual did not experience the event by a previous time period. These probabilities are sometimes called hazard probabilities or hazards. For a given time period, the hazard probabilities are computed by dividing the number of individuals who experienced the event by the number of individuals who were *at risk* of experiencing the event. Individuals at risk of experiencing the event are those who have not already experienced the event or who have not been censored for that time period.

The survivor function cumulates hazard probabilities across time and contains survival probabilities that represent the probability that an individual has not experienced the event, or survived the event, by a particular point in time. In many educational applications, it is actually the opposite of the survivor function that is of interest. For example, a researcher may be interested in the probability of a teacher being promoted; however, the survivor function gives probabilities of a teacher "surviving" their current status without promotion (i.e., remaining at the teacher's current rank). Subtracting the survivor probabilities from one will often give more informative probabilities, the probability of "not surviving."

Graphs of Survivor and Hazard Curves

Survivor and hazard probabilities can be plotted against time. The hazard plot shows the probability of an event for each point in time, whereas the survivor plot shows the probability that an individual has experienced the event by a particular point in time. Examples of hazard and survivor plots are given in Figure 1. The shape of the hazard plot will correspond to the values of the hazard probabilities at each time period; namely, the hazard plot may take on a variety of shapes depending on the hazard probabilities at each time period. For the example in Figure 1, the hazard probabilities increase to a particular time period and then decrease. Conversely, the survivor plot has a more distinctive shape. It is a decreasing function that begins at a survival probability of one because no one has experienced the event and decreases as individuals experience the event over time. The survivor plot gives a quick look at the proportion of individuals who have experienced the event at or before a particular time period.

Figure 1 Hypothetical example of hazard and survival plots



The Discrete-Time Survival Model

The discrete-time survival model is formulated using the hazard probabilities. Because the hazard probabilities range from 0 to 1, a logit (log odds) transformation is typically applied so the dependent variable is unbounded. The discrete-time survival model is written as:

$$\operatorname{logit} h_j = \alpha_j,$$

where α_j is a general representation of time (discussed more in the next section). This is called the baseline model because it does not contain any predictors; it models only the hazard function across time. The model considers only whether there is an effect of time.

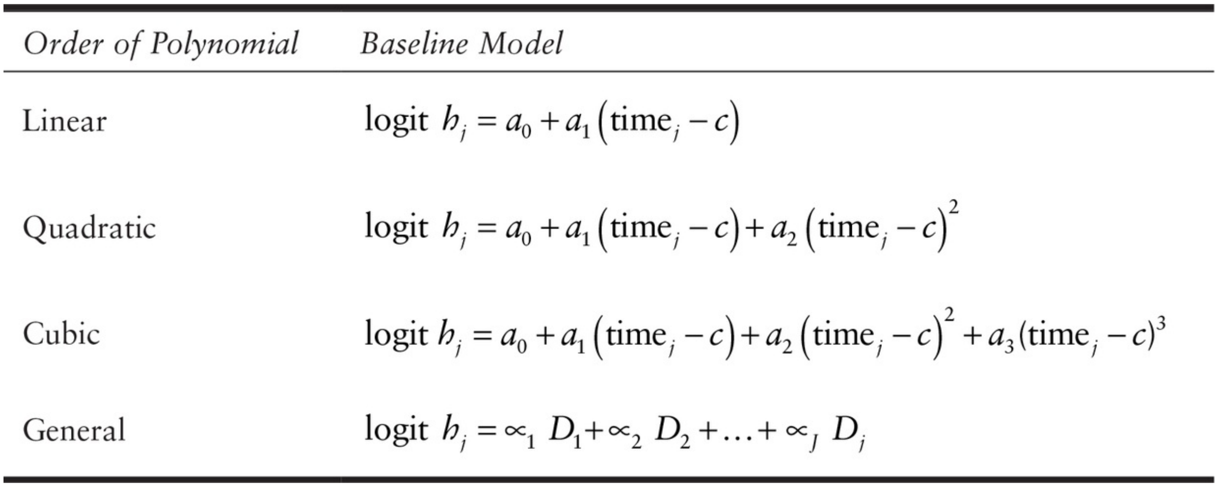
Representation of Time

Time, represented by $\alpha(j)$ in the aforementioned baseline model defined, is specified as either structured or unstructured. Structured time imposes constraints on the specification of time, whereas unstructured time does not impose any constraints on the specification. Unstructured time best represents the baseline logit hazard function and is specified by using dummy variables. Consider dummy variables, D_1, D_2, …, D_J, that represent the time periods j = 1, 2, …, J. Using this formulation, the baseline model is written as:

$$\text{logit } h_j = \alpha_1 D_1 + \alpha_2 D_2 + \ldots + \alpha_J D_j,$$

where α_j ($j = 1, 2, \ldots, J$) represents the intercept for each time period.

An alternative specification of time is to impose structure on the shape of the hazard function. This is typically performed for model parsimony when there are a large number of time periods or if the shape of the hazard function is theoretically known a priori. Structure is imposed using a polynomial specification of time (linear, quadratic, cubic, or higher). Table 1 shows the various specifications of time using a polynomial representation. The time variable is centered by subtracting a constant c; this makes a_0 more interpretable. Namely, a_0 represents the logit hazard for time period c. In addition, the nested structure of the polynomial orders allows for the researcher to perform a likelihood ratio test (i.e., the difference in deviance statistics for the nested models; discussed in a subsequent section) to compare models with different time structures.



Order of Polynomial	*Baseline Model*
Linear	$\text{logit } h_j = a_0 + a_1(\text{time}_j - c)$
Quadratic	$\text{logit } h_j = a_0 + a_1(\text{time}_j - c) + a_2(\text{time}_j - c)^2$
Cubic	$\text{logit } h_j = a_0 + a_1(\text{time}_j - c) + a_2(\text{time}_j - c)^2 + a_3(\text{time}_j - c)^3$
General	$\text{logit } h_j = \propto_1 D_1 + \propto_2 D_2 + \ldots + \propto_J D_j$

Incorporating Predictors

The survival model allows for the inclusion of both time-invariant and time-varying predictors. For example, in a college completion model, semester grade point average (GPA) is a time-varying predictor because GPA changes with each semester. However, high school GPA is a time-invariant predictor because it is not changing from semester to semester. Consider P time-invariant and Q time-varying predictors and let $X_1, X_2, \ldots, X_P$ represent the time-invariant predictors and Z_1, Z_2, …, Z_Q represent the time-varying predictors. The baseline model can be rewritten to include the predictors:

$$\text{logit } h_j = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \ldots \\ + \beta_P X_P + \kappa_1 Z_{1j} + \kappa_2 Z_{2j} + \ldots + \kappa_Q Z_{Qj},$$

where α_j represents the specification of time, the βs represent the slope parameters for time-invariant predictors and the κs are the slope parameters for the time-varying predictors. The effects of the time-invariant predictors are allowed to vary across time periods, but a proportional odds assumption can be invoked by constraining the effects to be the same across each time period.

Estimating the Model

Survival models can be estimated using the logistic regression procedure found in most statistical programs. It is important for the data to be formatted into a person-period format, which may be different than the original formatting of the data. A person-period format contains as many rows as time periods for each individual. The data set should contain a variable indexing each individual and a variable indexing the time period for each row. Columns for dummy-coded time variables or columns for the polynomial representation of time are needed. Finally, a variable identifying the time period during which the event occurred needs to be included. Time-invariant and time-varying predictors are included as additional columns in the data set.

Interpretation of Parameters

Parameter estimates in the survival model are interpreted similarly to those in a typical multiple regression model; however, the dependent variable is on the logit scale. Consider the survival model with dummy-coded time structure:

$$\text{logit } h_j = \alpha_1 D_1 + \alpha_2 D_2 + \ldots + \alpha_J D_j + \beta_1 X,$$

where *X* is a continuous predictor. The intercepts $\propto_j$ represent the value of the log odds for time period *j* when *X* is 0. The slope β_1 represents the change in log odds for a one-unit increase in *X*. In order to make the parameter estimates interpretable, the estimates can be transformed into odds ratios. Mathematically, this is accomplished by exponentiating the parameter estimates; most software will perform this calculation automatically. Thus, for each one-unit increase in *X*, the odds of *Y* is multiplied by eβ_1. Note that the intercepts $\propto_j$ are typically transformed into hazard probabilities.

Goodness of Fit

In most statistical software, the discrete-time survival model is estimated using maximum likelihood. The parameter estimates are determined by maximizing the log-likelihood function. Most software packages will output the −2 log likelihood, which is called the deviance statistic. The deviance statistic will always be positive, and smaller values represent a better fitting model. Deviance statistics are used not only to assess the fit of the model but also to compare models. A likelihood ratio test is performed to compare two nested models. Models are nested when one model includes at least one parameter more than the other model (i.e., one model is a subset of the other). Likelihood ratio tests can be used to assess the polynomial representations of time as well as competing models with different sets of predictors.

Extensions

The survival model can also be extended to include multivariate event histories; namely, the multiple spell model and competing risks model. The multiple spell model takes into account an event that can occur more than once (i.e., reoccurring events). For example, consider a model of student dropout; because students can potentially drop out multiple times (depending on school or university policy), a multiple spell model would be more appropriate than a single-event model. Alternatively, the competing risks model handles data where more than one event is possible, but only one can be experienced. That is, an individual is at risk of all events until one occurs. An application of this is college completion—that is, students may either graduate or depart (e.g., transfer and dropout) from the college.

R. Shane Hutton

See also Longitudinal Data Analysis; Multiple Linear Regression; Time Series Analysis

Further Readings

Allison, P. D. (2010). Survival analysis. In G. Hancock & R. Mueller (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 413–425). Routledge.

Allison, P. D. (2010). Survival analysis using SAS (2nd ed.). Cary, NC: SAS.

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. Psychological Bulletin, 110, 268–290.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics, 18, 155–195.

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY: Oxford University.

Tekle, F. B., & Vermunt, J. K. (2012). Event history analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskof, & K. J. Sher (Eds.), APA handbook of research methods in psychology: Data analysis and research publication (Vol. 3, pp. 267–290). Washington, DC: American Psychological Association.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. Review of Educational Research, 61(4), 407–450. Retrieved from https://doi.org/10.3102/00346543061004407

Breanna A. Wakar Breanna A. Wakar Wakar, Breanna A.

Dmitriy Poznyak Dmitriy Poznyak Poznyak, Dmitriy

Systematic Sampling Systematic sampling

1649

1650

Systematic Sampling

Systematic sampling (also called interval sampling) is a probability sampling technique that selects population elements at fixed intervals. This sampling method can be used when there is an available list of all elements of the population of interest or when a convenience sample is selected using fixed intervals (also called a flow sample, e.g., selecting every 10th student entering a school building). Systematic sampling is also practical when the sampling frame covers a well-defined spatial area (e.g., selecting every fifth room from a college dormitory). In education research, systematic sampling could be used to sample students from rosters within schools or to select parents for a survey using an administrative list of e-mail addresses, among other situations. This entry reviews principles of systematic sampling, the relationship between systematic sampling and select other sampling methods, and practical considerations for implementation.

Basic Principles and Estimation Procedures

When an ordered list of population elements is available to serve as a sampling frame (e.g., a list of all students in a school), systematic sampling is straightforward. In this case, every kth element from the sampling frame is selected, starting at an element chosen at random from the first k. First, determine a sampling interval, $k = N/n$, where n is the desired sample size and N is the number of elements in the population. For example, to select a sample of $n = 100$ students from $N = 1,315$ students, $k = 13.15$. Second, generate a random start between zero and k to determine the first element of the population to be sampled (can be done using Microsoft Excel, dedicated statistical software, or online tools; often these generate numbers between zero and one, in this case multiply by *k* such that the result is between zero and *k*). Third, add *k* to the random start repeatedly. For example, if 5.16 is the random start, numbers generated will be: 5.16, 18.31, 31.46, 44.61, …, 1307.01. Finally, round the numbers to integers, which are the list positions of the sampled elements: 5, 18, 31, 45, …, 1307. This procedure will result in selecting exactly 100 individuals at an approximately equal interval (exactly equal intervals if *k* is an integer). Note that if the random start is less than 0.50, rounding to an integer will yield zero, and the first list position will be selected after the first time *k* is added. In the case of a flow sample where a sampling frame is unavailable, establish a sampling interval based on the desired proportion of the population to be sampled.

The sample mean, , where y_i is the value of the outcome of interest for the *i*th sampled element, is an unbiased estimate of the population mean. With a randomly ordered population list, the population variance can be estimated by . However, nonrandom list orders (discussed later in this entry) require more complex variance estimates.

Relationship to Other Sampling Methods

If the population list order is random, systematic sampling has similarities to simple random sampling. However, using simple random sampling, any set of *n* population elements has the same chance of being selected. Many simple random samples will never be selected by systematic sampling (e.g., any sample including two adjacent population elements). The systematic sampling random start determines the rest of the sample, rather than each selection being made independently of all other selections as in simple random sampling. That being said, the probability of selection for each population element is equal for simple random and systematic sampling, 1*k*.

Systematic sampling can be thought of as a special case of stratified random sampling. Essentially, systematic sampling divides the population into sampling intervals (sometimes called zones), which are analogous to strata: The first zone contains the first *k* population elements, the second zone contains the next *k*, and so on. The systematic sample includes one element from each zone, although systematic sampling selects the element from the same relative position in each zone, which is not necessarily the case in a stratified random sample of one individual per stratum. An alternative to systematic sampling is a sequential sampling technique developed by J. R. Chromy, which uses zones as defined by systematic sampling but makes independent selections within each zone, eliminating the risk of nonrepresentative sample selection due to periodic list order.

Systematic sampling can also be combined with other sampling schemes, including explicit stratification and sampling with probability proportional to size.

Practical Considerations

If the list is ordered by an auxiliary variable that is related to key population subgroups, this will implicitly stratify the sample, increasing the likelihood that the sample will include elements with a range of values of the outcome of interest, in proportion to their representation in the population. Ordering the list by an auxiliary variable can also reduce sampling variation (the extent to which different samples produce different estimates) of the estimates of population parameters. Ideally, the variables used for list ordering are available prior to sampling for all population elements, such as student administrative data. Ordering variables that are closely related to the outcome of interest will yield the greatest reduction on sampling variation; when studying student test scores, a student’s previous test score may be a more useful ordering variable than the number of days the student was absent from school.

If the frame has a periodic order (e.g., a list of students ordered by homeroom and alphabetically within each homeroom), selecting a sample that is representative of the population requires a sampling interval that does not correspond to the periodic pattern. In the case of the student homeroom list, a sampling interval equal to the number of students in each homeroom will select only students from the same part of the alphabet. When the sampling frame is a spatial area, such as doors (rooms or apartments) along a building corridor, care must be given to determining list order to avoid such periodicity.

Common statistical software packages, including SAS and R, have built-in capabilities for selecting systematic samples. In cases where a population list exists, systematic sampling is easily executed using such software.

Breanna A. Wakar and Dmitriy Poznyak

See also Representativeness; Sample Size; Simple Random Sampling; Stratified Random Sampling; Variance

Further Readings

Bethlehem, J. G. (2009). Applied survey methods: A statistical perspective. Hoboken, NJ: Wiley.

Chromy, J. R. (1979). Sequential sample selection methods. Proceedings of the Section on Survey Research Methods (pp. 401–406). Washington, DC: American Statistical Association.

Kish, L. (1965). Survey sampling. New York, NY: Wiley.

Levy, P. S., & Lemeshow, S. (2008). Sampling of populations: Methods and applications. Hoboken, NJ: Wiley.

T

Jeanine Romano Jeanine Romano Romano, Jeanine

T Scores

T Scores

1651

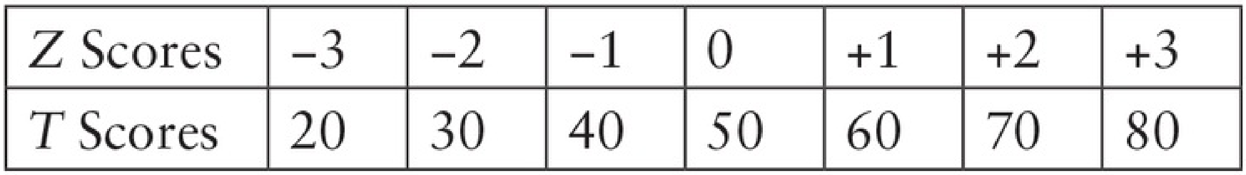
1651

T Scores

One of the most common standardized scores is T scores. Like Z scores, T scores are normally distributed and allow for consistent interpretation of a student's relative performance on an individual test and across tests. While Z scores have a mean of 0 and a standard deviation of 1, T scores have a mean of 50 and a standard deviation of 10, where the mean and standard deviation are held constant.

To convert any raw score into a T score, first transform the score into a Z score and then use the following basic formula: $T = 10z + 50$. To calculate the T score for a given score, it is important to first convert the raw score to a Z score, so that the Z score = (observed score − mean)/standard deviation.

For example, if a student's raw score is converted to a Z score, and if the student's Z score is 1, it is interpreted that the student scored 1 standard deviation above average at approximately the 84th percentile. T scores are similar when a T score of 60 would also be 1 standard deviation above the mean or approximately the 84th percentile. Because T scores are positive, it is fairly easy to report a student's score or scores. So instead of saying an individual had a Z score of −1, the equivalent would be that the individual had a T score of 40. Table 1 shows Z scores and their corresponding T scores:



Z Scores	–3	–2	–1	0	+1	+2	+3
T Scores	20	30	40	50	60	70	80

1

Jeanine Romano

See also Normal Distribution; Standardized Scores; Z Scores

Further Readings

Ebel, R. L. (1972). Essentials of educational measurement.

Linn, R. L. (1993). Educational measurement (American Council on Education Series on Higher Education). Phoenix, AZ: Oryx Press.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. Educational Measurement, 3, 221–262.

Olga Korosteleva Olga Korosteleva Korosteleva, Olga

Brian Song Brian Song Song, Brian

t Tests

t tests

1651

1654

t Tests

The *t* test is a statistical hypotheses in which the test statistic utilizes the *t* distribution (otherwise known as Student’s *t* distribution) to define the test conclusion. This test is used when the sample size is small and the distribution appears normal. The *z* test is very similar, except that it requires knowledge of the population variance (σ^2). Without knowing the true variance, the sample variance is used to approximate the population variance, but the *t* test must be used. Although the *t* distribution is generally used with unknown variation, the *t* test need not be used for larger sample sizes. As the sample size increases, the sample variance converges to the population variance, allowing the use of standard normal distribution for testing. In general, a sample size of 30 is used in distinguishing a small sample from a large sample.

William Sealy Gosset, a chemist for the Guinness Brewery in Dublin, Ireland, created the *t* test as an economical way of testing the quality of the stout beer. He submitted his work to the journal *Biometrika*, which was published in 1908. Because of the company’s policy restricting its chemists from publishing their findings, Gosset published his work under the pseudonym “Student.” The efficiency and accuracy of his method led to its wide usage in statistical hypothesis testing.

Depending on the setting, the *t* test can be approached differently. A one-sample test and two-sample test are used for either matched pairs, independent samples with equal standard deviation (*SD*), or independent samples with unequal *SD*s.

One-Sample *t* Test

A scenario for a one-sample *t* test can be formed as follows. For example, a student looking into gardening finds a source online that claims its tomato seeds will yield at minimum 30 tomatoes per plant, on average. Questioning such a claim, the student purchases the seeds to test the validity of the company's statement. Because the students' backyard has limited space, only 12 seeds could be planted in separate locations with similar conditions. After a couple of months, the tomatoes on the vines begin to ripen and the student records the data. The sample mean x comes out to 28.5 tomatoes with a sample *SD* (s) of 2.19 tomatoes. The student states the null hypothesis as the average tomato yield per plant to be $\mu = 30$, and the alternate hypothesis as $\mu < 30$. This way, if the student rejects the null, there is sufficient information that suggests a contradiction to the company's claim. The student also sets the level of significance at $\alpha = .05$. The student obtains the test statistic (t) through the equation , where μ_0 is the numeric value that the mean is being compared to (in this case, it is 30), and $n = 12$ is the sample size, which comes out to $t = -2.37$. Using the t table, the student calculates the critical value ($t_{0.05}$) at 11 degrees of freedom ($n - 1$) which is −1.796. The critical value is a predetermined area under the density curve to the right of it. The degrees of freedom is the size of the sample minus the number of estimated distribution parameters. In this instance, the population *SD* is the only parameter estimated from the data, thus the number of degrees of freedom is reduced by one. Further, because the test statistic falls in the rejection region ($t \leq t_{0.05}$), the student rejects the null and concludes that there is sufficient evidence against the company's claim.

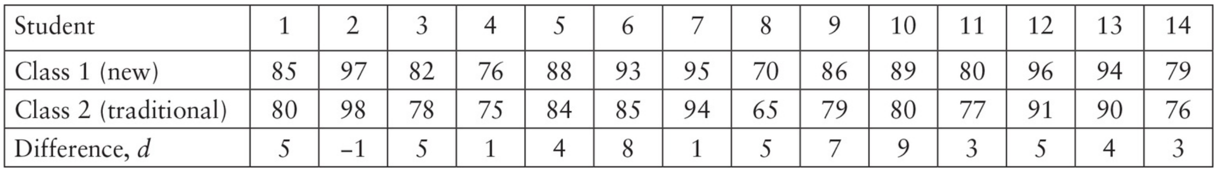
In a general setting, a one-sample t test for mean μ tests the null hypothesis H_0: $\mu = \mu_0$ against an upper tailed alternative hypothesis H_1: $\mu > \mu_0$, a lower tailed H_1: $\mu < \mu_0$, or a two-sided H_1: $\mu \neq \mu_0$. Assuming H_0 is true, the test statistic, , has a t distribution with $n - 1$ degrees of freedom. The null hypothesis is rejected if the test statistic falls in the reject region: $\{t > t\alpha\}$ for an upper tailed H_1, $\{t > -t\alpha\}$ for a lower tailed H_1, and for a two-tailed H_1.

Two-Sample Matched-Pair *t* Test

Two-sample *t* test can be processed differently based on the given samples. If each element in a random sample of a sample set is related to exactly one element in the second sample, and if the samples are of equal sizes, then the two samples are called paired data (or matched-pair data). By definition, these two data sets are dependent. Paired data can be used for hypotheses testing under the name matched-pair *t* test.

Let's assume a teacher wants to implement a new teaching method. The teacher manages to get two sets of 14 students, where each student in one group shares similar IQ (or other measure of intelligence) with another student in the second group. This way, the teacher can measure the difference between the similar pairs to see if there was a significant increase in knowledge through the new teaching method. The teacher teaches the same topics to each class, but one is taught with the new method and the other is taught the traditional way. The teacher gives the same test to the students at the end of the session and records the scores. The result is on the top of the next page.

The teacher subtracts the two measurements for each matched pair and uses the difference for test and sets the significance level α at .05. The mean of the difference (*xd*) is 4.214, with the *SD* of 2.751. The null hypothesis for this test states that $\mu d = 0$, and the alternative hypothesis is that $\mu d > 0$, where μd denotes the true population mean of the difference. To reject the null, there needs to be enough evidence from the data to support that the new method is significantly better than the traditional method on average. Note that in this case, $\mu_0 = 0$ and $n = 14$. Using the same test statistic as for one-sample *t* test, we obtain . Using $t_{0.05} = 1.771$ at $n - 1 = 14 - 1 = 13$ degrees of freedom, the teacher finds that the observed test statistic $5.733 > t_{0.05}$ falls in the rejection region of the test; thus, rejecting the null hypothesis and stating that there is a significant enough difference to say that the new method has a stronger impact on the students.



Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Class 1 (new)	85	97	82	76	88	93	95	70	86	89	80	96	94	79
Class 2 (traditional)	80	98	78	75	84	85	94	65	79	80	77	91	90	76
Difference, *d*	5	–1	5	1	4	8	1	5	7	9	3	5	4	3

The next example shows that the matched-pair test may be applied to one sample with two observations per sample element. That is, the two dependent samples may in fact be the same sample, where each element would play the role of a matched pair. For instance, consider the following situation. A kinesiology student is interested in finding out whether their left hand is, on average, stronger than their right hand for left-handed people. He rounds up seven left-handed students, randomly picked from the campus, and has them pull weights. He records the number of pounds pulled by each hand and, for each student, computes the difference between the left and right hands. The differences obtained are 23, 5, 12, 4, −13, 16, and 7. The kinesiology student then employs the matched-pair *t* test to test whether $\mu d > 0$ (the alternative hypothesis). The null hypothesis in this case is H_0: $\mu d = 0$. He computes $xd = 7.714$, $SD = 11.339$, $n = 7$, $\mu_0 = 0$, and hence the test statistic becomes . The critical value for the rejection region that corresponds to $n - 1 = 7 - 1 = 6$ degrees of freedom and .05 level of significance is $t_{0.05} = 1.943$. Therefore, the test statistic falls outside of the rejection region ($t = 1.800 < t_{0.05}$), and as the student concluded from this test, the data failed to support his supposition for left-handed people, regarding left hand being on average stronger than the right one. Going back to the measurements, we can see that the left hand was stronger for all but one respondent, who, as it turned out, “spoiled the whole thing.”

Two-Sample *t* Test for Independent Samples

Another situation in which a *t* test is applicable is when one wants to compare μ_1 and μ_2, the true means of two independent populations. Two cases are distinguished: Population *SD*s are assumed equal or unequal. Generally speaking, populations have distinct *SD*s, but in some special cases, these *SD*s may be considered equal. For example, if the two populations are very similar, or in fact are the same, population before and after some intervention could have affected the means but not the *SD*s.

Suppose two samples of sizes n_1 and n_2 are available. The respective sample means and *SD*s are x_1, x_2, s_1, and s_2. When the population *SD*s are assumed equal, both samples are pooled together to obtain a more precise estimate of this common *SD*, resulting in the estimate , which is called the pooled *SD*. The test statistic then can be found with the equation , which under the null hypothesis H_0: $\mu_1 = \mu_2$ has a *t* distribution with $n_1 + n_2 - 2$ degrees of freedom. If the *SD*s of the two samples cannot be assumed equal, then the equation for the test statistic is . The number of degrees of freedom in this case is computed as the largest integer not exceeding .

Imagine a scenario to test whether the average reaction times differ between males and females. The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is two sided, $H_1:\mu_1 \neq \mu_2$. To measure the reaction time, a sample of 15 male and 15 female participants were asked to press a button when the light shows up on a machine. The men's average reaction time (x_1) was recorded as 0.38 seconds and the women's average reaction time (x_2) was 0.35 seconds. The sample *SD*s of male (s_1) and female (s_2) reaction times were recorded at 0.12 and 0.14 seconds, respectively. Because there is no evidence that the population *SD*s for men and women would differ, they are assumed equal. The pooled *SD* is calculated as and the test statistic as . The critical value for *t* distribution with $n_1 + n_2 - 2 = 15 + 15 - 2 = 28$ degrees of freedom that has the area of 0.025 to its right is $t_{0.025} = 2.0481$. The observed test statistic is smaller than the critical value, so we can say that we fail to reject the null. The conclusion would be that there is no sufficient evidence that male and female average reaction times are different.

The next example considers the case of unequal population *SD*s. Let's say a researcher wants to test the difference between Tylenol and a generic brand from a drug store. The researcher hypothesizes that Tylenol activates faster than the generic brand and administers Tylenol and a generic brand to two groups of 10 people each, wherein two people of one group experienced nausea and were dismissed from the study, thus resulting in two groups of sizes $n_1 = 10$ and $n_2 = 8$. The mean activation time in the Tylenol group (x_1) was recorded to be 21 minutes with the *SD* (s_1) of 3 minutes. The group that received the generic brand had the mean activation time (x_2) of 33 minutes and an *SD* (s_2) of 10 minutes. The researcher tests a one-tailed alternative $H_1: \mu_1 < \mu_2$ and assumes unequal population *SD*s. He then finds the test statistic . The number of degrees of freedom for this test is the largest integer less than and thus equal to 8. The critical value $t_{0.05}$ that corresponds to 8 degrees of freedom is 1.8595. Because the test statistic is smaller than −1.8595, the researcher rejects the null hypothesis and concludes that Tylenol acts faster than the generic brand.

Limitations

Implementation of *t* tests is restricted to normally distributed observations with unknown *SD* (which has to be estimated from the data) and a small sample size (not more than 30 per sample). Also, *t* tests can accommodate at most two samples.

Olga Korosteleva and Brian Song

See also Hypothesis Testing; Normal Distribution

Further Readings

Wackerly, D. D., William, M., & Richard, L. S. (2008). Mathematical statistics with applications (7th ed.). Belmont, CA: Brookes/Cole.

Walpole, R. E., Raymond, H. M., Sharon, L. M., & Keying, E. Y. (2011). Probability and statistics for engineers and scientists (9th ed.). Pearson.

Helenrose Fives Helenrose Fives Fives, Helenrose

Nicole Barnes Nicole Barnes Barnes, Nicole

Table of Specifications Table of specifications

1654

1657

Table of Specifications

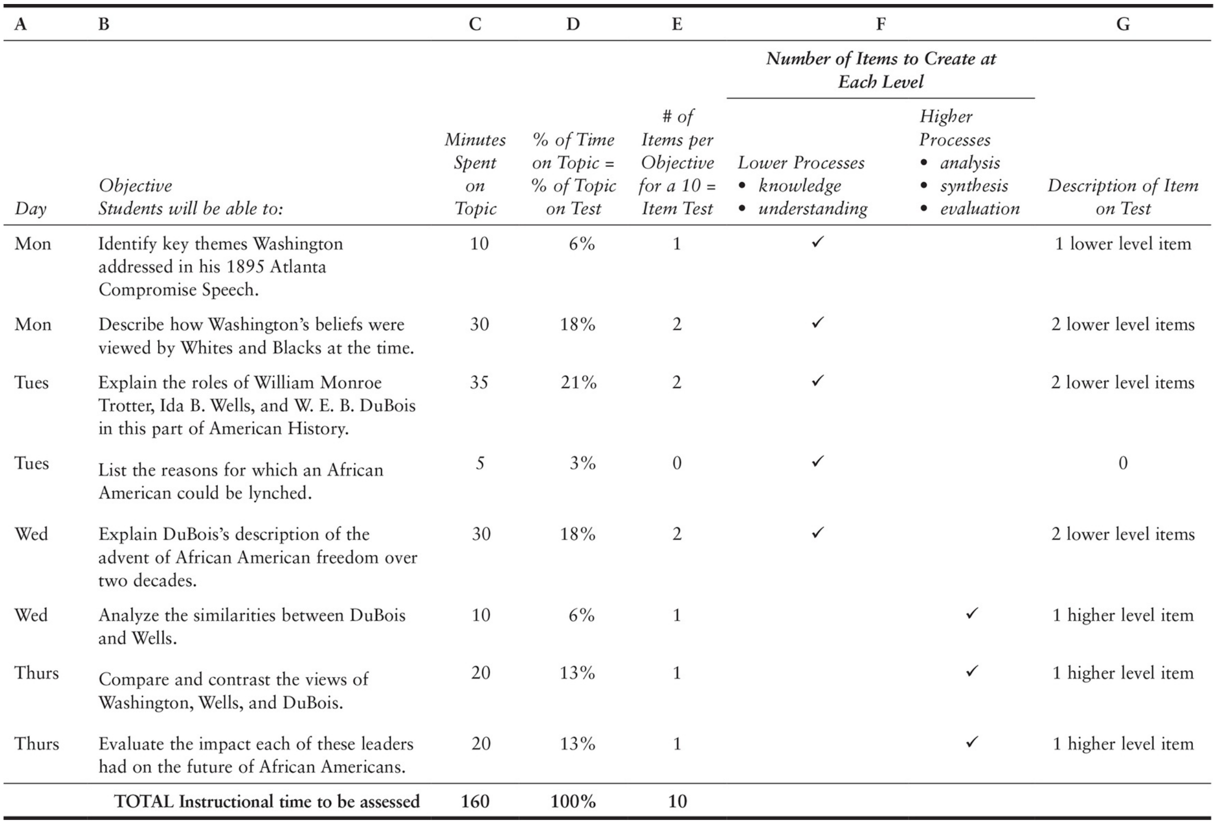
The table of specifications (TOS) is a tool used to ensure that a test or assessment measures the content and thinking skills that the test intends to measure. Thus, when used appropriately, it can provide response content and construct (i.e., response process) validity evidence. A TOS may be used for large-scale test construction, classroom-level assessments by teachers, and psychometric scale development. It is a foundational tool in designing tests or measures for research and educational purposes.

The primary purpose of a TOS is to ensure alignment between the items or elements of an assessment and the content, skills, or constructs that the assessment intends to assess. That is, a TOS helps test constructors to focus on issue of response content, ensuring that the test or assessment measures what it intends to measure. For example, if a teacher is interested in assessing the students’ understanding of lunar phases, then it would be appropriate to have a test item asking them to draw the phases of the moon. However, a test item asking them to identify the first person to walk on the moon would not have the same content validity to assess students’ knowledge of lunar phases.

In addition, a TOS can also be used to provide response process validity evidence for test constructors. Response process refers to the kind of thinking that is expected of the test taker in completing the assessment. For the lunar phases, for example, a teacher may expect students to memorize the phases of the moon and therefore a knowledge-level (relying on recognition or memory) question would be appropriate. Alternatively, if the teacher taught the lessons such that students tracked the moon for a month, developed lunar journals, and discussed the reasons for the different phases, then the assessment should target discussed the reasons for the different phases, then the assessment should target higher level thinking such as analysis, evaluation, and synthesis. As such, asking students to draw a model of the lunar phases with annotated explanations would be better aligned to the kind of thinking that students experienced during instruction.

The TOS is typically constructed as a table that includes key information to help teachers align the learning objectives that represent the content and cognitive levels intended for students to achieve with class time spent and the number of test items. Table 1 provides an example of a TOS for a chapter test on “New Ideas for a New Century,” from Molefi Kete Asante’s (1995) *African American History: A Journey of Liberation*. This entry explored the roles of prominent African American leaders from 1895 to 1919. Before constructing the TOS, the teacher decided the total number of items to include (i.e., 10) and quantity and type of those items (i.e., five multiple-choice and five short answers), and the decision was made based on the time allocated for students to complete the test and students’ general test-taking abilities. Next, the teacher referred to the lesson plans and notes to determine the content in columns A–C (i.e., day, learning objectives, time spent on objective). To calculate the percentage of class time for each objective (column D), the teacher divided the minutes spent teaching each objective (column C) by the total minutes for the unit and multiplied by 100. Determining the percentage of time spent in class on each objective is one approach to identifying how many items on the test should address any particular objective and enhances test content validity evidence.

Next, the teacher multiplied the percentage of time on topic (column D) by the total number of items on the test (10) to determine the number of items needed to measure each objective. Note that the teacher rounded to whole numbers when appropriate. In some instances (see Objective 4), none of the test items was used to assess that objective. In other words, not enough instructional time was spent teaching that content to justify assessing it on the unit test. Column F shows the classification whether each objective measured lower or higher order thinking processes. Lower level thinking processes require students to remember or understand, whereas higher level thinking processes requires students to apply, analyze, synthesize, and evaluate. Finally, with the information in columns E and F, the teacher determines the information in column G. Recall that prior to TOS construction, the teacher decided that both multiple-choice and short-answer items would be distributed evenly. The teacher used knowledge of the content and cognitive level along with professional judgment to determine the best one for each item.



A	B	C	D	E	F		G
					Number of Items to Create at Each Level		
Day	*Objective* *Students will be able to:*	*Minutes Spent on Topic*	*% of Time on Topic = % of Topic on Test*	*# of Items per Objective for a 10 = Item Test*	*Lower Processes* • *knowledge* • *understanding*	*Higher Processes* • *analysis* • *synthesis* • *evaluation*	*Description of Item on Test*
Mon	Identify key themes Washington addressed in his 1895 Atlanta Compromise Speech.	10	6%	1	✓		1 lower level item
Mon	Describe how Washington's beliefs were viewed by Whites and Blacks at the time.	30	18%	2	✓		2 lower level items
Tues	Explain the roles of William Monroe Trotter, Ida B. Wells, and W. E. B. DuBois in this part of American History.	35	21%	2	✓		2 lower level items
Tues	List the reasons for which an African American could be lynched.	5	3%	0	✓		0
Wed	Explain DuBois's description of the advent of African American freedom over two decades.	30	18%	2	✓		2 lower level items
Wed	Analyze the similarities between DuBois and Wells.	10	6%	1		✓	1 higher level item
Thurs	Compare and contrast the views of Washington, Wells, and DuBois.	20	13%	1		✓	1 higher level item
Thurs	Evaluate the impact each of these leaders had on the future of African Americans.	20	13%	1		✓	1 higher level item
	TOTAL Instructional time to be assessed	160	100%	10			

*MC: multiple choice item/ SA: short answer item

Helenrose Fives and Nicole Barnes

See also Alignment; Bloom's Taxonomy; Classroom Assessment; Construct-Related Validity Evidence; Content-Related Validity Evidence; Curriculum-Based Assessment; Instructional Objectives; Item Analysis; Multiple-Choice Items; Standards-Based Assessment; Tests; Validity

Further Readings

Barnes, N., & Dacey, C. M (in press). Using traditional assessments to effectively inform your teaching. In J. Grinberg & D. Schwarzer (Eds.), Successful teaching: What every novice teacher needs to know. Rowman & Littlefield.

DiDonato-Barnes, N. C., Fives, H., & Krause, E. (2013). Using a table of specifications to improve teacher constructed traditional tests: An experimental design. Assessment in Education: Principles, Policy, and Practice, 21(1), 90–108. doi:10.1080/0969594X.2013.808173

Fives, H., & DiDonato-Barnes, N. C. (2013). Classroom test construction: The power of a table of specifications. Practical, Assessment, Research, and Evaluation, 18(1). Retrieved from http://pareonline.net/pdf/v18n3.pdf

Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. Journal of Instructional Psychology, 31, 115–129.

Teacher Certification

See Certification

Patricia A. Jenkins Patricia A. Jenkins Jenkins, Patricia A.

Teacher Evaluation

Teacher evaluation

1657

1658

Teacher Evaluation

Teacher evaluation is a systematic, ongoing process used to assess teachers' competence, performance, and effectiveness in the classroom. Teacher evaluations also include an assessment of data to determine appropriate pathways for developing highly skilled teachers. Systematic evaluations give evidence of teachers' performance, address teachers' accountability, and influence professional development. This entry describes the procedures for teacher evaluations and how teacher evaluations are aligned with professional growth.

In most states, school districts evaluate teachers at least annually. In some states, school districts evaluate nontenured teachers twice a year. The evaluator is assigned by the school district and is usually the school principal. The procedures for teacher evaluation occur in several phases. First, there is a preconference between the teacher and the evaluator. The preconference or preobservation phase serves to provide a conceptual framework for the actual observation. During the preconference, date(s) are set for classroom observation(s) and the postconference, and the evaluation tool and expectations are discussed, with time allotted for questions. A teacher's self-assessment may be part of the preconference and/or the postconference.

The next step or phase is the classroom observation. The evaluator observes the teacher using the framework and assessment tool discussed in phase one. For the next phase, the evaluator organizes and analyzes data from the observation; data include documented artifacts relevant to selected domains prior to the scheduled observation, such as lesson plans, students' test results, classroom management style, and nonteaching responsibilities. The evaluator drafts a formal assessment style, and nonteaching responsibilities. The evaluator drafts a formal assessment of the teacher to provide an overall evaluation and a plan for continuous professional development.

A postconference follows with the teacher and evaluator to have a dialogue about the data and the formal assessment. The dialogue includes a discussion about the teacher's effectiveness, the teacher's self-assessment of the class the evaluator observed, and implications for professional growth based on all data collected. Elements for improvement are discussed and selected with specific growth-target dates.

The professional growth plan is a part of the systematic teacher evaluation. Evidence from the teacher evaluation informs the professional growth or professional development plan, which provides opportunities based not only on the teacher's individual needs but also those of the students, school, and district. In addition, the plan is aligned with the state's standards for curriculum and instruction, the assessment tools used to gauge students' achievements, and professional learning standards.

Patricia A. Jenkins

See also Certification; Framework for Teaching; Personnel Evaluation; Portfolio Assessment; Professional Development of Teachers

Further Readings

The Danielson Group. (2013). The framework. Retrieved from https://www.danielsongroup.org/framework/.

Gulamhussien, A. (2013). Teaching the teacher: Effective professional development in an era of high stakes accountability. Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Teaching-the-Teachers-Effective-Professional-Development-in-an-Era-of-High-Stakes-Accountability/Teaching-the-Teachers-Full-Report.pdf

Learning Sciences International. (n.d.). Marzano teacher evaluation model. Retrieved from http://www.marzanoevaluation.com/evaluation/causal_teacher_evaluation_mo

Marzano, R., Frontier, T., & Livingston, D. Effective supervision, Chapter 2. A Brief History of Supervision and Evaluation. Retrieved from http://www.ascd.org/publications/books/110019/chapters/A-Brief-History-of-Supervision-and-Evaluation.aspx

Patricia A. Jenkins Patricia A. Jenkins Jenkins, Patricia A.

Teachers’ Associations

Teachers’ associations

1658

1659

Teachers’ Associations

Teachers’ associations are professional organizations that provide a variety of services for teachers and advocate on teachers’ behalf. As many use the terms *teacher* and *educator* interchangeably, this entry includes associations that serve those other than classroom teachers, such as school counselors.

The best-known teachers’ associations are the two major teachers’ unions, the National Education Association and the American Federation of Teachers. Both work to influence state and national policies, to link research to teacher practice, and to set professional standards. Both have local chapters that engage in collective bargaining with school districts and postsecondary institutions to negotiate contracts that govern members’ salaries and working conditions. Membership dues vary and benefits tend to be more extensive than those of other types of teachers’ associations.

Discipline-specific associations support teachers’ professional development/growth in their field of specialty and are a voice for teachers. They center on subject areas such as language arts, social studies, science, math, and music. The subject areas are usually denoted in their titles. For example, the Music Teachers National Association, the National Council of Teachers of Mathematics, the National Science Teachers Association, the National Council for the Social Studies, and the International Literacy Association are all discipline-specific organizations. Almost all subjects taught have a professional association. Although not exhaustive, other examples include the National Council of Teachers of English, National Art Education Association, American Council on the Teaching of Foreign Languages, and the Association for Education in Journalism and Mass Communication.

Education in Journalism and Mass Communication.

Membership fees and benefits vary among these groups. Nearly all have websites with resource tools for teachers, such as lesson plans. Some also offer discounts from various merchants. In addition, information about research, issues, and trends in the discipline is available. Information on grants, employment openings, and professional development programs also can often be found on these associations' websites. Moreover, the organizations sponsor conventions and conferences, some with programs that allow teachers to receive continuing education units. Continuing education units are a measure recognizing participation in formal, noncredit education courses that may be applied to certification renewal. The conventions and conferences also provide members networking opportunities as well as exhibits from educational vendors. Almost all discipline-specific associations offer publications, such as professional journals, books, magazines, and newsletters.

Another type of organization involves teachers of particular grade levels or those who teach children with specific types of needs. Examples include the National Association for the Education of Young Children, the National Middle School Association, and the Council for Exceptional Children. As with discipline-specific organizations, membership fees vary. Many have similar benefits to the discipline-specific organizations.

Some associations cater exclusively to professionals in education whose responsibilities are not predominantly classroom teaching. For example, the American School Counselor Association supports school counselors' efforts to address children's academic and social–emotional development and to help them to prepare for careers.

Patricia A. Jenkins

See also Certification; Professional Development of Teachers; Stakeholders

Further Readings

Brigham Young University Department of Teacher Education. (n.d.). Professional Organizations. Retrieved from http://education.byu.edu/ted/professional.html

Dyrli, O. E. (2013, August 15). A list of professional organizations for K12 leaders. District Administration. Retrieved from https://www.districtadministration.com/article/list-professional-organizations-k12-leaders

Websites

American Federation of Teachers http://www.aft.org/

National Education Association http://www.nea.org/

Jana Craig-Hare Jana Craig-Hare Craig-Hare, Jana

Technology in Classroom Assessment Technology in classroom assessment

1659

1661

Technology in Classroom Assessment

Teaching in today’s modern classroom involves more than just standing in front of the class to cover content. Instead, teachers are learning facilitators, guiding students with curriculum-based activities designed to meet the individual needs and learning styles of each student. Many of these classrooms employ 1:1 laptop or tablet approaches, meaning that every student in the classroom has access to a mobile device (laptop or tablet-type device) every day in every classroom and oftentimes with the ability to take the device home. Technology provides not only an opportunity for teachers and students to be innovative but also an effective and efficient means for classroom assessment.

Historically, classroom assessments have been utilized as a summative measure for students to demonstrate content understanding at the end of a unit or lesson. Present-day classroom practice shares a more formative approach by encouraging the use of assessments *for* learning, rather than assessments *of* learning. Although assessment practices will most likely continue to have the external influence of standardized tests, technology has the potential to play a positive role in classroom assessment, using both formative and summative approaches to assess student learning, as well as create digital records of student learning that can follow the student throughout his or her educational journey.

Through constant monitoring of student learning and providing feedback to reinforce learning or to correct misconceptions, effective teachers must know when student learning is correct or incorrect. Classroom assessments that serve as meaningful sources of information reinforce the concepts, skills, and criteria learned in the classroom. Well-designed learning experiences involve feedback loops to assess learning. These feedback loops can be enhanced through the use of technology in the classroom. Within each of the types of assessments lies specific digital tools that help facilitate the assessment of student learning specific digital tools that help facilitate the assessment of student learning.

Formative and Summative Approaches

As a foundation, technology can be used to enable the delivery of online tests. Teachers oftentimes utilize online testing sites and/or software to create and deliver summative and formative tests for students. These tests might consist of multiple items to assess students’ understanding of a unit or lesson or a simple quiz. Creating an electronic test can be labor-intensive; however, the advantage for teachers is that once the test is developed, it is rather easy to modify the test for future use. Objective-type questions, such as multiple choice and true-false, can be automatically scored, providing immediate feedback to the student and saving time for the teacher who will need to score only essay-type questions. Testing in an online environment can be more interactive than traditional paper-and-pencil testing. Randomizing questions, including skip-logic, and embedding multimedia are just a few of the affordances that technology provides for classroom assessment. In addition, online tests provide accessible options for students with learning differences such as text-to-speech options for reading the questions or speech-to-text options for answering essay questions.

Formative classroom assessment provides a means for teachers to consistently check for understanding. Digital tools paired with research-based strategies and techniques provide teachers with valuable feedback to adjust their teaching and move student learning forward. Similar to online testing, websites and/or software can be used for formative assessment to assess whether a student is grasping the curriculum concept or knowledge being taught. Polling technologies (e.g., clickers, student response systems, and free online resources) can be utilized for students to respond to questions within a classroom environment. These technologies can potentially eliminate the feedback gap, as students are provided immediate feedback on their responses. With teacher guidance, students are able to reflect on their thinking and revise their knowledge and understanding before further misconceptions occur. Many of the free online resources provide game-like environments to engage students in multiplayer competitive formats. Digital graphic organizers can be used to help students visualize their understandings. Although some graphic organizers may only be provided in a digital format for students to access, print, and complete, others are online and may be accessed by a collaborative team of students and the teacher. Technology provides an authentic audience for student writing through the use of blogs, wikis, and individual websites. Writing passages can be shared online allowing for the teacher, classmates, and others to post comments to the student author. Additional classroom discourse opportunities can be provided through posting content-related questions to social media sites or a backchannel chat for students to respond. Collaborative note-taking offers further means for teachers to assess student's understanding of topics during the course. If used appropriately, these digital formative assessment tools can strengthen the feedback loop and allow instructors the opportunity to check for understanding, guide instruction, and develop student mastery.

Classroom Projects and Performance-Based Activities

Classroom projects and performance-based activities are often assessed differently than formative or summative tests. Technology provides many options for assessing these types of classroom activities. Classroom projects are enhanced by student's use of technology and digital tools. Assessing classroom projects oftentimes necessitates the use of scoring rubric. Online rubric creation sites allow for the collaborative development of a scoring rubric to be used for a classroom project. Oftentimes, these sites provide generic rubric language that can be customized to match individual classroom projects, thereby eliminating the tedious tasks of creating categories and criteria language across all quality ratings within the rubric. Students can use digital video tools on their smartphones to document a performance or demonstrate their fluency with a new process or skill. Although the teacher can use this recording to assess student performance or knowledge, these digital video/audio files are also self-assessment tools for students to review and evaluate their own performance. In addition, screencasting technology can be utilized for students to create a reflection of their learning or a mini-tutorial sharing their understanding of concepts they are learning. These screencasts can use images, text, and audio to demonstrate their understandings in an engaging, multimedia format.

Technology can also be utilized in the production of student portfolios. Utilizing digital tools, students can create evidence of their learning in a multimedia portfolio. Digital photographs, movies, audio clips, and user-created content can be added to scanned objects to create a comprehensive view of student learning and document growth and development. Blogs and wikis can showcase student writing examples over time and engage students in online discussions. Smartphones have created relatively accessible recording devices for capturing photos, movies, and audio clips to broaden the ways in which students can demonstrate their knowledge and skills. These digital portfolios can be shared demonstrate their knowledge and skills. These digital portfolios can be shared with a variety of audiences and might be used by the student for self-reflection, the teacher for assessment and grading, and even external organizations to determine evidence of readiness for future college and career opportunities. By providing online access to digital portfolios, student work can become part of a larger conversation, involving a broader scope of stakeholders and communities.

Integrating Technology Into Assessment

Teachers wanting to integrate technology into assessment should begin by exploring the variety of technology solutions available, especially in the area of formative assessment and creative ways for students to demonstrate their understanding of a topic. Recognizing that a classroom may consist of personal learning environments for each student, technology can be a key solution to effectively and efficiently assess individual student learning to help students manage their learning goals and document their learning progress. Digital tools for classroom assessment are readily available, oftentimes free or inexpensive, and can be a critical time-saving tool for teachers. However, it should be noted that these tools rely on technology and/or access to the Internet. Teachers should consider alternative assessment options to allow for Internet outages or limited student access to the Internet.

The use of technology for classroom assessment may require additional building or district policies to provide guidance for teachers and students. Such policies may define what constitutes a “submitted” assignment, acceptable file types for attachments, or even assessment offenses such as students not authoring their own work. In addition, teachers may create policies for online testing regarding the submission window and deadline, late submissions, and quality of submitted work. Exceptions for extenuating circumstances should be considered on a case-by-case basis for system failure or individual student hardships. In such cases, reasonable adjustment to the assigned assessment or due date can be made.

Benefits

There are many benefits of using technology in classroom assessment. Students are more likely to be engaged in the curriculum and interested in the outcome when they see their test results immediately. Digital tools also allow for students to access a test for summative or formative results or self-practice from anywhere that provides Internet access. Because many standardized tests including K–12 state assessments and the ACT can now be completed on a computer, using technology for classroom assessments provides an opportunity for students to practice with technology-based test formats and emerging test strategies required for digital tests. Digital tools used for classroom assessment provide an easy way for teachers to begin to use technology on a regular basis. These tools also prove to be a time-saver by eliminating or reducing teachers' grading time. These benefits can present challenges for classroom teachers as well. Although many safeguards may be in place, Internet communications may allow for someone other than the intended recipient to view student information and/or responses. Assigning a randomized number or pseudonym to students may help protect student confidentiality. Although the capability of computerized scoring has advanced in recent years, digital tools may still present some limitations for rigid scoring processes of online tests or quizzes. Choosing multiple-choice or true-false formats, rather than short-answer or fill in the blanks, may provide a more accurate scoring process.

Utilizing classroom assessments as information sources, followed by corrective instruction and the opportunity to relearn concepts, is a natural process for teachers working with students in their classrooms. Of all the options that technology provides for classroom assessment, providing immediate feedback and engaging students in authentic ways to demonstrate their learning through the use of digital tools may be the most significant advantages for classroom teachers and will add value to the learning experience for students.

Jana Craig-Hare

See also Classroom Assessment; Formative Assessment; Game-Based Assessment; Performance-Based Assessment; Portfolio Assessment; Summative Assessment

Further Readings

Confrey, J., & Maloney, A. (2012). Next-generation digital classroom assessment based on learning trajectories. In C. Dede & J. Richards (Eds.), Digital teaching platforms: Customizing classroom learning for each student. New York, NY: Teachers College Press.

Frey, B. B. (2013). Modern classroom assessment. Thousand Oaks, CA: Sage.

Gullen, K. (2014). Are our kids ready for computerized tests? Educational Leadership, 71(6), 68–71.

Magana, S., & Marzano, R. J. (2014). Art & science of teaching: Using polling technologies to close the feedback gap. Educational Leadership, 71(6), 82–83.

Magana, S., & Marzano, R. J. (2013). Enhancing the art & science of teaching with technology. Bloomington, IN: Marzano Research.

Meldrum, K. (2016). Assessment that matters. Irvine, CA: EdTechTeam Press.

Voce, J. (2015). Reviewing institutional policies for electronic management of assessment. Higher Education, 69(6), 915–929.

Michelle L. Boyer Michelle L. Boyer Boyer, Michelle L.

April L. Zenisky April L. Zenisky Zenisky, April L.

Technology-Enhanced Items Technology-Enhanced items

1662

1665

Technology-Enhanced Items

A technology-enhanced item is one that leverages available technologies in the context of computerized test delivery to efficiently measure elements of a construct domain. Such technological enhancements may exist in the item stem as a means for the examinee to access content, or in the nature of the response action by which the examinee provides a response. Through technology, examinees can be provided with access to item content and resource materials through text scrolling, playing a video, and/or using a calculator or glossary. At the response level, technological enhancements allow response manipulations that demonstrate applied knowledge, skills, and abilities, such as by manipulating text or objects through key strokes, point and click, touchscreen movements, or voice capture. This integration of technology is intended to foster and encourage more direct examinee application or demonstration of what the examinee knows and can do.

Technology-enhanced items expand the measurement possibilities of tests, but they likewise offer testing programs the opportunity to reduce scoring and testing burdens for more complex constructs. For example, productive language assessment and individually executed scientific experiments traditionally require one-on-one administration and scoring environments. Technologies such as voice recognition and automated scoring make it possible to realize more efficient processes over larger numbers of examinees, in a more limited time frame. This entry reviews the types, applications, uses, and development of technology-enhanced items and discusses concerns and challenges associated with their use.

Types and Applications

Types and Applications

The term *technology-enhanced item* describes a range of item types, existing and future. What is common across the various technology-enhanced item types is that they were made operationally possible through the technological advances that have occurred to make test administration, and increasingly complex examinee interactions with test content, widely available on computers and other devices, such as tablets and cell phones. The focus, then, is on the technology enhancements that are implemented in different aspects of test items and not a single item type.

Arguably, technology enhancements for large-scale assessment began with the first computer-based test administrations in the 1970s. At that time, computerized administration *was* the technological enhancement. As technology has continued to improve and has grown increasingly commonplace, the use of technology and technology-enhanced items has become widespread: Indeed, certification and licensure programs in a variety of professional contexts have been exploring the use of simulations since the 1990s. One early example of the use of a complex technology-enhanced item is the Architectural Registration Examination in the United States, in which the examinees must produce architectural designs using a computer-based interactive design tool, which was built to simulate, with reasonable fidelity, the actual practice of architectural design by asking candidates to produce design drafts given certain specifications and building regulations. More recently, both the Uniform CPA Exam and Step 3 of the United States Medical Licensure Examination have developed computerized simulation-based tasks as well in the areas of accounting and medical licensure, respectively. In the case of the Uniform CPA Exam, the test uses a wide range of item formats and situates the items in the context of accounting documents typically encountered in the practice of accounting, such as spreadsheets and other financial forms.

Not surprisingly, interest in the use of technology enhancements for test items has also grown in K–12 summative assessment programs. In educational settings, technology-enhanced items have incorporated a wide range of test item types and response actions and have found particular use not only in the mathematics and science domains but also in reading and writing. Items might ask examinees to order elements in a particular sequence onscreen or manipulate presented items to carry out simulated science experiments and describe the results observed.

Uses

There are multiple intersecting conditions that underlie this growing interest in technology-enhanced items, including the advancing technology itself, a growing demand for assessments that measure increasingly complex cognitive processes in efficient ways, calls for more engaging test content, and changes in the knowledge and skills required for graduates to be competitive in the job market. Although this list is certainly not exhaustive, it provides a context to describe the emergence and continued evolution of technology-enhanced items.

Another underlying condition that is motivating the use of technology enhancements for test content and delivery is the need for efficient measures of complex cognitive processes and applied skills that have been historically difficult to measure efficiently in large-scale assessment. To measure such traits, traditional large-scale assessment relies heavily on open-ended items whereby examinees construct written responses to demonstrate their knowledge, skills, and abilities. Such items ask examinees to respond with very few constraints. Although such items are widely considered to make strong contributions to test validity arguments, they also represent an efficiency challenge through the large, expensive, and time-consuming effort required to accurately score examinee responses. One potential promise of technology enhancement is that these items may offer similar opportunities for examinees to demonstrate complex, cognitive, and applied skills, but with fewer constraints than traditional multiple-choice items, and while allowing rapid scoring based on predefined rules.

Also motivating the use of technology enhancements for test content and delivery is a demand for measures of “new” traits, including those considered necessary for job competitiveness. The very technological and global nature of current and future jobs calls for graduates who have skills in using technology to access, create, assimilate, and share increasingly large volumes of information in both traditional and novel ways. In addition, interest has grown in measures of so-called soft skills, as they continue to be recognized as having a strong relationship with academic and job-related success. Examples of such traits include task perseverance, collaboration, and continuous learning.

Development

Technological advances that support the wider development of technology-enhanced items include dramatically increased device memory, storage, and enhanced items include dramatically increased device memory, storage, and speed, all of which support more rapid and flexible test content and delivery. For example, cloud-based computing has made it possible to bypass loading content in local clients, which in turn supports more rapid deployment of new test content. More complex content, such as video, voice, and interactive graphics, is also supported. The adaptive assignment of content to examinees under large numbers of test construction decisions and constraints is made possible through the availability and use of linear equation solvers. Online testing environments also provide for rich data capture of examinee interaction with test content such as timing, keystrokes, response patterns, and answer changes—all of which can now be made available for use in decisions about examinee performance, as well as assist in explaining it.

Concerns and Challenges

There are, however, some important psychometric and operational issues that must be taken into account when considering implementation of technology-enhanced items. Chief among these is the nature of the data gathered and the extent to which such valid and reliable inferences can be made on the basis of such data. In contrast to traditional static test items, with technology-enhanced items, there is often an enormous data trail that can be collected. These data can be set to include not only the very final answer or response but also everything that an examinee types in or clicks on with the mouse, from authoritative literature to tabs displayed onscreen, as well as a timestamp for each click. These are all data, but an emerging issue for testing agencies is to try and parse out the construct-relevant data from the construct-irrelevant noise to figure out how best to take advantage of these items as valid and reliable measurement opportunities.

Part of the challenge is that the data collected can be viewed as falling into two general categories: *process* and *outcomes*. Outcomes from technology-enhanced items—the actual final answers examinees provide—are relatively easier to handle from a scoring perspective, given current dichotomous and polytomous models. The psychometric literature has a long history of well-developed models for creating rubrics and scoring examinee work, even for complex products such as essays, portfolios, and performances. On the side of process data, however, there is relatively less guidance for evaluation. Because the processes that examinees may follow as they go about answering a technology-enhanced item may vary considerably, more research is needed to develop strategies for evaluating process from both qualitative and quantitative perspectives. In some cases, agencies have tried to develop profiles of examinees based on patterns observed in process data, but such process data are complex by nature and difficult to categorize in ways that can be readily communicated and used. To be clear, process-related information may indeed be valuable in terms of summative results and/or for diagnostic or formative purposes, though it remains to be seen the extent to which the useful information can be distilled and presented in a functionally helpful way. It may well be that the continued use of technology-enhanced items affects a new branch of psychometric theory in which indices of reliability and validity evidence are reconceptualized because traditional formulations may not be well equipped to handle these data.

Another area of concern regarding the use of technology-enhanced items involves construct representation and construct-irrelevant variance, as these are the primary validity issues that can impact the extent to which these kinds of items facilitate appropriate interpretations. The degree to which technology-enhanced items actually measure the domain of skills or interest in a specific testing context must be documented because agencies must be careful to ensure that the full intended domain is assessed. This is especially important, given that many technology-enhanced items can be time-consuming to complete. As compared to traditional test item formats, tests comprised of technology-enhanced items may have fewer “measurement opportunities” due to the trade-off of including more complex items while being constrained by limited administration time, but the data collected from these kinds of items may be richer and differently informative. In terms of construct-irrelevant variance, a concern for technology-enhanced items is the potential for the presence of characteristics that may affect students’ performance on a test that are extraneous to the construct measured. The consequence of construct representation and construct-irrelevant variance as validity issues is that validity evidence must be obtained to show that the item formats either increase construct representation or at least maintain the same level achieved by other available testing formats. It also means that such item formats should minimize or eliminate measurement of proficiencies that are unrelated to the construct targeted by the test.

As some item types that fall under the technology-enhanced item umbrella are highly akin to performance assessment tasks, some of the challenges that affect those tasks are relevant to the present discussion. For some of the highly extended scenario-based, technology-enhanced items, one open question for research and practice concerns task specificity and the question of whether patterns of performance observed in one particular scenario can and should generalize to other scenarios. Speededness is another consideration: Many generalize to other scenarios. Speededness is another consideration: Many technology-enhanced items can be designed and implemented in such a way as to require a degree of familiarity with the specific user interface in order to proceed through the task, and examinees must be given adequate opportunity to fully understand how to navigate through and respond to these kinds of tasks, including instruction on how to access and use resources that might be made available.

Accessibility and fairness represent another important challenge for technology-enhanced items. This is a great concern in many testing contexts, but this is especially relevant in education, where the examinee population may be quite diverse in terms of language background and learners with disabilities. Incorporating the principles of universal test design at the outset of test development can help to improve the accessibility of technology-enhanced items, but at the same time, many technology-enhanced items rely on highly physical interaction between the examinee and the user interface to provide a response action, and this must be addressed as a matter of both research and practice regarding the use of technology-enhanced items.

Michelle L. Boyer and April L. Zenisky

See also Computer-Based Testing; Computerized Adaptive Testing; Multiple-Choice Items

Further Readings

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke, J.,… Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. Journal of Technology, Learning, and Assessment, 10(5), 1–52.

Bennett, R. E. (1999). Using new technology to improve assessment. Educational Measurement: Issues and Practice, 18, 5–12. doi:10.1111/j.1745–3992.1999.tb00266.x

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. Journal of

Technology, Learning, and Assessment, 8(8). Retrieved January 2, 2011, from http://www.jtla.org

Huff, K. L, & Sireci, S. G. (2001). Validity issues in computer-based testing. Educational Measurement: Issues and Practice, 20, 16–25. doi:10.1111/j.1745–3992.2001.tb00066.x

Quellmalz, E. S., & Pellegrino, J. (2009). Technology and testing. Science, 323(5910), 75–79.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. Journal of Technology, Learning, and Assessment, 8(4). Retrieved January 2, 2011, from http://www.jtla.org

Sireci, S. G., & Zenisky, A. L. (2015). Computerized innovative item formats: Achievement and credentialing. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 313–334). New York, NY: Routledge.

Wendt, A., & Harmes, J. C. (2009). Evaluating innovative items for the NCLEX, Part 1: Usability and pilot testing. Nurse Educator, 34(2), 56–59.

Harrison J. Kell Harrison J. Kell Kell, Harrison J.

Jonathan Wai Jonathan Wai Wai, Jonathan

Terman Study of the Gifted Terman study of the gifted

1665

1667

Terman Study of the Gifted

The Terman Study of the Gifted (originally known as Genetic Studies of Genius) is one of the most famous longitudinal studies in the history of psychology. In 1921, Lewis M. Terman, professor of psychology in Stanford University, initiated the study and its sample was comprised of 1,528 children (11 years old, on average), all with IQs of 135 or above—placing them in the top 1% of the population at the time. Its participants (Termites) were systematically followed for over 80 years: Comprehensive surveys and interviews investigated all aspects of their lives, including educational and occupational achievements, mental and physical health, marital and parental status, and mortality.

The results of Terman's study provide important insights into the long-term, real-world influence of intelligence as defined and assessed by standardized tests. Terman's work also inspired subsequent studies of the gifted (e.g., Study of Mathematically Precocious Youth), which have replicated and extended many of his findings. This entry describes the genesis and rationale for the study, summarizes the general trends of its findings, and concludes with brief descriptions of some of the study's most notable members.

Origins

Terman's interest in intelligence long predated his Study of the Gifted: His 1905 dissertation compared the mental and physical abilities of boys identified as being of very high and very low intelligence. In 1916, Terman and his colleagues published their translation of the original Binet–Simon intelligence test. They relied primarily on Stanford–Binet, one of the most widely used IQ tests, to identify children for the Study of the Gifted.

Nearly all the children identified for the study lived in California cities (where the search was largely limited to) and the majority came from middle-to upper-class households. A major impetus for the study was to disprove the stereotype that highly intelligent children were physically frail, socially incompetent, and emotionally maladjusted; “early ripe, early rot” was a phrase often used to describe them. Terman believed that gifted children were in fact superior in many ways to children of average intelligence and that by identifying them early on, they could be given the appropriate opportunities that would allow them to develop into society’s leaders. (Like many of his contemporaries, Terman was a proponent of eugenics, although his views were not as extreme as those many others held at the time.)

Summary of Findings

The study produced an avalanche of data: five books, one monograph, and hundreds of articles. When they were initially assessed, the Termites put the lie to the early ripe, early rot stereotype: Compared to children of the same age, but of average intelligence, participants were taller, heavier, and stronger; had the same rate of contagious diseases; better nutrition; and were equally emotionally well adjusted. Gifted children were as interested in sports as children of average intelligence, reported spending an average of over 2 hours per day with children outside of school, and did not report being teased significantly more than “normal” children. The gifted did, however, evince slightly less interest in competitive games, were rated as somewhat less sociable, and reported playing alone slightly more than children from the general population.

As they matured, the Termites obtained numerous positive outcomes at many times the rate of individuals of average intelligence: two thirds earned bachelor’s degrees (10 times the rate of the general population at the time) and 8 times as many earned doctoral degrees as typical college graduates. Their occupational attainment was similarly impressive, with 95% of men working in jobs categorized as “professional” or “high-level business” by the U.S. Census Bureau and receiving income that was 4 times greater than that of the general population. (Owing to the lack of opportunities at the time, women’s career outcomes were less impressive.) Over 90% of participants married and over 80% had children. The gifted remained healthier than the general population as they aged, lived an average of approximately 10 years longer, and retained their place in the top 1% of intelligence, as evidenced by IQ tests they were given later in life. The gifted did not exceed the general population in all ways, however, and they fared no better in terms of alcoholism, suicide, and divorce.

Notable Participants

Many Termites were remarkably successful. Lee J. Cronbach and Robert R. Sears were two eminent psychologists—and their own subjects, as they were highly involved in the study in its later years. Ancel Keys invented the K-ration. Jess Oppenheimer created, produced, and wrote *I Love Lucy*. Edward Dmytryk directed 23 films, one of which (*Crossfire*) earned Academy Award nominations for Best Picture and Best Director. L. Sprague de Camp was an award-winning fantasy and science fiction writer. Norris Bradbury was director of Los Alamos National Laboratory. Shelley Smith Mydans was a novelist and reporter for *Life* and *Time*. William A. P. White, writing under the pen name “Anthony Boucher,” was one of the original editors of *The Magazine of Fantasy and Science Fiction*. Douglas McGlashan Kelley was the chief psychiatrist during the Nuremberg trials.

Despite the incredible accomplishments of an elite sample of Termites, and the study population as a whole, the Study of the Gifted did not produce any “indisputable geniuses,” and none of its members won a Nobel Prize. (It is worth noting that the two winners of the Nobel Prize in Physics, William Shockley and Luis Alvarez, were tested but failed to qualify for the study.) Terman was very pleased with the extraordinary accomplishments of his Termites—but also concluded that the relationship between intelligence and achievement is far from perfect.

Harrison J. Kell and Jonathan Wai

See also Ability Tests; Aptitude Tests; Giftedness; Intelligence Quotient; Intelligence Tests; Standardized Tests

Further Readings

Fancher, R. E. (1985). The intelligence men: Makers of the IQ controversy. New York, NY: W. W. Norton & Company.

Friedman, H. S., & Martin, L. R. (2011). The longevity project. New York, NY: Hudson Street Press.

Holahan, C. K., Sears, R. R., & Cronbach, L. J. (1995). The gifted group in later maturity. Stanford, CA: Stanford University Press.

Minton, H. L. (1988). Lewis Terman: Pioneer in psychological testing. New York: New York University Press.

Oden, M. H. (1968). The fulfillment of promise: 40-year follow-up of the Terman gifted group. Genetic Psychology Monographs, 77, 3–93.

Shurkin, J. N. (1992). Terman’s kids. Boston, MA: Little, Brown.

Terman, L. M. (1925–1959). Genetic studies of genius (Vols. 1–5). Stanford, CA: Stanford University Press.

Terman, L. M. (1930). Autobiography of Lewis M. Terman. In C. Murchison (Ed.), History of psychology in autobiography (Vol. 2, pp. 297–331). Worcester, MA: Clark University Press.

Krystal Mendez Krystal Mendez Mendez, Krystal

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Test Battery

Test battery

1667

1667

Test Battery

A test battery consists of a series of tests administered to assess different facets of a child’s or adult’s functioning (e.g., psychological functioning). A test battery is utilized by a professional (e.g., a psychologist) to assist in decision making, such as making a diagnosis, about an individual and determining whether there is a need for services and supports for that person. Introduced by Francis Galton in 1884, the initial test battery was used to measure an individual’s sensory and motor abilities. Once the battery of tests was administered, a report was written to summarize the findings. Test batteries have evolved over time.

Test Battery Process

There are a variety of test batteries that can be used to collect meaningful data on an individual. Test batteries often consist of norm-referenced measures and informal assessments. Norm-referenced measures are well standardized and psychometrically sound tests that allow an examinee to be compared to a normative group, whereas informal assessments tend to be less psychometrically sound. Examples of norm-referenced tests include standardized intelligence and academic achievement tests, whereas examples of informal assessments are projective and curriculum-based measures.

Test batteries can be used in many different fields. For example, educators and school professionals have relied on cognitive, academic achievement, behavioral, and social–emotional measures to determine appropriate services and supports for students with disabilities. A standard battery approach or a process approach may be used in the administration of a group of tests to an individual. A standard battery approach involves selecting and administering a group of tests based on the reason for referral (i.e., the reason for testing the individual) and the professional's hypotheses generated concerning the difficulties the person is experiencing. The battery of tests selected does not change once the administration of the measures begins. In the process approach, selection and administration of a group of tests are also based on the reason for referral and the professional's hypotheses, but the actual tests used in the assessment are altered in the process as more information is gleaned about the individual during the assessment.

The administration of a test battery may take less than 2 hours or occur over several days. Once the test battery is complete, the measures are scored and a report is issued summarizing the findings and making recommendations (e.g., services, supports, and treatments suggested for the individual) based on the assessment results. After the report is completed, a feedback session is usually conducted with the individual assessed or, if a child is the examinee, the parents and the child to explain the results of the assessment in layperson's terms. The information gleaned from test batteries can assist professionals in helping individuals develop a better understanding of their strengths and/or difficulties and determine whether services, supports, or treatments are needed, so that individuals can lead more productive and self-fulfilling lives.

Krystal Mendez and Patricia A. Lowe

See also Psychometrics; Reliability; Standardized Tests; Tests

Further Readings

Gregory, R. J. (2004). The history of psychological testing. In R. J. Gregory (Ed.), Psychological testing: History, principles, and applications (4th ed., pp. 1–28). Needham Heights, MA: Allyn & Bacon.

Irvine, P. (1986). Sir Francis Galton (1822–1911). The Journal of Special Education, 20(1), 6–7.

Stephen W. Loke Stephen W. Loke Loke, Stephen W.

Patricia A. Lowe Patricia A. Lowe Lowe, Patricia A.

Test Bias

Test bias

1667

1670

Test Bias

Test bias is one of the most important issues in the development of measures. However, it is often confused with fairness. On one hand, fairness is a social concept that is concerned with whether one views test scores as being used in an appropriate manner. There is no right way to examine whether test scores are used appropriately, as this is based on an individual’s subjective perception. On the other hand, bias is viewed as a statistical issue and is concerned about whether there is systematic error in measuring a trait or attribute across groups. If systematic differences due to group membership on a test exist, this would suggest that bias is present in the test. This article outlines several methods used to examine test bias.

Differential Item Functioning (DIF)

DIF is a method to determine whether a measure (e.g., a personality, intelligence, or academic achievement measure) is equivalent across groups. DIF occurs when an item on a measure is responded to differently by individuals in different groups, such as different gender, age, or ethnic groups, who have the same amount of an attribute or a latent trait. The latent trait or attribute could be an ability, skill, or personality characteristic. If DIF exists on an item, then the item may be biased. DIF is important to investigate because group comparisons, such as age, gender, or ethnic differences, on a measure cannot be made unless the items are found to be equivalent across the groups of interest. Equivalence of items across groups should be examined when one develops new instruments or it can be examined with measures already existing in the field.

There are two types of DIF: uniform DIF and nonuniform DIF. Uniform DIF is when the probability of a specific response to an item (e.g., the probability of endorsing a yes response on an item) is higher for one group than for another group (e.g., females than males) at each level of the attribute (e.g., anxiety) that is being measured. In contrast, nonuniform DIF is when the probability of a specific response to an item differs at different levels of the attribute. For example, males may be more likely to endorse a no response on an item at lower levels of an attribute but are more likely to endorse a yes response on the item at higher levels of the attribute.

Different procedures exist for detecting DIF. Some of these approaches are nonparametric and others are parametric. One of the most common nonparametric approaches for detecting DIF is the Mantel-Haenszel method. The Mantel-Haenszel is a contingency table-based approach that uses odds ratios to determine whether one group outperforms the other group on each of the items. If a common odds ratio indicates that one group outperforms the other group across all levels of the trait or attribute for a specific item, then DIF is said to be present for that item. Besides the nonparametric approach, there are two common parametric methods to detect DIF: the logistic regression and item response theory (IRT) approaches. Hariharan Swaminathan and H. Jane Rogers indicate that nested models can be compared in the logistic regression approach. One model, referred to as the augmented model, that includes the group (e.g., gender), the trait (e.g., anxiety), and the interaction between the group and the trait variable is tested against another model, referred to as the compact model, that includes the group and the trait variable, but not the interaction term. Jeanne A. Teresi and John A. Fleishman assert that when the augmented and compact models are estimated using the maximum likelihood parameter estimator, a likelihood value is obtained for the augmented and the compact models, and the difference in the log-likelihood values between these models is examined using a chi-square test. If the chi-square test is significant, indicating a significant group by trait interaction, then nonuniform DIF is present. If the chi-square test is not significant, then the compact model is compared to a model where no group effect is assumed. If the chi-square test is significant, indicating a group effect exists, then uniform DIF is present. This procedure is repeated for items on the measure. However, it should be noted that variations do exist in the logistic regression approach to detect DIF. IRT is another method that can be used to detect DIF, and there are variations in this approach too. In IRT, the item characteristic curve, an S-shaped curve, represents the graphic relationship between the probability of giving a certain response on an item on a measure and an individual’s position on the latent trait continuum. The shape of the curve is determined by its parameters (discrimination, difficulty, and if applicable, guessing). IRT models are derived from these parameters, including one-, two-, and three-parameter models. To detect DIF, the item characteristic curves of two groups are compared and if one of the parameters is different for the two groups, then DIF is likely to be present. Different statistical tests, such as a likelihood ratio test, or magnitude measures are used to determine the salience of DIF.

A significant DIF obtained through one of the parametric or nonparametric approaches does not mean that an item is biased necessarily. Cecil R. Reynolds and Patricia A. Lowe state that an item flagged because of significant DIF is considered to be biased only after further research has been conducted and careful consideration has been made as to whether the item is not tapping into the intended construct of interest.

Factor Analysis

Factor analytic methods are used to group items or subtests that are highly correlated with one another. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are the two main factor analytic methods utilized.

EFA

An EFA seeks to identify whether the latent structure of a test is similar across groups. Harry Harman mentioned that coefficient of congruence values can be computed between groups after an EFA is performed on the data of each of the groups. To compute this value, factor coefficients for each item on a corresponding factor for each group are multiplied and then added together. This value is then divided by the square root of the sum of squared factor coefficients for each group. Should the coefficient of congruence value be at least .90, this would suggest no evidence of construct bias between groups. Additionally, Raymond B. Cattell stated that the salient variable similarity index is frequently used in conjunction with the coefficient of congruence values. In this method, a threshold of salience (e.g., +.15) for the factor coefficient to be salient is selected. Factor coefficients that exceed this threshold are considered positively salient, whereas factor coefficients that are below this threshold are considered negatively salient. The positive or negative salient factor coefficients on each item for each group are paired. The frequency of each pairing is then entered into a matrix and the salient variable similarity index is computed via a formula. A salient variable similarity index value that is closer to +1.00 would suggest that similar constructs are measured across groups; however, if the salient variable similarity index is closer to −1.00, this would suggest that relatively different constructs are measured across groups.

CFA

A CFA seeks to explore whether there are differences in the latent structure of a test across groups. Specifically, multigroup CFAs are performed by testing for measurement invariance across groups. In this process, a least restricted model is tested against another model that consists of additional parameter constraints. Matthew R. Reynolds and Timothy Z. Keith asserted that testing for measurement invariance is one of the most important methods in assessing for test bias and that measurement should be invariant across groups if there is no bias in a test across groups. In other words, individuals in different groups with similar latent traits or attributes should have similar observed scores. To test for measurement invariance, Brown mentioned that tests for configural invariance, weak factorial invariance, strong factorial invariance, and strict factorial invariance are performed in a stepwise manner in this order.

Configural invariance is also referred to as the test of equal factor structures, as it pertains to the measurement model being similar across groups. This suggests that the number of factors and the pattern of factor-indicator (e.g., item) correspondence are similar across groups. Once configural invariance has been established, a test for weak factorial invariance is performed. Weak factorial invariance is also referred to as the test of equal factor loadings. This entails adding an additional constraint to equate unstandardized factor loadings across groups. If weak factorial invariance is tenable, this would suggest that the factor loadings across groups are proportionally similar to one another. Once weak factorial invariance is tenable, a test for strong factorial invariance is performed. Strong factorial invariance is also referred to as a test of equal indicator intercepts. In this model, an additional equality constraint beyond the weak factorial invariance model—that is, constraining the intercepts (or thresholds, depending on parameter estimator used) across groups to be equal—is imposed. If strong factorial invariance is tenable, this suggests that the latent factors for both groups have the same unit of measurement (or are on the same scale). Once strong factorial invariance is tenable, a test for strict factorial invariance is performed. To test for strict factorial invariance, an additional constraint beyond the strong factorial invariance model—that is, constraining the residual variances and covariances across groups to be equal—is imposed. If strict factorial invariance is tenable, this suggests that any differences in observed scores across groups are accounted for by group differences in the factor means and variances of the latent variable.

To evaluate whether the competing nested model fits the data just as well as the less restricted model, multiple goodness-of-fit indices are used. Researchers have proposed guidelines to be used to demonstrate adequate model fit for the following goodness-of-fit indices. For example, Gordon W. Cheung and Roger B. Rensvold had mentioned that a nonsignificant change in chi-square and a decrease in the comparative fit index of less than or equal to .01 between models, and Todd D. Little recommended that the root mean square error of approximation of the nested model should fall within the 90% confidence interval of the less restricted model to indicate invariance across groups. Taken together, if a preponderance of information across multiple goodness-of-fit indices indicates at least acceptable model fit when a multigroup CFA is performed, this would suggest that there is no evidence of bias between groups. Alternatively, if most fit indices indicate a poor model fit, this would suggest that there is bias between groups on the test.

Final Thoughts

When developing a test, it is critical that test developers ensure that no evidence of bias is present in tests based on a preponderance of evidence using the different methods discussed herein to assess for bias. Test bias should also be examined on existing measures. In particular, Reynolds and Lowe asserted that despite bias being less frequently examined in psychological measures, it is necessary to examine bias in these measures too, as a biased instrument may influence the interpretation of scores for individuals of different groups. Furthermore, with cross-cultural research becoming more prominent, the issue of test bias is even more important, as researchers have to be more cognizant of test bias when using measures that are developed in one culture and used in other cultures.

Stephen W. Loke and Patricia A. Lowe

See also Differential Item Functioning; Measurement Invariance

Further Readings

Cattell, R. B. (1978). The scientific use of factor analysis in behavioral and life sciences. New York, NY: Plenum.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9, 233–255. doi:10.1207/s15328007sem0902_5

Harman, H. (1976). Modern factor analysis (2nd ed.). Chicago, IL: University of Chicago Press.

Keith, T. Z., & Reynolds, C. R. (1990). Measurement and design issues in child assessment research. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children. New York, NY: Guilford.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research, 32, 53–76. doi:10.1207/s15327906mbr3201_3

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), The handbook of school psychology (4th ed., pp. 332–374). New York, NY: John Wiley.

Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child assessment research. In D. H. Saklofske, C. R. Reynolds, & V. Schwean (Eds.), Oxford handbook of child and adolescent assessment (pp. 48–83). New York, NY: Oxford University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 26, 361–370.

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. Quality Life Research, 16, 33–42.

Matthew S. Johnson Matthew S. Johnson Johnson, Matthew S.

Test Information Function Test information function

1671

1674

Test Information Function

The test information function (TIF), a function of the unknown ability or true score θ, is a measure of the amount of information provided by the item responses on a test about θ. The TIF is defined as the Fisher information for θ contained in the item response vector ***X***, and under the typical item response theory (IRT), assumption of local independence is the sum of the item information functions (IIFs). It is important in the context of ability or true score estimation because the TIF serves as an estimate of the precisions of the maximum likelihood estimator (MLE) of the ability θ, or equivalently, the inverse of the TIF is an estimate of the variance of the MLE. Test developers often use the TIF for test construction purposes and to compare two competing tests of the same construct.

Formal Definition

Let ***X*** = $(X_1,X_2,\ldots,X_J)$ denote the random vector of item responses from an examinee to a *J*-item test and assume that we have a model that describes how the item responses are related as a function of the unobservable or latent ability θ. This model defines the likelihood function of θ given the item responses; let denote this likelihood function. Taking the natural logarithm of the likelihood function produces the log-likelihood function of θ given the random item responses and is denoted . If the unknown parameter θ is known to lie in some open interval of the real line, then the Fisher information in the item response vector ***X*** about θ is defined as:

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;\boldsymbol{X})\right)^2 \mid \theta\right],$$

where *E*[ċ] denotes the expectation, which in this case is taken with respect to the conditional distribution of item responses given θ. When used within the context of educational or psychological testing, is usually called the TIF.

There are two alternative forms of the Fisher or test information that are often used when the set of possible item responses *X* is not restricted by the value of the unknown parameter θ; this condition holds for any typical IRT model. In this case, the following two results hold:

$$\frac{\partial}{\partial\theta}\sum_x L(\theta;\boldsymbol{X}) = \sum_x \frac{\partial}{\partial\theta} L(\theta;\boldsymbol{X})$$

$$\frac{\partial^2}{\partial\theta^2}\sum_x L(\theta;\boldsymbol{X}) = \sum_x \frac{\partial^2}{\partial\theta^2} L(\theta;\boldsymbol{X}),$$

and the TIF can be equivalently found by either finding the variance of the first derivative of the log likelihood or the negative expected value of the second derivative of the log likelihood:

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta}\ell(\theta;\boldsymbol{X}) \mid \theta\right) = -E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta;\boldsymbol{X}) \mid \theta\right]$$

Consider a test consisting of *J* dichotomously scored items. Let denote the item response function or item characteristic curve for item *j*, and let denote the logit of the item response function or log odds of a correct response given the ability parameter θ. Then the log-likelihood function of θ given the item response vector ***X*** is:

$$\ell(\theta;\boldsymbol{X}) = \log\left(\prod\nolimits_j p_j(\theta)^{X_j}(1-p_j(\theta))^{1-X_j}\right)$$

$$= \sum\nolimits_j X_j \psi_j(\theta) - \sum\nolimits_j \log(1+e^{\psi_j(\theta)}).$$

Taking the derivative of this log likelihood with respect to θ produces the score function:

$$\frac{\partial}{\partial\theta}\ell(\theta;\boldsymbol{X}) = \sum\nolimits_j X_j \psi'_j(\theta) - \sum\nolimits_j \left(\frac{\psi'_j(\theta)e^{\psi_j(\theta)}}{1+e^{\psi_j(\theta)}}\right).$$

The TIF is then found by finding the variance of the quantity above over the distribution of item responses given the latent ability θ. The second term is constant with respect to the item responses and, therefore, has no impact on the variance of the score function. So, the TIF of a test with binary items is . Under the typical IRT model, assumption that item responses are conditionally independent, given the latent ability θ, produces the following form for the TIF:

$$I(\theta) = \sum\nolimits_j [\psi'_j \theta)]^2 Var(X_j \mid \theta)$$

$$= \sum\nolimits_j [\psi'_j(\theta)]^2 p_j(\theta)(1-p_j(\theta)).$$

Notice that this is a summation over the items *j* = 1, … , *J* with the summands equal to:

$$I_j(\theta) = [\psi'_j(\theta)]^2 p_j(\theta)(1-p_j(\theta))$$

$$= \frac{[p'_j(\theta)]^2}{p_j(\theta)(1-p_j(\theta))};$$

is the item information for item *j*, the amount of information about θ contained in the response to item *j*. This result holds in general as long as the local independence of items given the latent ability θ holds. That is, under the assumption of local independence the TIF is the sum of the IIFs, .

Consider, for example, a test consisting of two-parameter logistic (2PL) items. The item response function or item characteristic curve of the 2PL is equal to:

$$p_j(\theta) = \frac{1}{1+e^{-a_j(\theta-b_j)}},$$

and the log odds function is equal to:

$$\psi_j(\theta) = a_j(\theta - b_j).$$

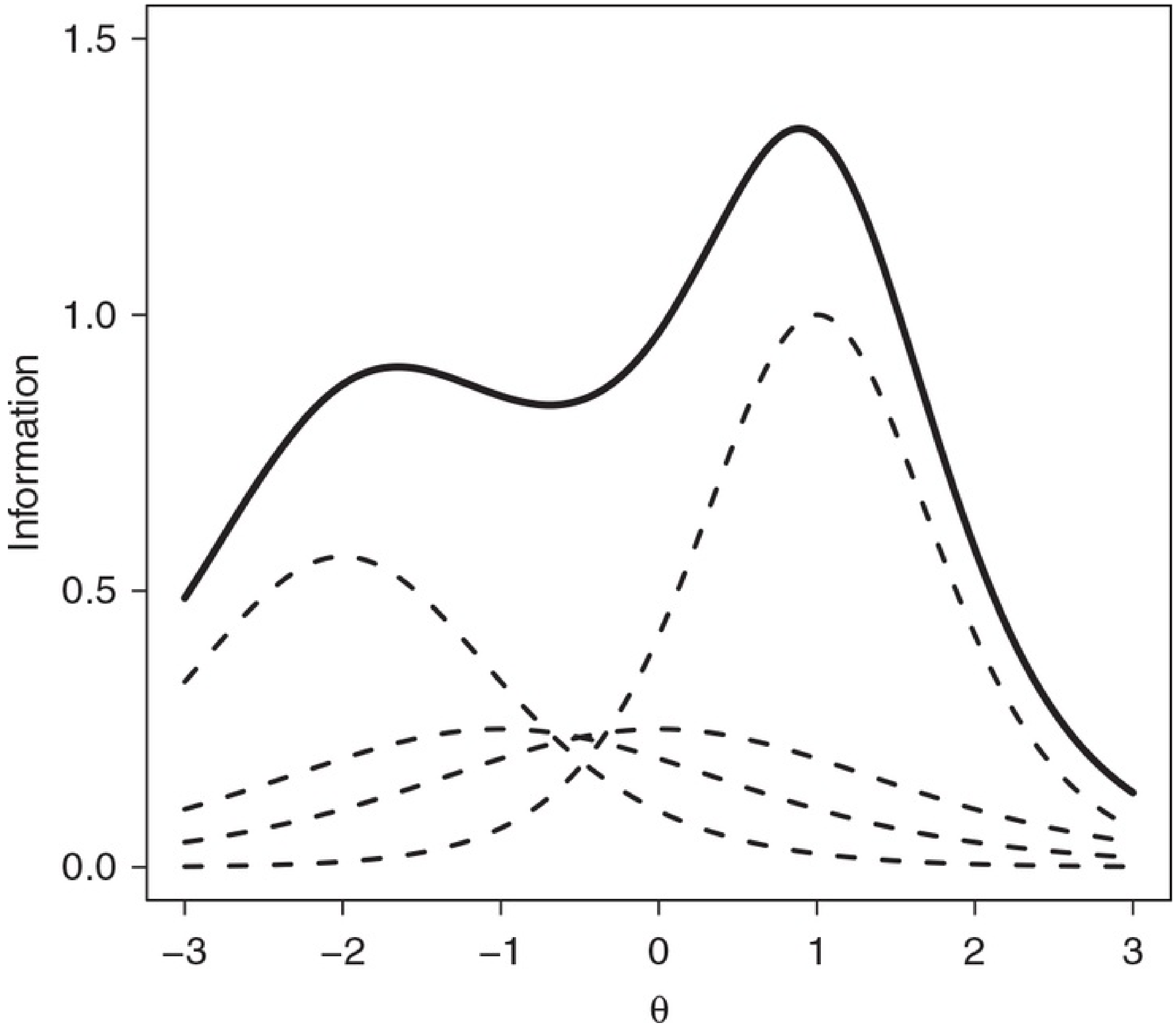
Taking the first derivative of the log-odds function, we have , so the IIF for a 2PL item is . The IIFs and TIF for four 2PL items with *a* parameters equal to ***a*** = (1.5, 1.0, 1.0, 2.0) and *b* parameters equal to ***b*** = (−2.0,−1.0, 0.0, 1.0) are displayed in Figure 1.

The dashed curves in Figure 1 represent the IIFs for the four hypothetical 2PL items, and the solid curve represents the TIF. The IIFs are symmetric around the difficulty or *b* parameter values for the item, and the height of the information function at the *b* parameter is equal to one fourth of the squared discrimination (*a* parameter) value. The TIF is simply the sum of the four IIFs.

Use as a Measure of Measurement Precision

One important use of the TIF is as a measure of the precision of the estimate of the unobserved ability θ. For example, if is the MLE of θ, then the conditional variance of the MLE given θ will approach as the number of items increases; therefore, the conditional standard error of measurement given θ is approximately equal to SEθ = 1Iθ for sufficiently long tests.

Figure 1 Item information functions (dashed curves) and test information function (solid curve) for five items with *a* parameters equal to 1.5, 1.0, 1.0, 2.0 and *b* parameters equal to −2.0, −1.0, 0.0, 1.0.



Because the information changes as a function of θ, measurement precision is not constant with respect to θ. This is in contrast to results from classical test theory, which, under the assumptions of that theory, produce precision measures that are constant with respect to the unknown ability measure θ. Furthermore, because the test information is a function of the true ability level θ it will always be unknown; as such, common practice is to use the observed information function , found by calculating the information at the maximum likelihood estimate, as an estimate of the true information *I*θ.

Another important result involving the Fisher information in general, and the TIF in the context of IRT, is the asymptotic normality of the MLE. Under the standard assumptions of IRT, the MLE is approximately normally distributed with mean equal to the true ability θ and variance equal to , that is, approximately for sufficiently long tests. The asymptotic normality of the MLE, allows for a relatively simple approach to hypothesis testing and construction of confidence intervals.

Because the distribution of is approximately standard normal $N(0,1)$, a Wald z test statistic can be constructed for testing the null hypothesis $H_0:\theta = \theta_0$ by calculating the z statistic and comparing it to the appropriate quantiles of the standard normal distribution to determine statistical significance, or finding the probability that a normal random variable exceeds the observed z statistic value to determine the p value of the test.

Similarly, the information function is useful for constructing confidence intervals for the ability parameter θ. There are two slightly different approaches for constructing the confidence intervals. One approach is to invert the Wald test statistics described in the previous paragraph to find the set of θ_0's that would not be rejected in a hypothesis test. For example, the $(1 - \alpha) \times 100\%$ two-sided confidence is defined as the set , where is the quantile of the standard normal distribution. The second approach uses the observed information function to construct the interval .

Efficiency of a Test

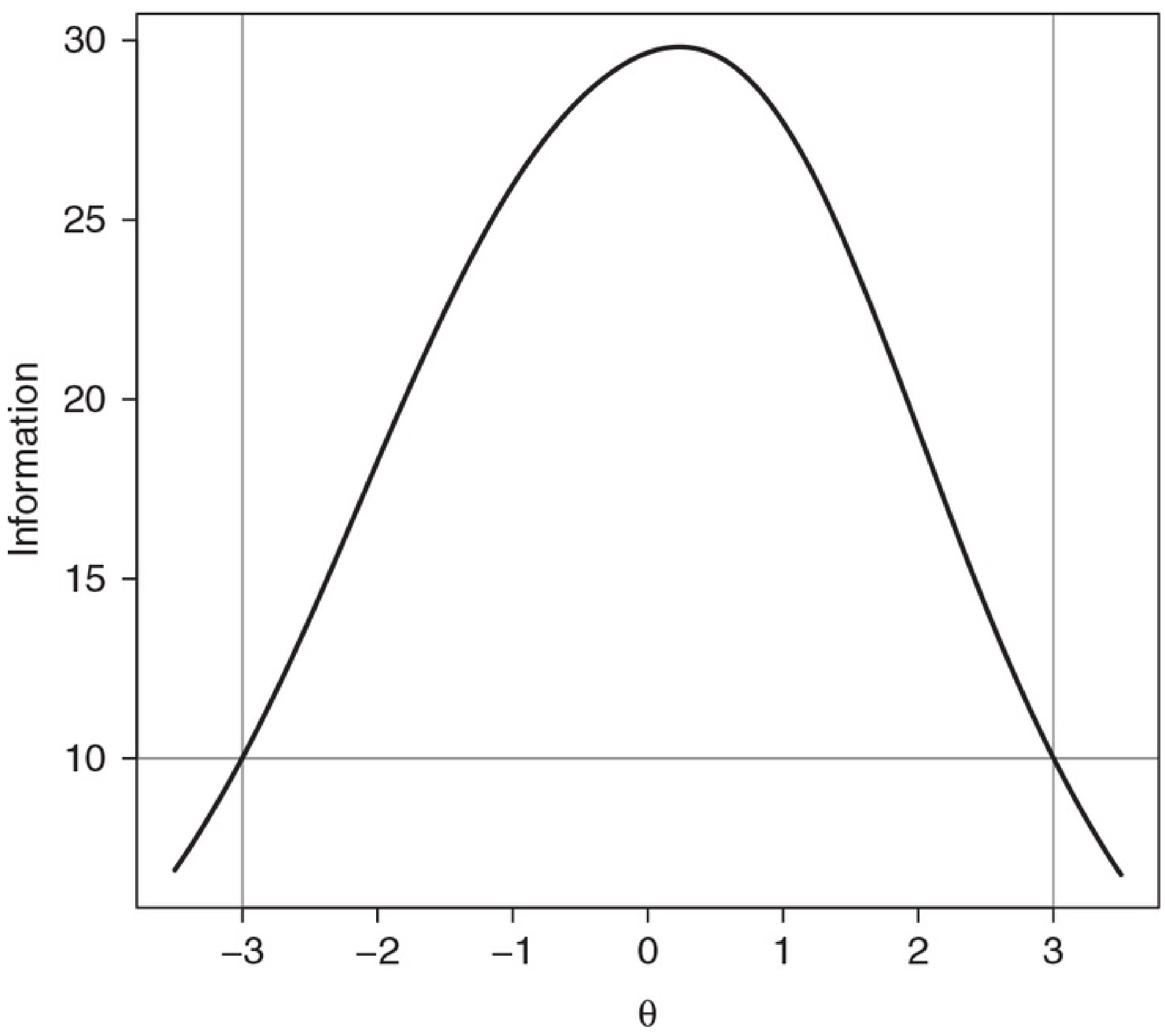
Statistical efficiency can be defined in many ways, but for point estimation, it is usually defined in terms of the mean squared error, . Assuming all other things remain constant (e.g., cost of the test and testing time), the goal would be to select the test that has the lower mean squared error. Unfortunately, finding the exact form of the mean squared error for ability estimates in IRT is difficult. However, as discussed earlier, the inverse of the TIF provides an asymptotic estimator of the mean squared error and, thus, is often used as the measure of efficiency of the test.

Using the TIF as a measure of the efficiency of a test makes it very easy for test developers and consumers to compare two competing assessments of the same construct. Suppose there are two tests with TIFs and . Then the relative efficiency function of Test 1 compared to Test 2 is defined as the ratio of the two information functions, . For example, if the information in the two tests about $\theta = 0$ is , then the relative efficiency of Test 1 compared to Test 2 is 1.2. One way that this is interpreted is that at $\theta = 0$, Test 1 is functioning as if it is 20% longer than Test 2 than Test 2.

Use in Test Construction

Because the TIF is a measure of how well a given test measures the latent construct θ, it is useful when constructing a test. For example, suppose test developers have a large pool of items from which they wish to construct a test. Furthermore, suppose that there is a desire to construct a test such that the conditional standard error of measurement is no greater than for ability levels in the range $-3 \leq \theta \leq 3$. Then, the test developer might attempt to find the smallest set of test items from the pool that produce a TIF that satisfies the constraint for all . For example, the TIF in Figure 2 was generated by selecting the smallest number of items that satisfy the aforementioned constraint from a pool of 200 2PL items. In this case, the test required 159 items to produce a test with test information greater than or equal to 10 across the range of ability levels between −3 and +3.

Figure 2 The TIF for the shortest test that has test information greater than 10 for all ability levels between −3 and 3 with items selected from a pool of 200 2PL items.



The test construction example in the previous paragraph assumes that the test developer wanted a test that would do an adequate job measuring ability over a broad range of the ability scale. In other situations, test developers might want a test that maximizes the information at a specific point on the ability scale. For example, suppose that a test developer wants to construct a test that would be used to identify individuals in the top 10% of the population. Then, if abilities are assumed to be normally distributed, the developer would want to develop a test that has a high level of test information at the 10th percentile of the normal distribution θ = 1.28. In this case, the developer would select the items with the highest IIFs at θ = 1.28.

Concluding Remarks

The TIF is extremely important as a measure of the quality of an educational or psychological test. However, it is not without its limitations. Firstly, it is important to understand that the TIF is not invariant to transformations of the latent ability. For example, suppose that a test developer wants to report test results on the transformed scale *g*(θ) instead of the original scale θ, then the information about *g*(θ) is equal to . Although the information itself is not invariant to transformations, the relative efficiency is, so, .

Secondly, using the inverse of the test information as estimate of the sampling variance of the ability estimate is really only appropriate for sufficiently long tests. While some research has shown that the inverse of the information works reasonably well for tests as short as 20 items, other estimates of the sampling variability should be explored for shorter tests.

Matthew S. Johnson

See also *a* Parameter; *b* Parameter; *c* Parameter; Item Information Function; Item Response Theory; Local Independence; Maximum Likelihood Estimation; Rasch Model; Standard Error of Measurement

Further Readings

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Wesley.

Casella, G. & Berger, R. L. (2002). Statistical inferece (2nd ed.) Chapter 7. Pacific Grove, CA: Duxbury.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory, Vol. 2. Sage Publications.

Samejiman, F. (1994). Some critical observations of the test information function as a measure of local accuracy in ability estimation. Psychometrika, 59(3), 307–329.

Jamie R. Mulkey Jamie R. Mulkey Mulkey, Jamie R.

Test Security

Test security

1674

1678

Test Security

Test security is the process of protecting assessments, so that the results from those assessments can be trusted and used to make important decisions about individual competence. According to the American Education Research Association/American Psychological Association/National Council on Measurement in Education standards, test users have the responsibility to protect the security of tests. Test security is particularly important when assessments are used, for example, in selection, accountability, credentialing, or diagnosis because of the inferences drawn from the assessment’s validity.

Test security is a relatively new field that has been gaining momentum since the early 2000s. For a long time, many testing organizations kept test cheating and theft to themselves. If there was a breach, such as a box of test booklets reported as missing or a test taker whose response patterns looked suspicious, the testing organization handled it quietly. Testing organizations did not want anyone outside their organization to know they had test security issues. However, since the early 2000s, the dialogue on cheating has emerged into a community-wide conversation. Many testing organizations now discuss test security issues in groups and forums to help each other understand the threats and risks around test fraud, theft, and cheating. Special interest groups work to create tools that help testing organizations address test security concerns, and benchmark studies are conducted to understand the breadth of test fraud as a problem. Process methodologies have been defined to help provide a structured way of thinking about test security. There are now published works, articles, and training that have helped to elevate the profession and professionalism of test security. There is even test security certification to qualify individuals as test security professionals. All this to say that test security has emerged as a subspecialty of great interest in the field of educational measurement.

What Is Test Fraud?

Test fraud can be broken down into two areas: test cheating and theft. Test cheating is the actual act of an individual or group of individuals who obtain unauthorized exam content prior to a testing event, giving them an unfair advantage because they have prior knowledge of the content to be tested. Examples of cheating can include a person copying the answers from another test taker, a teacher helping a student with the answers during a test, and an individual taking a test on someone else's behalf (proxy test taking). Test theft is the actual stealing of exam content. This act can occur by memorizing test content and then transcribing it for later use, downloading items from a test delivery system, stealing test booklets, or taking pictures of actual test questions. When test theft occurs, it is for the purpose of sharing or selling the test content, so that others may benefit from gaining prior knowledge before the test event.

Test cheating and theft are on the rise. Research studies by Rutgers University professor Donald McCabe suggest that graduate business college students tend to cheat more than their nonbusiness counterparts and that more cheating is going on because today's students do not consider what they are doing, cheating. There are occurrences of cheating and fraud that result from the unintended use of test scores, for example, tying school accountability and teacher performance to student assessment. There were many states and districts caught up in cheating scandals that stemmed from federal requirements of the No Child Left Behind Act of 2001. Under this act, schools and districts were required to show school improvement through annual statewide K–12 assessments or be subject to disciplinary action plans and sanctions. As a result, some school personnel coached students during testing, changed student answers, and identified low-performing students for noninclusion in testing events, in the name of making their school or district look like they were high performing. In 2015, the enactment of Every Student Succeeds Act introduced different school accountability measures, but it is unclear whether these new measures will drive similar behaviors.

When Do Test Cheating and Theft Occur?

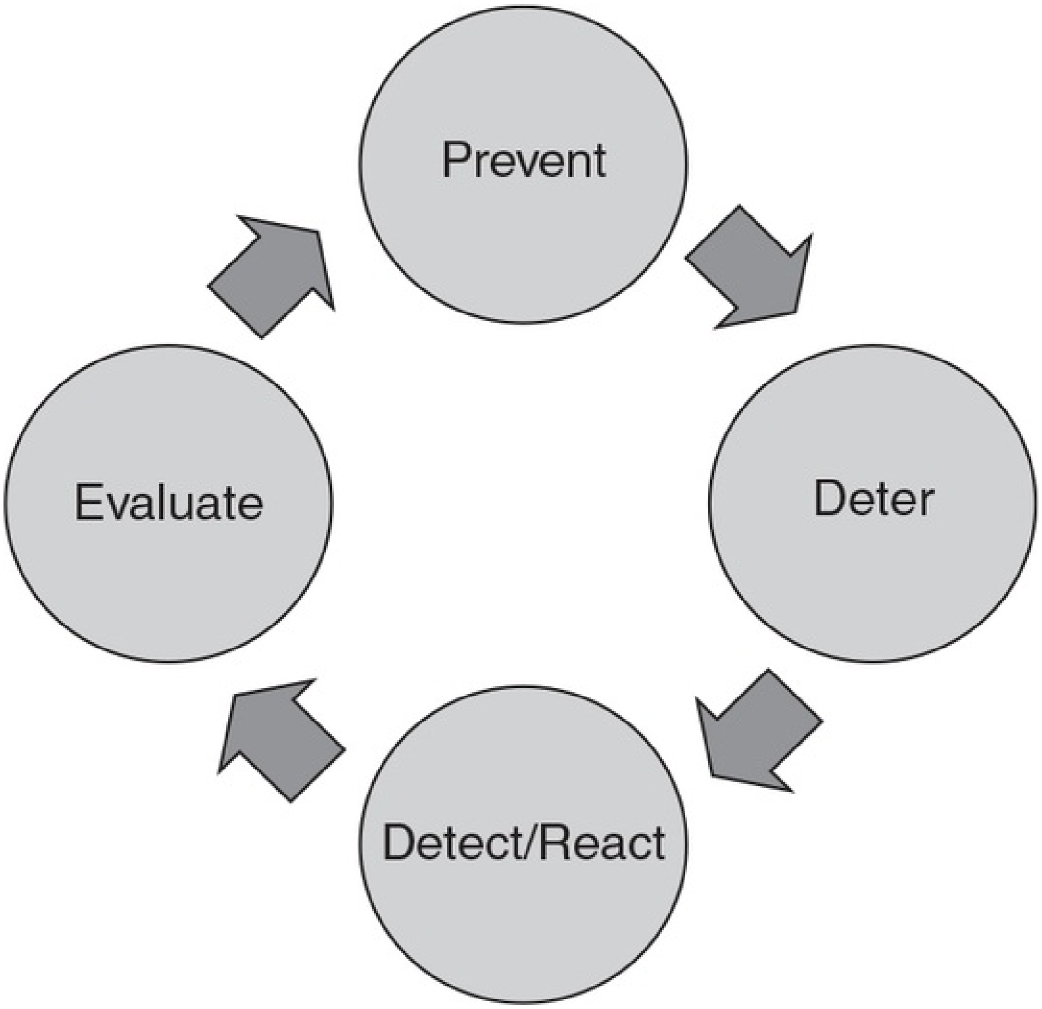
Test cheating and theft can occur when high stakes are associated with a test's outcome—that is, the outcome of the test will have an impact on test takers' lives or livelihood. Cheating and theft can occur in any profession or environment; no content area or genre is immune. There have been incidences of test theft in the military, where a department kept a filing cabinet of test answer keys available for up and coming midshipmen to successfully pass exams. Medical doctors have memorized test content and then provided it to a test preparation provider to help prepare other doctors for board certification exams. Employees from a test preparation company repeatedly took the multistate bar exam and then provided actual test questions as preparation to test candidates.

Test cheating and theft also cross cultural boundaries. In some cultures, such as China and India, collaborating on a test, sharing test content, or stealing test items is seen as a necessity. For example, competition in India and China to get into the next level of education is so fierce, students have received cheat sheets from parents, been caught with microscopic cameras and audio receiving devices, and used proxy test takers to complete these admissions exams on their behalf.

Is There a Framework for Thinking About Test Security?

Like any discipline, it helps to have a framework or model for thinking about how that discipline works. Test security is no different. Test security's framework is similar to that of W. Edward Deming's quality assurance circles: plan, do, check, and act. One test security framework defined by David Foster is prevent, deter, detect/react, and evaluate (see Figure 1). The security policies and practices are designed, communication strategies are used to communicate the consequences of test theft and cheating, test security incidents are detected, there is response and action plan to those incidents, and then refinements are made to help ensure a similar test security incident doesn't happen again.

Figure 1 The Test Security Framework



Source: Caveon Test Security.

Prevent

It is important for a testing organization to determine the level of risk around its test security vulnerabilities and to create a plan to address mitigating and managing that risk. The prevent phase of test security deals with protecting the intellectual property and preventing test fraud from occurring. Examples of prevention include creating a test security plan (e.g., a document that dictates an organization's policies, processes, and procedures around test security), creating robust candidate or test taker agreements that stipulate sanctions for cheating and test theft, and creating organizational awareness of test security by training staff on a regular basis.

Deter

Deterrence involves communication. One of the best ways to prevent test security incidences from occurring is to broadly communicate the impact of cheating and test theft to constituents. This communication acts to dissuade would-be cheaters and thieves from attempting inappropriate actions and behaviors. Other deterrence strategies include conveying agreements with stringent sanctions for test taker misconduct and frequently defining in newsletter articles and website notices what inappropriate test-taking behaviors are and discussing their associated consequences.

Detect/React

Even the best preventive measures may not stop those who are intent on gaining an unfair advantage when it comes to taking a test. The detect phase deals with investigating and measuring whether test fraud is occurring. Monitoring the Internet and social media sites routinely to detect for stolen or shared test content helps a testing organization know whether their test content is at risk. Are their actual test questions available? How long have they been available? What is the likelihood test takers are using this unauthorized test content for study purposes?

Data forensics, or conducting statistical analysis to detect for aberrance or irregular response patterns, helps a testing organization not only to determine the health of the exam but also to detect response patterns that are indicative of test cheating and theft. Data forensic analysis can detect such things as answer copying, proxy test taking, group cheating, and preexisting knowledge of test content.

More and more, data forensic results are being used as evidence in academic integrity inquiries, teacher licensure hearings, and courts of law. These analyses provide the statistical probability of whether or not the test results are valid. In some cases, the probability of an individual’s test response patterns is so extreme or so unlikely, the test results cannot be trusted and the score is invalidated. The important point is that the test results are irregular and therefore are not an adequate measure of an individual’s performance. The individual is not being labeled as a cheater. Rather, the score results are not valid and consequently must be thrown out.

When a testing program wants to know how a group of test takers has identical responses or who is to blame for stolen test booklets, an investigation may be required. Investigating can also be part of the react phase. It is the process of determining the root cause of the test security breach.

Evaluate

The evaluation phase reviews how test security incidents occurred and explores how improvements to the process can be made to prevent similar incidents from occurring again. For example, if an entire form of a test is found on a social media site, the testing organization should have a reserve form available to replace disclosed test content, should a similar incident happen again.

What Are the Biggest Areas of Concern for Test Security?

For the many established methods used to combat cheating and theft, the number of techniques used to thwart these practices is increasing as well. Cheating devices are getting smaller and smaller. There are wireless ear phones that are so small; it takes a magnet to remove them. Miniaturized spy cameras, as small as pin holes, can be found in pens, fake buttons, and ball caps. Watches are no longer allowed in many test administration environments for the fear of Internet access or the recording of test content. Although thorough test administration processes are practiced, there may still be individuals who slip by or slip into the restroom on a break to text a collaborator for test answers.

Test administration, which is a small percentage of the entire testing process (from design to item development, through score reporting), is likely to be the area where unauthorized use or illicit access to test content will occur. Testing professionals need to get ahead of the problem and be proactive versus reactive to test security incidents.

What Do Preventive Test Security Strategies Look Like?

Proactive and preventive test security strategies begin with secure test and item design. For example, a testing organization decides during the design phase that their test of 75 items will be delivered in a computer-based environment; it will have two forms, with an additional form in reserve (in the case that one form is breached). All the items and options when delivered will be scrambled, so that individuals seated adjacently will not be able to copy from each other. Secure design strategies, such as these, are used to preserve the hard work and resources used to develop items and to reduce item disclosure.

There are new item types that also help reduce item exposure and minimize memorization. An example is the discrete option multiple-choice item. This item type presents the question and then presents one answer option at a time. With the random delivery of multiple correct and incorrect responses, each test taker will receive a unique item each time the test is delivered, making it difficult to memorize or capture test content.

There are other strategies such as stronger identification protocols to identify high-risk test takers, development of item variants (clones) to continuously refresh test content, and strong test taker policies that deter test theft and cheating from occurring.

What Does the Future of Test Security Look Like?

There is no single answer to solving the problem of test fraud (cheating and theft). It will take a multipronged approach of educating test takers, developing new methods of test taker authentication, designing test and item construction strategies that limit test content memorization, and understanding the risk level for the assessment being used. Fundamentally, test security means there is confidence in the competence of the individual, the appropriate skills and knowledge are being measured, the assessment used to measure the competency is a valid, and the authenticated individual is the one being assessed.

Jamie R. Mulkey

See also Cheating; Computer-Based Testing; Computerized Adaptive Testing; Conflict of Interest; Every Student Succeeds Act; Item Analysis; No Child Left Behind Act; Testwiseness; Trustworthiness

Further Readings

Standards for educational and psychological testing. (2014). p. 168. Washington, DC: American Educational Research Association.

Associated Press. (2014, February 5). Exam cheating scandal hits naval nuclear training school. Retrieved June 4, 2016, from http://nypost.com/2014/02/05/exam-cheating-scandal-hits-navy-nuclear-training-school/

ATP Test Security Committee. (2013, September 10). Association of test publishers security survey report 2013. Retrieved May 29, 2016, from https://www.createspace.com/4418924

CBS/Associated Press. (2015, March 20). Indian parents scale school wall to help students cheat on exams. Retrieved June 11, 2016, from http://www.cbsnews.com/news/indian-parents-scale-school-wall-to-help-students-cheat-on-exams/

Certified Exam Security Professional program. (2014, June 06). Retrieved May 29, 2016, from http://www.cesp.org/

Foster, D. F. (n.d.). Discrete option multiple choice. Retrieved May 29, 2016, from http://www.trydomc.com/

Foster, D. F. (n.d.). Exam security breaches: Summaries and resolutions (pp. 2–45, Working paper). Retrieved from http://nebula.wsimg.com/21c04f160f3490dd071ef631f125cf66?AccessKeyId=1117C2B103CC5C76116D&disposition=0&alloworigin=1

Garger, K. (2015, August 8). Firefighters “cheated” on exam during bathroom breaks. Retrieved May 13, 2016, from http://nypost.com/2015/08/18/firefighters-accused-of-taking-bathroom-breaks-to-cheat-on-exam/

Jordan, M., & Belkin, D. (2016, June 5). Foreign students seen cheating more than domestic ones. The Wall Street Journal, pp. 1–14. Retrieved June 7, 2016, from http://www.wsj.com/articles/foreign-students-seen-cheating-more-than-domestic-ones-1465140141

Moneo, S. (2014, April 24). Cheating said to be on rise in North American B-schools. Retrieved June 11, 2016, from http://www.theglobeandmail.com/report-on-business/careers/business-education/cheating-said-to-be-on-rise-in-north-american-b-schools/article18077376/

Olson, J., & Fremer, J. (2013). TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities. Washington, DC: Council of Chief State School Officers.

Olson, J., & Fremer, J. (2015). TILSA test security: Lessons learned by state assessment programs in preventing, detecting, and investigating test security irregularities. Washington, DC: Council of Chief State School Officers.

Severson, K. (2011, July 5). Systematic cheating is found in Atlanta's school system. The New York Times. Retrieved June 4, 2016, from http://www.nytimes.com/2011/07/06/education/06atlanta.html?_r=2

Slass, L. (2010, June 17). ABIM reaches a settlement with Arora board review. Retrieved June 1, 2016, from https://www.abim.org/news/abim-reaches-settlement-with-arora-board-review.aspx

Slobodzian, J. A. (2006, August 6). Barexam prep firm out of line, judge says The ruling, awarding $12 million in damages, says Multistate Legal Studies used questions taken from tests, violating copyright. Judge finds against bar exam prep firm. Retrieved June 4, 2016, from http://articles.philly.com/2006–08-26/news/25396298_1_barexaminers-test-questions-fullam

Wollack, J. A., & Fremer, J. (Eds.). (2013). Handbook of test security (Chap. 3–

8, p. 10). New York, NY: Routledge.

Matthew B. Fuller Matthew B. Fuller Fuller, Matthew B.

Testing, History of

Testing, history of

1678

1685

Testing, History of

The history of testing is diverse and lengthy. This entry attempts to condense major historical developments in educational and psychological testing. To accomplish this, the entry first defines testing and offers definitions of just two interrelated variants of testing: psychological testing and testing in educational settings. The entry is organized according to different historical vignettes to demonstrate shifts in the development of the social and educational testing movements. Although the entry focuses on educational and psychological testing, these movements have always been connected to social, political, and technological developments. As such, the entry includes references to interrelated movements that influenced testing.

After defining the term *testing*, the entry focuses on similarities and uses in educational and psychological settings. Next, the bulk of the entry focuses on historical vignettes, starting with the Chinese Imperial Examination system and concluding with contemporary legislative and social movements in testing and measurement-based accountability. Throughout the entry, the following claims will be substantiated: (a) testing has always been connected to and influenced by technological advances, (b) testing has been one way in which social and legislative connections to educational institutions are made, and (c) tests have always been scrutinized for biases in their capabilities.

Definition of Testing

Testing is defined as the revelation of a person’s capabilities by examining their response to a situation, prompted problem, or question. Tests, therefore, are delineated as the instruments, sets of questions, problems, or physical responses used in determining a person’s capabilities. This definition is the one used in this entry because, as with any historical concept, one must consider what is being examined in order to examine its history.

Testing has been a concept of particular importance in psychological and educational settings. Tests have long been used to examine a patient’s or client’s mental or psychological state. For the purpose of this entry, a *psychological test* is defined as any examination or observation of an individual’s mental state, behaviors, or any number of cognitive or noncognitive constructs. For the purpose of this entry, *educational tests* are unique psychological tests defined by their use in educational institution settings and with the intent of examining students’, teachers’, or school leaders’ abilities. The following brief historical vignettes trace the chronology of psychological and educational tests, beginning first with the Chinese Imperial Examination system. Special attention is paid to social and political movements as well as technological advances influencing testing.

History of Psychological and Educational Testing

As early as the Zhou Dynasty (1046–249 BCE), citizens were given a promotion within the bureaucratic structure of the Emperor’s court based upon demonstrated skill in archery. It was not, however, until the Sui Dynasty (581–618 CE) that a system requiring performance on a written examination, as well as martial arts and archery, emerged to sort capable citizens into leadership positions in the Imperial court. These early civil service examinations required citizens from different precincts and regions to participate in standardized written, oral, and observed examinations of one’s ability to recite important moral and philosophical arguments, recite texts the Emperor wrote, and perform martial and military arts. The use of the Imperial Examination system was a direct result of social shifts away from a feudal system of patronage as a means to gain improved social status toward a more meritorious system. Test takers from each precinct would engage in the same test, which occurred in regular cycles and contained the same questions and instructions administered by specially trained test administrators. Each test progressed in difficulty, with the highest level, the Palace Examination, often being supervised by the Emperor.

During the Song Dynasty (960–1279), the school system was expanded considerably and along with it, the Imperial Examination system. In this time frame, the examinations contained standardized tests administered at the district, provincial, and metropolitan levels and were attached to the bestowing of an educational credential. Strict quotas allowed only a small number of test takers to pass each exam, and students often took the tests multiple times before passing them, often waiting three or more years before the next test cycle.

Exams were a test of the candidates’ physical abilities and intellect. By 115 CE, during the Han Dynasty, the school curriculum and examinations focused on music, math, writing, Chinese traditions and ceremonies, archery, and horsemanship. The curriculum and exam would eventually evolve to also include militaristic strategy, civil law, taxation, agriculture, geography, and Confucian philosophy. Following each examination a test proctor “called the roll” and announced each test taker’s scores, a practice familiar to modern instructors. The most accomplished students were said to take on God-like qualities once passing the highest levels of exams. This was due, in large part, to the fact that each exam was a grueling, 3-day experience. Across 3 days and 2 nights, exam takers were ushered to a tiny, outdoor cubicle wherein he would replicate the exact text of an entire essay made available from the Emperor. No interruptions were allowed. Candidates had to supply their own food, water, and bedding. If bad weather was present, tests were not rescheduled. Instead, test takers would simply have to make do with the constraints given them. Occasionally, test takers died during the exams and their bodies were simply thrown over the walls of the grounds so as not to distract other test takers.

By the start of the Sui Dynasty (518), the Imperial Examination system was institutionalized as an expectation for those citizens hoping to improve their social status and gain service in the Emperor’s Court. As the Imperial Examination system began to solidify in its structure as the earliest form of a nationwide standardized examination system, connections to technology, cheating, and bias also solidified. The intense pressure to succeed and the high stakes of the examination—at the highest level, successful candidates and their families were often invited to live in the Emperor’s Court—meant students went to extraordinary lengths to cheat on exams. Moreover, test proctors also implemented security technologies. In any given administration, test takers would have to provide specially colored ink made from octopus ink available to them only after traveling to the city for the exam. This was designed to discourage test takers from bringing in copies of essays they were to replicate during the exam. Test takers were inspected and some were found to be sneaking in copies of essays printed on the inside of their undergarments. To prevent situations in which bribes were written into exam responses or reviewers recognized family names or calligraphic pen strokes indicative of a specific students' hand, tests were copied by specially trained calligraphers who offered a normalized script for reviewers. Also, multiple readers reviewed the examinations and students were given numbers to reduce the likelihood of a family's name influencing the outcome of a grade.

Test takers would often have to arrange for travel to the cities to complete the exam, often taking months or years to complete. Once in the cities, test takers would secure long-term housing and food arrangements, meaning only those candidates with sufficient financial means could afford to participate in the highest examinations. Similarly, more affluent test takers could afford to purchase copies of essays to study and a viable economy for tutors was in place by the middle of the Sui Dynasty and continued well into the Song Dynasty. However, this growth did not cease with the Song Dynasty. Throughout the Ming Dynasty (1368–1644) and Qing Dynasty (1644–1911), exams continued to grow in complexity and significance. By the start of the 20th century, calls for educational reform and the development of new nationwide standards for education in China saw many reformers focused on calls for a new examination system. In 1905, the Imperial Examination system was formally disbanded with the promise of a newly developed exam in the near future. However, within 6 years, the Qing Dynasty was overthrown and exam reforms were left unresolved.

However, the influence of the Imperial Examination on modern examination systems are not too remote. Political leaders in Vietnam, Japan, India, and the United Kingdom took note of the system as an effective means of implementing meritocratic bureaucracies. In 1808, Napoleon's founding of le baccalaureate exam was influenced in part by the success of the Imperial System. In England, in February 1854, Sir Charles Trevelyan and Sir Stafford Northcote published a refutation of the British system of patronage. Besides refuting the system of patronage, the Northcote-Trevelyan Report (1854, p. 6) recommends: "The first step toward carrying this principle into effect should be the establishment of a proper system of examination before appointment [to civil service]." Ssu-yu Teng notes that Trevelyan and Northcote relied on the Imperial Examination system when crafting their report.

Following the British lead, at least four U.S. Presidents—Grover Cleveland, William McKinley, Theodore Roosevelt, and Chester Arthur—in their State of the Union addresses would hold up the competitive examination as a success of their administration. The idea of examinations adorned American philosophical discussions for decades. Thomas Jefferson's *Notes on the State of Virginia* (1832, p. 13) laid out a plan for education in Virginia: "twenty of the best geniuses will be raked from the rubbish annually and be instructed, at the public expense." Examinations were to be Jefferson's tool for accomplishing this meritocratic end.

English proficiency examinations were also recommended in the 1907 Dillingham's Congressional Commission on Immigration, a precursor to the Immigration Act of 1917. The act required literacy and citizenship tests, physicals, and mental examinations of immigrants seeking citizenship. According to Raymond Fancher, noted eugenicist Henry Goddard authored the Citizenship Exam, which was written only in English and favored an understanding of American social norms. By the start of the 20th century, testing had a strong foundation on which to rely as it made more direct connections to educational settings.

Measuring General Intellect

By the mid-19th century, concerted efforts on the part of eugenicists and statisticians to classify students and citizens according to *general intellect* emerged. In the United Kingdom, Sir Francis Galton, a eugenicist and statistical pioneer, introduced the use of questionnaires to collect data on his theory that intelligence was inherited. Galton recognized the need to measure mass numbers of citizens and, in the late 19th century, developed a standardized measure of participants' hereditary traits and reasoned that these and other hereditary traits correlated to intellect. Although Galton is widely considered the father of mental measurement, his hypotheses were never sufficiently examined or proven.

In the United States, James McKeen Cattell began administering a battery of tests of human memory and reaction time to a series of problems to students at the University of Pennsylvania. In 1890, when Cattell published *Mental Tests and Measurements*, he coined the term *mental measures* and offered his opinion of the burgeoning profession of psychology. When Cattell moved to Columbia University, Cattell's tests were required of all freshman students.

During the late 19th century, France mandated compulsory education for all children between 6 and 14 years of age. French psychologist Alfred Binet devised a battery of questions aimed at categorizing intellectually slow children for exclusion from schooling. In 1903, Binet published his methods for examining students in *L'Etude experimentale de l'intelligence* (*Experimental Studies of Intelligence*). Binet, with the help of his young research assistant, Theodore Simon, began expanding and revising his scale to include age-relevant measures. The Binet and Simon Test of Intellectual Capacity was a list of 30 tasks arranged in order of increasing difficulty and specialized for a variety of age ranges. A test administrator would ask test takers to perform a variety of tasks or respond to questions orally. A student's score on the test would reveal the student's mental age, which would in turn be compared to the student's chronological age. Those students with the greatest disparities between their mental and chronological age were said to be unfit for schooling.

Goddard, himself a renowned eugenicist, was commissioned by the superintendent of the New Jersey Training School for FeebleMinded Girls and Boys to develop a system for examining mentally challenged students (then called, feebleminded and morons). During the summer of 1908, Goddard traveled to Europe, engaging other scholars who were testing mentally challenged students. During his travels, he learned of Binet and Simon's work and began translating their test to English and modifying it for American settings. By December 1908, Goddard published his version of the test and began promoting its use aggressively throughout U.S. schools. At Stanford University, Lewis Terman learned of the Binet and Simon Test of Intellectual Capacity and, in 1916, offered a revised version of the test, just 11 years following Binet's first use of his test. Considering challenges in communication at the time, the development of such mass testing of human intellectual abilities from Binet to Goddard to Terman begets the importance scientists placed on mental tests.

At approximately the same time Binet was developing his Test of Intellectual Ability, American psychologist Edward Thorndike was developing standardized examinations in reading, writing, and arithmetic. Binet's original test contained 30 tasks, verbal questions, or pictorial questions. For example, students, according to their age, might have been asked to cut a picture out of a piece of paper, do math to compare weights, write a word that rhymed with another word, name or describe a picture, or circle an object that a test proctor named.

Thorndike marketed his examinations specifically to schools and later predicted that schools and parents would engage in competitive test taking to improve individual students' abilities to gain access to better school programs and more individual students' abilities to gain access to better school programs and more prestigious colleges. In 1900, the College Entrance Examination Board, known today at the College Board, was founded by 12 Ivy League and exclusive institutions in the Northeast of the United States and sought a means of sorting the most intelligent students out from less intelligent students.

In 1917, the relatively young American Psychological Association, through its Committee on the Psychological Examination of Recruits, began developing a number of tests aimed at sorting U.S. Military men during World War I. The committee Chair, Robert Yerkes, led the development of two group-normed examinations of mental intelligence aimed at sorting men into officer and general enlistment ranks. Yerkes's work required the support of the Army and, through the Surgeon General's office, Yerkes was introduced to a young 1st lieutenant in the Sanitary Corps, Carl Brigham. Brigham worked closely with Yerkes to develop a system for testing men as they entered the army and would later publish an influential text on the study, *A Study of American Intelligence*. Yerkes's Army Test Alpha and Army Test Beta consisted of timed batteries of responses to prompts and questions. Groups of men, sometimes as large as 500, were seated in an examination hall and given a blank piece of examination paper that had questions or prompts. Men had roughly 50 minutes to circle, underline, or cross out the appropriate response to every question. The tests made use of Guttmann scaling techniques and arranged questions in order of increasing difficulty. Following the initial development and administrations of Army Test Alpha, Army Test Beta was developed in response to criticism that illiterate or non-English-speaking men were unfairly assessed via the English-only Army Test Alpha and thus often placed into front-line service in World War I. A committee of seven, including Yerkes, developed a test for illiterate and non-English-speaking service men, drawing heavily on pictorial representations and gestures to administer exams. Both examinations were criticized for bias and errors that had tremendously dire results. Those who were illiterate and those without an in-depth knowledge of American social norms or marketing campaigns scored lower on the exams. Still, by the close of World War I, Army Tests Alpha and Beta had been administered to 1.7 million men. These tests were administered by a single, human proctor to several men, but technological advances in the coming decades would provide for the advancement of a system of mass testing in educational organizations and society. In particular, the founding of IBM in 1911 through a merger of three successful companies would influence testing for decades to come.

Spurred by his work in testing Army recruits, Brigham developed his own test for use in educational settings in 1925. In his 1926 text, *A Study of American Intelligence*, Brigham stated, “American intelligence is declining, and will proceed with an accelerating rate as the racial admixture becomes more and more extensive” (p. 210). Brigham also noted the increasing specialization of psychological testing as “rather hard to explain to the layman, who is familiar only with the ‘school teacher’ type of examination” (p. 57). Brigham developed a test consisting of math, reading, vocabulary, and grammar-related questions. For example, in the math section, test takers would respond to a line of numbers arranged in a pattern, and the test taker would have to discern which numbers would come next in the pattern. For vocabulary, test takers would have to circle worlds that were related or synonyms or antonyms of a given word. Later revisions of the test would include comprehension questions, wherein the reader was asked to respond to questions after reading a paragraph. In 1925, Brigham administered his test, the Princeton Test, to incoming Princeton freshmen. One year later, the College Entrance Examination Board asked Brigham to develop a test that could be used across all College Board institutions as a means of selecting the brightest students for admission. What emerged was a revised exam drawing heavily from Brigham’s work with Army tests named the Scholastic Aptitude Test (SAT). By 1926, the SAT was administered to high school seniors as a means of identifying the best and brightest for college admission.

In the late 1920s through the 1940s, while the rest of the country was dealing with the crippling effects of the Great Depression and World War II, mass testing as an industry was just beginning to blossom in the United States. In the early 1930s, Brigham’s work at Princeton was introduced to Harvard University President, James Bryant Conant. Conant was hoping to find a test that could be used to award scholarships to incoming freshmen at Harvard. Nicholas Lemann argued that Conant’s use of Brigham’s test and work coincided with growing pressure to limit the number of Jewish students at Harvard. Jewish student enrollment at Harvard tripled to 21% of the freshman class in 1922 from about 7% in 1900. Brigham’s eugenics-driven research supported Harvard’s ability to exclude Jewish students from scholarship awards, thereby limiting the size of the Jewish student population. Conant assigned Henry Chauncey and Wilbur Bender the tasks of finding a test that could be used to select intelligent students for scholarships. Conant, Chauncey, and Bender traveled to Princeton and met with Brigham to learn more about his work. By 1934, the SAT was administered to Harvard freshmen as a means of selecting students for scholarships. A year later, Harvard required the examination of all candidates. Eventually most institutions of higher education followed suit.

In the same year, high school science teacher Reynold Johnson devised a system for reading pencil marks on a piece of paper and burning and later punching out a hole if a question was correctly answered. This marked or graded paper would allow a test administrator to score a respondent’s test (i.e., how many questions were correctly answered) in a fraction of the time typically required to score tests. Columbia University Professor Benjamin Wood sensed the potential of Johnson’s machine, called the Mark-O-Graph, and began administering his tests using Johnson’s machine. Wood even developed a new system for testing that included a test booklet consisting of several multiple-choice questions that were to be marked by the test taker on a separate answer sheet that could be scored by a machine or a teacher using a punch card stencil containing the proper answers. Test grading, which had formerly required many hours of a highly paid, skilled expert’s time, could now be completed for more test takers, in a fraction of the time, and with little or no skill requirements for the grader.

Toward the end of his life, Brigham’s beliefs in eugenics had begun to shift and he recognized the bias in his exams. In 1930, Brigham recanted his beliefs in racial superiority of Caucasian races and acknowledged his prior conclusions were “without foundation” and stated “that [his work with Army test,] with its entire hypothetical superstructure of racial differences collapses completely” (p. 164). Brigham would continue this line of argumentation and withhold the SAT from mass testing until his death in 1943. His death coupled with IBM’s purchase of Johnson’s Mark-O-Graph design, however, paved the way for the SAT’s mass use as a college admissions exam beginning in 1943. In this same year, the Army-Navy College Qualifying Test was administered to 316,000 U.S. high school seniors, demonstrating that mass, standardized multiple-choice tests can be administered successfully. Throughout World War II, American GIs would continue to take Army tests.

Emboldened by these successes, Chauncey founded a corporation aiming at advancing the cause of the fledgling testing industry. Inspired by the acronym of his father’s alma mater, the Episcopal Theological Society, Chauncey created the Educational Testing Service (ETS) and would serve as ETS’s first President while Conant served as Chairman of the board. Increasing college enrollments due to compulsory public education laws in decades prior and the GI Bill meant more men and, for the first time, women needed to be tested. The young corporation had a tremendous supply of test takers, technology at its disposal to support large administrations, and a demand for test scores, making it a billion support large administrations, and a demand for test scores, making it a billion dollar organization by 1969.

But almost immediately, racial and ethnic biases were suspected in the SAT. As almost every researcher who does so finds, when SAT scores are viewed according to racial and ethnic groups, cultural biases are noted. In the 1930s and 1940s, landmark studies by anthropologists Franz Boas and Ruth Benedict characterized the whole notion of intelligence testing and tests in general as biased and systemically racist. Racial, gender, and socioeconomic bias has been a consistent criticism of testing efforts.

In 1946, after a distraught Jewish woman who had been denied admission to college asked Stanley Kaplan to help her prepare for the SAT, Kaplan spent the next few years devoted to standardized test tutoring. Most of his students would be the grandchildren of Eastern European Jewish immigrants from Brooklyn, the very people Brigham feared would water down American intelligence and whom Chauncey and Conant sought to exclude. Kaplan's SAT tutoring business had a significant influence on the national conversation about tests. In the eyes of ETS's leadership, the SAT, like Galton's concept of intelligence, was uncoachable and ETS psychologists called upon validity and reliability studies to discredit Kaplan's efforts. However, nearly three decades of studies determined that coached test takers did score better than if they were not coached. By the 1970s, ETS and American College Testing Program both began offering test preparation services.

Following World War II, ETS began to develop the familiar structure of the flagship test, the SAT. The familiar structure of the SAT Reading exam—complete with reading comprehension, analogies, antonyms, and sentence completion questions—was in place by 1952. By 1957, the number of test takers in a given year surpassed half a million. Today, approximately 1.65 million high school students take the SAT each year. ETS developed a unique industry and several organizations joined the mix in the 1940s and 1950s. In 1959, the American College Testing Program emerged as a lead competitor to the SAT in the testing marker.

The 1960s and 1970s saw an increased legislative focus on testing as a form of educational accountability. Since the 1947 Truman Report, presidential commissions and governmental hearings on education and testing have been commonplace. President Eisenhower's Committee on Education Beyond the High School (1956) and Kennedy's Task Force on Education (1960) are just a few examples of government commissions that made an impact in their time. In 1965, President Lyndon B. Johnson signed into law the Elementary and Secondary Education Act (ESEA) and the Higher Education Act. Both pumped new financial resources into education. However, both also called for increasing systems of assessment on behalf of the states. In elementary and secondary education, the National Assessment of Educational Progress was developed as “the nation’s report card.” Testing, particularly standardized testing in each state, became a popular means for policy makers to gauge return on investment in educational agencies. During the 1960s, when U.S. Attorney General Robert Kennedy brought hundreds of suits against schools he claimed were segregated and offering inequitable schooling, he called upon test scores in leveling his claims.

The 1970s and 1980s saw a flurry of state-level responses to federal pressures for increased accountability. States such as Texas, California, Iowa, Pennsylvania, and Michigan began developing comprehensive systems of regular, statewide testing throughout the 1970s and 1980s. By the 1990s, every state had developed comprehensive, statewide systems for testing school-aged students with a few developing tests for college students as well. Many states began developing a systematic approach to testing and accountability that educators began characterizing as high-stakes testing. This concept holds that it is not the test itself that is high stakes but dire consequences of poor individual or group performance on the tests. Under a high-stakes testing system of accountability, schools with poor school-wide or subgroup performance on tests face penalties or state agency takeovers of the school district. Proponents of such systems claim the tests are a clear articulation of necessary educational attainment goals and outcomes. Opponents often argue that to deprive poor performing schools of resources only further reinforces challenges they face. Moreover, poor performing schools point to student demographics as a means of illuminating challenges they face in educating traditionally underprepared, underrepresented student populations.

The Elementary and Secondary Education Act was the foundation of most modern legislation pertaining to education and educational testing. In particular, the Elementary and Secondary Education Act faced reauthorization in 2001 under President George W. Bush. The reauthorization, commonly known as the No Child Left Behind Act of 2001, called for increased accountability and governmental oversight for schools that did not meet Adequate Yearly Progress. Schools had to have an increasing percentage of fifth graders, for example, who met statewide standards in test performance. However, criticism over a lack of met statewide standards in test performance. However, criticism over a lack of funding to support school improvement and philosophical differences saw the U.S Congress reverse elements of No Child Left Behind Act in 2015.

More recently, political movements aimed at nationwide coordination of standards, under the banner of Common Core State Standards, have led to a cadre of subject area tests on which schools in participating states have chosen to test students. A number of statewide and nationwide advocacy organizations have emerged calling for more responsible, learner-centered approaches to testing. Moreover, a number of states have also passed laws limiting the number of tests in which students can participate. Clearly, the history of testing is a complex and storied endeavor. Testing has always been influenced by social developments, technology, legislative guidance, and student and family needs. Time will tell as to the direction of future efforts in educational testing.

Matthew B. Fuller

See also ACT; Admissions Tests; Educational Testing Service; Ethical Issues in Testing; High-Stakes Tests; SAT; *Standards for Educational and Psychological Testing*; Stanford-Binet Intelligence Scales; Test Battery; Tests

Further Readings

Benedict, R. (1943). The races of mankind. New York, NY: Public Affairs Committee Incorporated.

Berliner, D. C., & Biddle, B. J. (1995). The manufactured crisis. New York, NY: Addison Wesley.

Boas, F. (1938). The mind of primitive man. New York, NY: Kessinger.

Brigham, C. (1923). A study of American intelligence. Princeton, NJ: Princeton University Press.

Brigham, C. (1930). Intelligence tests of immigrant groups. The Psychological Review, 37(1), 158–165.

Cross, T. (1998). Explaining the gap in black-white scores and IQ and college admissions tests. Journal of blacks in higher education, 01(18), 84–97.

Fancher, R. (1985). The intelligence men: Makers of the IQ controversy. New York, NY: W. W. Norton & Company.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. Harvard Educational Review, 73(1), 1–43.

Giordano, G. (2005). How testing came to dominate American schools: The history of educational assessment. New York, NY: Peter Lang Publishers.

Green, D. R. (1981). Racial and ethnic bias in achievement tests and what to do about it. Journal of educational measurement, 18(2), 1–6.

Jefferson, T. (1832). Notes on the State of Virginia. Boston, MA: Lilly and Wait.

Lemann, N. (1999). The big test: The secret history of the American meritocracy. New York, NY: Farrar, Straus, and Giroux.

Northcote, S. H., & Trevelyan, C. E. (1854). Northcote–Trevelyan report. London, UK.

Ravitch, D. (2011). Death and life of the great American school system: How testing and choice are undermining education. New York, NY: Basic Books.

Ravitch, D. (2014). Reign of error: The hoax of the privatization movement and the danger to America's public schools. New York, NY: Vintage Books.

Teng, S. Y. (1943). Chinese influence on the Western examination system.

Harvard Journal of Asiatic Studies, 7, 267–312.

Young, M. (1958). The rise of meritocracy (T. Parsons, Trans.). London, UK: Thames and Hudson.

Hong Jiao Hong Jiao Jiao, Hong

Dandan Liao Dandan Liao Liao, Dandan

Testlet Response Theory Testlet response theory

1685

1688

Testlet Response Theory

A testlet or an item bundle refers to a group of interrelated items presented as a single unit. Often, a testlet consists of several items following a single stimulus. It is a commonly used test construction unit in large-scale assessments and often provides a context or situation for assessing knowledge, skills, and ability. Passages in reading comprehension tests, scenarios in science tests, and graphs and/or tables in math tests are such examples in practice. When items are constructed around such a common stimulus, items associated with the same stimulus are connected by the common context. Item connection or clustering may affect an examinee’s performance on those items due to the common contextual effects. Thus, local item dependence (LID) or testlet effects may be induced. When LID is present, an examinee’s response to an item may affect the examinee’s response to other items in the same testlet given the person and item parameters. Testlet response theory models these effects. This entry describes testlet response theory and its models then discusses the estimation methods of the model parameters.

Standard item response theory (IRT) models are not robust to the violation of the local item independence assumption. LID affects model parameter estimation, equating, and estimation of test reliability. Possible causes for LID include passage dependence, item chaining, explanation of previous answers such as clueing, item or response format (multiple-choice *vs*. constructed response items), scoring rubrics, fatigue, speededness, and practice effects. Passage dependence refers generally to item clustering around a common stimulus.

One method documented in the literature to account for LID among dichotomously scored items within a testlet is to treat it as a single super-item, score it polytomously, and apply polytomous item response models such as the partial credit model, the graded response models, or the generalized partial credit model. This method may lead to loss of information due to the sum of correct responses to the dichotomous items within a testlet thus reducing measurement precision. If items are scored polytomously, it is not practical to sum up the polytomous scores to get a giant polytomous super-item with even more item response categories.

Testlet Response Theory Models

Testlet effects can be conceptualized from multiple perspectives, an interaction between a testlet and persons, multidimensionality, or contextual effects of item groups on items nested within a testlet. In accordance with these conceptualizations, different testlet response theory models have been proposed.

Bayesian Random-Effects Testlet Model

Eric Bradlow, Howard Wainer, and Xiaohui Wang proposed a two-parameter Bayesian random-effects testlet model by incorporating a random-effect parameter into the unidimensional two-parameter item response model, indicating the interaction between a person and a testlet. Extensions of this model have been made to a three-parameter IRT model as well as to the graded response model.

Rasch Testlet Model

In another attempt to model testlet effects, Wen-Chung Wang and Mark Wilson proposed the Rasch testlet model as a special case of the multidimensional random coefficients multinomial logit model by including one more dimension or latent trait for each testlet. Essentially, each testlet introduces one additional dimension to an item. For each item, two latent traits are underlying the item performance for a specific examinee. These two latent traits are the general latent trait the test is intended to measure and the testlet-specific latent trait. For the items within the same testlet, the testlet-specific latent trait remains the same across these items. For different testlets, the testlet-specific latent traits will be different. If there are six testlets on a test, overall the test has seven dimensions, but for each item, it assesses two dimensions.

but for each item, it assesses two dimensions.

Three-Level One-Parameter Testlet Model

Hong Jiao, Shudong Wang, and Akihito Kamata developed a three-level one-parameter testlet model from the hierarchical generalized linear modeling framework for item analysis. This modeling framework conceptualizes the testlet effects as the item clustering effects where a testlet or an item cluster/group will exert contextual effects on items within the same item cluster or testlet. More specifically, item effects are modeled at Level 1, testlet or item group effects modeled at Level 2, and person effects modeled at Level 3. When the three-level model is combined into one model, it is equivalent to the Rasch testlet model.

Generalized Testlet Model

More generalized testlet models have been proposed to allow the discrimination parameters to be different for the general ability and testlet-specific ability. This generalization is essentially the application of the bifactor multidimensional IRT model to testlet-based tests. This more generalized testlet model allows more flexibility and increases model fit. Essentially, the two-parameter testlet response theory model is a second-order factor analysis model or a restricted bifactor model. Frank Rijmen elaborated the formal relations among the bifactor, the testlet, and a second order multidimensional IRT model and provided an empirical comparison.

Cross-Classified Random Effects Modeling

In real application settings, researchers have frequently observed other complex structures where more than one source of dependence is involved, and the relationship among multiple dependence sources is not purely hierarchical. One such example would be that items from the same scenario do not necessarily measure the same content, whereas items measuring the same content area are not necessarily nested within the same scenario. In this case, items are cross-classified by content areas and scenarios. Such LID from two sources is referred to as dual LID (DLID).

To model data with cross-classified structure, cross-classified random effects modeling has been developed to account for variance in test scores contributed by multiple nonnested clustering factors. When the cross-classified structure is fitted with hierarchical structure, the standard error estimates associated with the incorrectly modeled clustering variable will be underestimated. Recently, cross-classified IRT models have been proposed to account for DLID. Specifically, cross-classified testlet model for DLID can be expressed using multilevel modeling parameterization.

Following the multilevel modeling framework, Level 1 models item effects. Level 2 models item clustering effects cross-classified by two clustering factors. The two sources of clustering factors, content and scenario, will lead to the addition of content-specific and scenario-specific abilities, respectively. Level 3 models person-specific effects, which have to do with a person's ability. Combining the three levels leads to a three-dimensional testlet model for each item. Essentially, the probability of a correct response for a person to an item in a content area and a scenario is determined by the person-specific general ability, the content-specific ability for this person, and the scenario-specific ability for this person and item parameter(s).

Overall, the number of dimensions of the whole test is the sum of the number of content areas, the number of scenarios, and the number for the general ability. The content-specific ability is random effects, which is the same for the items assessing the same content area. The same is true with the scenario-specific ability. Although the whole test has many other dimensions, each item is only three-dimensional, which simplifies the computation and estimation of model parameters.

Multidimensional, Multilevel, Multigroup, and Mixture Testlet Models

Other extensions of the testlet models include the multidimensional, multilevel, multigroup, and mixture testlet models. More specifically, Li Cai proposed a two-tier item factor model with two general dimensions and one secondary dimension that accounts for testlet effects. Hong Jiao, Akihito Kamata, Shudong Wang, and Ying Jin explored the multilevel extension of the Rasch testlet model to simultaneously account for dual local dependence, namely LID and local person dependence, due to item and person clustering, respectively, in testlet-based assessments with individual persons nested with clusters such as classes, schools, and countries.

Following work on the multigroup IRT framework, “group-level” IRT models could be fitted for different populations to release the measurement invariance assumption across groups. Such multigroup IRT modeling offers a unified approach to problems such as differential item functioning detection, item parameter drift, test linking and equating among nonequivalent groups, and vertical equating. For example, Minjeong Jeon, Rijmen, and Sophia Rabe-Hesketh proposed a generalization of the multigroup bifactor model that extends the classical bifactor model by relaxing the typical assumption of independence of the specific dimensions. In addition to group-specific item parameters, means, and variances of all dimensions, the correlations among specific dimensions are also allowed to be different across groups.

In the multigroup testlet models, group membership for each person is a manifest variable that is known prior to data analysis. When persons cluster due to latent variables, their membership will be latent and needs to be estimated. The mixture testlet models extend the multigroup testlet models to account for latent person clustering. Thus, the modeling approach allows for latent differential item functioning due to the latent group differences in testlet-based assessments. The Rasch mixture testlet models for both dichotomous and polytomous item response data have been proposed in addition to the 3PL mixture testlet model.

Recent studies proposed more extensions of the testlet models. For instance, a multilevel cross-classified testlet model was explored to account for person clustering and dual item clustering in items cross-classified by two grouping variables. Also, a multigroup cross-classified Rasch testlet model was proposed for DLID in the presence of differential item functioning.

Estimation Methods

Estimation methods differ for these testlet models. The marginalized maximum likelihood estimation method with the expectation–maximization algorithm was used to estimate parameters for the Rasch testlet models using ConQuest. The sixth-order approximation Laplace (Laplace) method was explored for the three-level one-parameter testlet model using HLM6. Further, the Markov chain Monte Carlo method was demonstrated for the two-parameter, three-parameter, and the graded-response IRT models using the SCORIGHT. These three estimation methods were compared for the Rasch testlet model.

Essentially, the estimation method does not make a practical difference in the Rasch testlet model parameter estimation accuracy. Li Cai proposed the Metropolis–Hastings Robbins–Monro algorithm, which can be utilized to estimate the testlet and the more generalized bifactor model parameters. Further, other software programs could also be used in estimating testlet model parameters including WinBUGS, OpenBUGS, mdltm, IRTPRO, flexMIRT, SAS, and Mplus 7.

Hong Jiao and Dandan Liao

See also Item Response Theory; Reading Comprehension; Reliability

Further Readings

Jiao, H. (2014). Differential item functioning analysis for testlet-based assessments. (Research Report submitted to the Governing Board for the AERA Grants Program). College Park, MD: University of Maryland.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. Journal of Applied Measurement, 6, 311–321.

Liao, D., & Jiao, H. (2016, April). A multigroup cross-classified testlet model for dual local item dependence in the presence of DIF items. Paper presented at the 18th International Objective Measurement Workshop, Washington, DC.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). User’s guide for SCORIGHT (Version 3.0): A computer program for scoring tests built of testlets including a module for covariates analysis (Research Report 04–49). Princeton, NJ: Educational Testing Service.

Xie, C. (2014). Cross-classified modeling of dual local item dependence. Unpublished Doctoral Dissertation, University of Maryland, College Park.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187–213.

Paul B. Ingram Paul B. Ingram Ingram, Paul B.

Michael S. Ternes Michael S. Ternes Ternes, Michael S.

Test–Retest Reliability Test–Retest reliability

1688

1691

Test–Retest Reliability

Reliability estimates utilizing a test–retest approach measure the degree to which the same testing instrument produces similar results when administered to the same individual in as similar a manner as possible over a period of time. Test–retest reliability is a popular form of reliability estimation for the development and validation of test instruments and is based on correlation. Test–retest reliability falls behind only internal consistency estimates (e.g., coefficient α) in popularity for the evaluation of reliability. Test–retest reliability is a measure of test consistency and score fluctuation emphasizing the psychometric assessment of test form stability over a period of time. For instance, if an intelligence test is administered twice to the same individual within a short period of time, then a high test–retest reliability coefficient would be expected because of the general stability of the measured intellectual functioning and the standardized testing procedures. In contrast with other consistency estimates, such as those that examine internal consistency (e.g., coefficient α or split-half reliability), test–retest reliability is a measure of temporal stability. Because the instrument used during the calculation of test–retest reliability is the same during both administrations, this approach to reliability estimation assesses measurement error as the degree to which changes happen across administrations. If different but supposedly related testing instruments were administered in the same manner over a period of time, this administration would depend on an alternative form of reliability. If reliability is calculated using responses from a single test administration based on how similar responses are to one another, this administration would depend on the coefficient α. This entry discusses the theoretical approach and assumptions of as well as issues associated with test–retest reliability and then provides information on notation and interpretation.

Theoretical Approach and Underlying Assumptions

Test–retest is a measure of reliability as seen through the lens of classical test theory. In this approach to classical test theory, the closer obtained scores are to one another over two administrations, the higher the test–retest reliability coefficient. Higher reliability coefficients indicate a greater portion of true score measurement and lesser amount of error. Thus, higher reliability coefficients indicate more precise and stable measurement. For instance, if 80% of variability in test scores is attributable to systematic performance, then the instrument would have a .80 test–retest reliability coefficient, indicating 20% of variability being the result of error. Some examples of error that may occur causing variability between scores include variations in attention and concentration to the task at hand, learning as a result of test exposure, approaches to testing that are indicative of haphazard or of careless responding, and problems with item comprehension. Although it is impossible to remove all variability from measurement, the expectation is that well-designed tests for a stable trait will be able to obtain a consistently reliable measurement of the underlying true score.

Test–retest reliability relies on two underlying assumptions. Test–retest assumes that true scores of the measured characteristic for an individual do not change over time and that all variation in an observed score is due to either random or systematic error. Not all characteristics are ideal for this assumption because some are expected to change over time. Whereas major personality characteristics (e.g., the Big Five personality traits such as extraversion and agreeableness) are generally conceptually stable over the lifetime and thus appropriate for test–retest reliability measurement, other state-based attributes are not. For instance, depression is a mood state and would be expected to fluctuate over a course of time, therefore use of test–retest coefficients to demonstrate evidence of reliability would be less appropriate. In the interim between separate administrations of a depression test, individuals are likely to experience a change in their stress (e.g., receive parking tickets, have disagreements with loved ones, enjoy a rewarding day at work) and may even experience major life events. All of these would be expected to impact the amount of depression the person reports because the underlying level of depression experienced would have changed.

Test–retest reliability also assumes that a long enough period of time has passed between test administrations to ensure that there are no carryover effects that would impact how the respondent selects answers. For instance, people may remember their answers during the second test administration, biasing how they opt to respond. Another possibility, particularly salient with psychological measures, is that the experience of testing may alter the underlying state being assessed. Taking a depression inventory might, for instance, increase individuals’ experienced level of depression by priming them to think about their mood and recent behaviors. In both of these instances, it is important for test–retest reliability that a sufficient time period between the administrations exist in order to reduce error and establish an accurate reliability coefficient. Because test–retest reliability is affected not only by error associated with internal consistency but also by sample specific error (e.g., people remembering their answers or giving different responses as a function of test-influenced mood state), test–retest is prone to higher rates of error than other reliability estimates. At least three response categories are recommended for each item in order to maximize the opportunity a test has to achieve high test–retest reliability. However, it is important to note that a lower number of test item response categories does not necessarily result in lower reliability.

Issues in the Use of Test–Retest Reliability

Using test–retest reliability offers some advantage over other methods of assessing reliability evidence. Test–retest maintains a stable structural form for the instrument because it utilizes the same instrument. This use of the same instrument means that items are held as a constant and provide a stable evaluative context for their intended construct without risk of construct drift due to item changes. This consistency of testing form assuages the concern over systematic error due to measurement variability by ensuring that any error present is there in equal amounts during both test administrations. Test–retest also requires only a single instrument and does not require development or identification of an appropriate alternative form for comparison; therefore, test–retest is easier and more readily accessible to conduct during test development.

However, test–retest reliability also has some disadvantages. As the interval of time between test administrations is selected, carryover effects, which can inflate the reliability estimates, must be considered. If a period of time is too short, then responses may simply reflect the participants remembering their previous answers and not the actual stability of the measured construct. Similarly, because test–retest reliability relies on the administration of the same items, it is possible that practice effects occur during which test scores systematically increase. A student who is administered a math test twice, for instance, may learn the content of the test or develop an improved approach to the test's tasks as a response to earlier test exposure. Both of these errors could possibly occur together, making it difficult to parse the causal reason behind observed changes in trait stability. If a psychologist administers a math skills test to a group of children to measure test–retest's reliability, some students possibly will do better the second time merely as a result of chance, while some will do worse because of differing levels of influence from practice and carryover effects. Other effects (e.g., regression to the mean or number of item anchors) may also influence test–retest reliability.

Increasing the time between test administrations is an option for balancing carryover effects. However, this approach can create other problems. For instance, although a longer period of time between tests decreases the likelihood that participants will remember the responses they provided to items during the first administration, it also increases the chance for change to occur on the measured characteristic. This risk for change is particularly pronounced for state characteristics. Consider anxiety as an example. Over a brief period of time, one would expect that the anxiety level would remain relatively stable. If tests are administered several weeks apart, however, anxiety would be expected to shift as a result of situational changes. A period of 1 or 2 weeks is frequently recommended for test–retest reliability calculation to balance these contrasting concerns.

Some researchers administer interventions between test administrations. These interventions are aimed at increasing the measured trait to calculate a change or gain score. Gain scores measure the ability of a given characteristic to be influenced by outside factors (e.g., random error), and so it is especially important to consider measurement error. One common approach is to evaluate the standard error of measurement using the test–retest coefficient of the two tests. This approach of utilizing the test–retest coefficient instead of the internal reliability coefficient for a single administration allows for the inclusion of measurement of error for between test stability and does not simply include standard error of measurement during a single test's administration. One example of a gain approach is the reliable change index.

Notation and Interpretation

Test–retest reliability provides a coefficient ranging from 0 to 1.00 with higher scores being increasingly stable and more preferred. Scores obtained using a scores being increasingly stable and more preferred. Scores obtained using a test–retest measurement approach are frequently represented using the notation T1 for the first time the test is given and T2 as the second administration of the test. For instance, the reliable change index is calculated using the following formula:

$$\mathrm{RCI} = (\mathrm{T2} - \mathrm{T1}) \Big/ \left(\mathrm{SEM}\sqrt{2}\right).$$

One difficulty with using test–retest reliability stems from the fact that there are few, if any, empirical standards to judge how a reliability estimate should be interpreted. The most frequently used cut value for acceptable test–retest reliability is .70. While the range of .70 to .90 is frequently described as acceptable for demonstrating test–retest reliability, the expected performance of a given test will depend upon a myriad of factors including the type of characteristic being measured (e.g., state vs. trait), the length of time between administrations of the test, and the type of population being sampled. Although there are numerous meta-analyses evaluating the frequency and distribution of test–retest estimates, these reviews typically focus on a single construct. This focus on single measurement constructs makes it difficult to establish standard interpretive guidelines but emphasizes the importance of considering test–retest reliability interpretation within the established evaluative context for which it is used. For instance, in a meta-analytic review of neuropsychology assessment instruments, some assessed areas showed evidence of reliability that is considered adequate (>.70), whereas others are not. Variations in reliability below .70 were commonly attributed to the common problems with test–retest reliability discussed earlier, such as practice effects and instrument design factors (e.g., ceiling and floor effects within the population).

Paul B. Ingram and Michael S. Ternes

See also Classical Test Theory; Coefficient Alpha; Correlation; Reliability; Split-Half Reliability

Further Readings

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency and use of various types. Educational and Psychological Measurement, 60(4), 523–531.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Toward a standard definition of clinically significant change. Behavior Therapy, 17, 308–311.

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educational and Psychological Measurement, 64(6), 956–972.

Stephanie Schmitz Stephanie Schmitz Schmitz, Stephanie

Tests

Tests

1691

1692

Tests

A test can be defined as an instrument, tool, or procedure that is used to obtain information about a particular outcome. Other definitions include reference to its capacity to measure specific variables and including standardized procedures to gather data on underlying constructs for drawing conclusions or developing hypotheses for further examination. Currently, there are thousands of available tests in many different areas, such as achievement, intelligence, personality, aptitude, and vocational. This entry discusses the various types of test and scores obtained and the psychometric properties of tests.

Types of Tests and Scores

Tests can be considered norm referenced or criterion referenced. For norm-referenced tests, an individual's scores are compared to scores from a particular normative group, which is a sample of individuals who should be representative of the individuals for whom the test was developed. When interpreting criterion-referenced tests, however, the emphasis is on determining what the individual knows, which can be done by comparing the examinee's performance to a particular standard, or level of performance.

Multiple types of scores can be obtained from tests. Raw scores, such as the number of correct and incorrect items produced by an examinee, can be obtained on any type of test. Although raw scores are, by themselves, relatively meaningless, they can be transformed into more refined scores that carry different types and levels of meaning. On norm-referenced tests, raw scores are often transformed into standard scores, which are scores that are standardized according to a certain metric and are compared to a group’s mean score in reference to their standard deviation. Percentile ranks are often used to describe a student’s performance on a norm-referenced test, as they provide information about an individual’s position within a distribution or set of scores from a particular group. Percentile ranks indicate what percentage of individuals within a particular group received scores that fell at or below the examinee’s scores. Percentile ranks are somewhat limited, however, as they have unequal scale units and the differences between these units will vary, affecting their meaning.

Due to the inherent presence of measurement error, norm-referenced scores are often presented within confidence intervals. Confidence intervals can be of various sizes, with their width indicating how much one should expect an individual’s actual test score to vary from the true score. The larger the confidence interval, the more confident one can be that the true score falls within that range. Although the confidence interval can be calculated by multiplying the standard error of measurement by *Z* scores, they are often presented within test manuals, such as at the 68%, 90%, 95%, and sometimes 99% level.

Psychometric Properties

Psychometric properties are critical components to consider when selecting a test. Reliability represents the consistency of measurement across administrations, or replications, of a particular test. Typically, reliability is used when discussing an individual’s obtained score, as the true score is directly influenced by measurement error. Different types of methods evaluate reliability and produce reliability coefficients. One such type is test–retest reliability, which is determined using the Pearson product–moment correlation technique, showing how consistent scores are across administrations. A second type of reliability is an alternate form of reliability, in which scores from one form of a test are compared to those from an alternate form of the same test, each containing the same content. A third type of reliability coefficient is internal consistency, which refers to the level of consistency across test items. Internal consistency typically provides the highest reliability coefficients, which reflects the level of consistency across items. Internal consistency is reported by the Cronbach’s coefficient α, representing the relationship between items across the entire test. Finally, interrater reliability can also be obtained, which refers to the consistency across ratings. This particular type of reliability addresses measurement error associated with an examiner’s subjective evaluations of an examinee’s test responses. Interrater reliability can be represented through either interrater agreement, in which the percentage of agreement is calculated, or through an interrater reliability coefficient, often resulting from Pearson product–moment correlations.

Reliability is a prerequisite for validity. In other words, scores from a particular test cannot be considered valid if they are not first found to be reliable. Validity has often been defined as whether a test sufficiently measures the underlying constructs on which it was developed, providing meaning for test scores. However, more recent definitions indicate validity to refer to whether the constructs underlying the test are represented in a meaningful way through the scores obtained. John Kranzler and Randy Floyd in their 2013 work explained that evidence of validity has historically been reported through the "classic tripartite method," consisting of content, criterion-related, and construct validity, each considered to be a distinct type of validity (p. 74). However, recent conceptualizations suggest validity to be a "unitary concept," with all types providing information on construct validity related to obtained scores and corresponding inferences and interpretations (p. 74).

Evidence of validity has been divided into five strands, including evidence on (1) content validity, referring to whether the content of the test accurately represents the underlying constructs of the test; (2) response processes, confirming examinee testing behaviors to give information on cognitive operations and processes being used to provide responses; (3) internal structure, referring to whether the correlation between test items and other test variables was as expected; (4) external relations, examining various relationship patterns between scores on the test and how they correlate with other tests said to measure the same underlying constructs; and (5) consequences, referring to a determination of whether there were "unintended consequences" for a particular examinee and/or group taking the test. It is important to recognize threats to validity as well as indications of test bias toward a particular group.

Stephanie Schmitz

See also Reliability; Standardized Tests; Test Bias, Testing, History of; Validity

Further Readings

Cohen, R. J., & Swerdlik, M. E. (2005). Psychological testing and assessment:

An introduction to tests and measurement (6th ed.). Boston, MA: McGraw-Hill.

Kranzler, J. H., & Floyd, R. G. (2013). Assessing intelligence in children and adolescents: A practical guide. New York, NY: Guilford Press.

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). Measurement and assessment in education. Boston, MA: Pearson.

Salkind, N. J. (2006). Tests & measurement for people who (think they) hate tests and measurement. Thousand Oaks, CA: Sage.

Wright, R. J. (2008). Educational assessment: Tests and measurements in the age of accountability. Los Angeles, CA: Sage.

Mary M. Chittooran Mary M. Chittooran Chittooran, Mary M.

Testwiseness

Testwiseness

1693

1695

Testwiseness

Testwiseness, also referred to in the literature as testwiseness, was first conceptualized in the early 1950s by Robert Thorndike, who described it as a factor that differed across individuals, had an influence on test performance, and contributed significantly to test score variance among individuals. The earliest and most well-known definition of testwiseness comes from the work of Jason Millman, Carol Bishop, and Robert Ebel who in 1965 defined testwiseness as

> a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score. Testwiseness is logically independent of the examinee's knowledge of the subject matter for which items are supposedly measures … . it will be restricted to the actual taking of (not preparing for) objective achievement and aptitude tests. (p. 707)

Testwiseness is a particularly salient concept in the prevailing climate whereby results of high-stakes testing determine educational outcomes, and there is a call for increased accountability among both students and teachers. It becomes especially important to examine all potential contributors to students' test scores. This entry describes the nature and characteristics of testwiseness, discusses the trainability of testwiseness, examines some of the controversies surrounding this topic, and finally, offers some testwiseness strategies reported to enhance test performance.

Nature and Characteristics

Nature and Characteristics

There are three common schools of thought regarding testwiseness; one perspective views it simply as a source of variance in test scores, a second argues that it is a persistent trait of the examinee, and yet a third suggests a synthesis of these two approaches. Testwiseness is thought to exist independently of general cognitive ability, but it has also been linked with specific cognitive skills as well as with general verbal ability. Since Millman and his associates first defined testwiseness, other researchers have confirmed that objective or multiple-choice tests—whether commercially developed or teacher made—are most susceptible to the influence of testwiseness. This may be particularly problematic among test users with large groups of examinees for whom they are responsible or those who have to assess mastery of a great deal of information (which lends itself better to objective tests rather than short answer or essay tests).

Trainability

Despite ongoing debate about its composition, testwiseness is reported to be trainable in individuals from preschool through adulthood in a variety of settings and disciplines. Certainly, not all examinees show improvement in test performance following training in testwiseness strategies; such improvement, if it occurs, is thought to be dependent on various factors such as examinee’s attitude, length and intensity of training, prior knowledge, familiarity with application of testwiseness skills, and the use of metacognition about one’s test performance. What is clear is that a testwise individual tends to perform better on items that are “testwise susceptible” than another examinee, who may be equally knowledgeable about test content, but who lacks testwiseness. Individuals who are lower in testwiseness may, therefore, be at a considerable disadvantage in a testing situation.

Controversies

A major controversy that surrounds testwiseness relates to whether it should be taught at all. If a test score is a snapshot of a student’s functioning at a particular moment in time, then teaching testwiseness strategies that are based on something other than ability and knowledge of test content might lead to inaccurate determinations about that student’s actual learning. It is also argued that training in testwiseness benefits already privileged students whose families can afford the substantial costs charged by test preparation companies but does not similarly privilege students with low income, who might be equally deserving, but for whose families those costs might be prohibitive. Testwiseness also compromises test validity and predictive utility because the test does not measure what it purports to measure but instead assesses the test taker’s ability to “work” the test. If testwiseness influences test validity, then it also follows that it implicates test reliability, that is, consistency in measurement.

Test developers who have become sensitized to the potential impact of testwiseness on test scores have attempted to circumvent this influence and to minimize the resultant variance across test scores by identifying common sources of testwiseness, modifying test items, clarifying directions, and cautioning consumers about this phenomenon. Despite these efforts, however, the problem continues unabated.

Strategies

Various classifications for testwiseness strategies have been proposed over the years, including taxonomies that describe strategies for effective time management, error avoidance, and guessing, as well as deductive reasoning and use of cues. Other taxonomies have described categories that address strategies used before, during, and after a test. This section proposes a five-category taxonomy of testwiseness strategies: (1) understanding objective tests, (2) organization and time management, (3) utilizing the format and characteristics of tests, (4) making use of test cues, and (5) benefiting from flaws in test construction. It also offers a sampling of strategies that are culled from various sources thought to increase testwiseness. The following section describes these five categories in more detail.

Understanding Objective Tests

Objective tests of the multiple-choice variety differ from other types of tests in that they test recognition, not recall of information, attempt to sample the greatest number of behaviors in the shortest space of time, require both speed and accuracy, and may evaluate more than one type of cognitive functioning, such as comprehension and application. Each multiple-choice item includes a stem and a variety of options or alternatives, at least some of which are distractors (i.e., options that lead the unsuspecting examinee away from the correct answer). Testwise individuals have learned to approach multiple-choice tests differently than they would essay or short-answer tests.

Organization and Time Management

Testwise individuals are organized; plan their time before, during, and after the test; monitor their progress during the test; and review the completed test once they are done. They pay close attention to directions, read and answer each question carefully, skip (but then return to) difficult questions, so as to minimize anxiety and frustration, and quickly eliminate obviously incorrect options. On the other hand, ineffective test takers often use what Abdullah Al Fraidan referred to as “test-unwise” strategies such as impulsively changing answers, managing time poorly, and not reading questions and directions carefully.

Utilizing the Format and Characteristics of Tests

Familiarity with test construction principles allows testwise individuals to use that knowledge to enhance test performance. For example, test developers usually place easy test items first, so as to build confidence in test takers; therefore, testwise individuals get through these quickly, reserving time and energy for the more difficult items that are sure to follow. If there is no penalty for guessing, testwise examinees do not leave an answer blank, but instead, try to reason out an answer, even on an item with which they are not familiar. Instead of using blind guessing techniques, which often lead to incorrect answers, testwise examinees know that educated guessing, whereby one can easily eliminate half of the available options, can maximize the chances of finding the right answer. Ineffective test takers may impulsively change a correct answer to an incorrect one, often at the last minute; however, effective test takers only change answers if new information on subsequent items suggests that a change is in order.

Making Use of Test Cues

Many tests contain cues, both within and across items, that lead the testwise examinee toward the correct answer. Items containing absolutes and specific determiners, such as always, never, and should, or phrases such as “All of the above" and "None of the above" are as likely to be incorrect as are absurd, funny, or clearly incorrect options, whereas grammatical agreement on stem and options often leads to the right answer. Knowledge of common prefixes and suffixes (e.g., hyper-/hypo-) can help the examinee identify the right answer, even when test content is unfamiliar. Testwise examinees know that attractive distractors (i.e., incorrect answers) are often placed first because it is the first thing test takers see and remember (the so-called primacy effect), or last, because it is the last thing they see and remember (the recency effect). Also worthy of attention are options that seem different until they are reworded, that differ on only one dimension (pupils constricting or dilating), are mutually exclusive (life and death), or are emotionally charged (terrible or awful). Common errors that ineffective test takers make have to do with misreading words (e.g., mm for cm), overlooking key words such as "occasionally" and "usually," and not paying attention to wording such as "Which of the following is *not* true?" or "Pick the *best* answer," in a situation where all the options are true but only one is the best.

Benefiting From Flaws in Test Construction

Occasionally, in the case of commercial assessments and more frequently on teacher-made tests, it is possible to use flaws in test construction to benefit the examinee. For example, studies examining both commercial and teacher-made objective tests have found a significant and disproportionate number of errors that could conceivably be used by a sophisticated test taker to improve test performance. Some common flaws on teacher-made tests, for instance, include language that does not match the instructor's teaching style, spelling errors, and lack of grammatical agreement between stem and alternatives.

Future Research

Testwiseness, which is a source of examinee variation on test performance and is not attributable to random error, has been discussed in the literature over the past several decades. Because it is not yet fully understood, it is important that further research be conducted to clarify its components and understand how it works to influence test performance.

Mary M. Chittooran

See also Classroom Assessment; Metacognition; Multiple-Choice Items; Standardized Tests; Tests; Validity

Further Readings

Al Fraidan, A. (2014). Test-unwiseness strategies: What are they? Journal of Applied Sciences, 14(8), 828–832.

Amer, A. A. (2007). ESL/EFL testwiseness and test-taking strategies. Retrieved from ERIC Document Reproduction Service (ED497399). http://www.eric.ed.gov/PDFS/ED497339.pdf

Benson, J. (1988). The psychometric and cognitive aspects of testwiseness: A review of the literature. In M. Kean (Ed.), Testwiseness. Bloomington, IN: Phi Delta Kappan.

Bicak, B. (2013). Scale of test taking and test preparation strategies. Educational sciences: Theory and practice, 13(1), 279–289.

Chittooran, M. M., & Miles, D. D. (2001). Test-taking skills for multiple-choice formats: Implications for school psychologists. Resources in Education. (ED455488).

Frey, B. B., Peterson. S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. Teaching and Teacher Education, 21(4), 357–364.

Houston, S. E. (2005). Testwiseness training: An investigation of the impact of testwiseness in an employment setting (Dissertation). University of Akron, OH.

Millman, J., Bishop, C., & Ebel, R. (1965). An analysis of test wiseness. Educational and Psychological Measurement, 25(1), 707–726.

Rogers, W. T., & Yang. P. (1996). Testwiseness: Its nature and application. European Journal of Psychological Assessment, 12(3), 325–335.

Xu, Y., & Wu, Z. (2011). A review of the research on test-taking strategies in the past 50 years. Foreign Language Learning Theory and Practice, 1(1), 42–51.

Magdalena Bielenia-Grajewska Magdalena Bielenia-Grajewska Bielenia-Grajewska, Magdalena

Threats to Research Validity Threats to research validity

1696

1698

Threats to Research Validity

Research is an indispensable part of modern reality, facilitating the progress and development of societies, economies, and individuals. Thus, it is crucial to ensure the quality of research at all its stages. The most critical aspect of quality is research validity, or whether the results of studies are interpreted and understood correctly. Threats to validity are characteristics of research designs that lessen the degree to which results are interpreted correctly. This entry first discusses conceptual frameworks of research validity and then looks at specific types of validity threats and ways of avoiding validity threats.

Conceptual Frameworks of Research Validity

In quantitative research, the term *validity* is often used to discuss measurement validity, which refers to the degree to which an instrument or test measures what is supposed to be measured. In qualitative research, validity is concerned with whether findings are representative of participants' experiences. Although they are not precisely the same concepts, researchers also use the terms *trustworthiness*, *quality*, and *rigor* when discussing research validity. Isadore Newman and Carolyn R. Benz suggest the term *legitimation* to denote a broader understanding of truth value.

Three key concepts related to validity are transferability, dependability, and confirmability. *Transferability* stresses that validity is anchored in a given context. Thus, transferability shows the way studies can be applied or transferred to other environments. *Dependability* refers to the stability of data and the degree to which data are collected in a way that is precise and reliable.

Dependability can be achieved by triangulation and sequencing of methods. The third concept is *confirmability*, which highlights the qualitative side of objectivity. Thus, confirmability is aimed at recognizing and investigating systematic biases that take place in research, which are threats to research validity that can be either intended or unintended.

Donald T. Campbell and Julian C. Stanley, who introduced the concept of threats to validity in a 1963 book, discuss three types of research validity: internal validity, external validity, and construct validity. Internal validity is understood as the causal relationship between one or more independent variables and one or more dependent variables. External validity involves how the results can be used in other contexts. Construct validity refers to the degree to which a construct under investigation is accurately measured and interpreted.

Types of Threats

Some threats to validity take into account the role of researchers engaged in studies. *Researcher-dependent variability* encompasses the factors influencing trustworthiness that depend on scientists. These factors include, among others, the researchers' diligence in creating and conducting experiments and the procedures they use in sampling. Researcher-independent variability includes factors that do not depend on the performance of researchers. Threats to validity that are independent of the performance of researchers include, among others, technological failures or fake answers provided by respondents.

Language can also be analyzed as a potential threat to research validity. First, language barriers may determine the understanding of questions. When studies are conducted in a language that is not the participants' native language, researchers should pay attention to participants' level of fluency with the language. Even when participants are fluent in the language of the study, if it is not their native language, their perception of phenomena and speed of selecting answers may be different than if the same questions were asked in their mother tongue. Another crucial factor is technology; validity may be influenced by access to technology. For example, Internet surveys may exclude people who have no Internet access or do not have a high-speed connection. In addition, the speed of individuals' Internet connection may influence their choices when taking a survey.

Validity in studies may also be examined from the perspective of stage. For example, sampling may be conducted in the wrong way and this will influence results. Distribution and collection of questionnaires may also influence the validity of an experiment. The final stage of research, such as presenting results at conferences or publishing papers, is also prone to potential failures. An example may include presenting only one side of the phenomenon in a short paper or showing only a piece of an experiment during a speech at a congress due to limited presentation time. Moreover, research validity may also be viewed through the perspective of models.

Threats to validity can be studied through the perspective of a given type of validity. Donald H. McBurney and Theresa L. White list different threats to internal validity, external validity, and construct validity. Threats to internal validity are ambiguous temporal precedence, events outside the laboratory (history), maturation, effects of testing, regression effect, selection, and mortality.

Ambiguous temporal precedence is connected with difficulty in determining which variable is the cause and which is the effect. Events outside the laboratory (history) encompass situations that take place outside experiments but may influence them. Maturation involves errors occurring due to the flow of time between experiments. This notion is especially important in an experiment involving children who grow up during the course of the experiment. Maturation and history are often confused although historical effects are external, whereas maturation effects are internal.

Effects of testing involve changes in the participants that occur as a result of being tested. The regression effect refers to the likelihood that individuals who achieve high scores during the first testing will be closer to the mean during the second testing. Selection involves comparing different groups on some dependent variable where the two groups are not comparable. Mortality, also called selective subject loss, involves participants dropping out of a study or becoming unavailable to participate, so that the remaining sample is no longer representative of the group being studied.

Two threats to construct validity are a loose connection between theory and method, when the correspondence between theoretical concepts and the way to check them is vague, and ambiguous effect of independent variables, when the perception of a situation is not the same among participants and the experimenter. Threats to external validity include problems arising from experimenter. Threats to external validity include problems arising from generalizing findings of a study to other subjects, other times, and other settings.

Research Validity Tools

Sharon M. Ravitch and Nicole Mittenfelner Carl have proposed reflexive validity questions to determine research validity that include questions about credibility, transferability, dependability, confirmability, descriptive validity, interpretive validity, theoretical validity, generalizability and evaluative validity. Questions to consider to determine the credibility of research include those about the correlation between methods and research questions, selection criteria and sampling strategies, richness of data, challenging biases and assumptions, and the roles of research participants in research and its interpretation. Transferability involves questions on contextual data and contextual relevance as well as questions about the description of settings and participants.

Dependability focuses on the relation between methods and research questions as well as challenges concerning study design, data collection processes, and analyses. Confirmability includes questions about challenging one's thinking and similarities with other studies. Descriptive validity deals with the accuracy of data reported by the researchers. Interpretive validity involves the researcher accurately interpreting participants' thoughts and beliefs as they relate to the research. Theoretical validity involves the degree to which the theoretical explanation developed from a study fits the data. Generalizability involves the degree to which results of a research study can be applied to populations and settings other than those represented by the study's sample. Evaluative validity deals with judgments and their influence on research.

Ravitch and Carl discuss the application of strategies and processes that aim at reaching validity, such as triangulation, participant validation, strategic sequencing of methods, thick description, dialogic engagement, multiple coding, structured reflexivity practices, and mixed methods research. Triangulation can be described as using different processes to ensure the validity of research, with subcategories such as methodological triangulation, data triangulation, investigator triangulation, theoretical triangulation, and perspectival triangulation. Participant validation strategies (member checks) are used to observe how participants perceive the study; they may focus on technical or relational aspects of study, offering research feedback.

Strategic sequencing of methods involves organizing complex studies in an

Strategic sequencing of methods involves organizing complex studies in an efficient way and can be subcategorized into within-methods sequencing, which involves how questions are grouped and ordered, and between-methods sequencing, which involves how questions may be informed by other methods such as observations and focus groups. Another strategy is thick description, aimed at a detailed description of the participants and research context.

Dialogic engagement (also called peer debriefers, critical friends, and critical inquiry groups) involves sharing studies with other colleagues or stakeholders. Multiple coding (or interrater reliability) is used to check how other researchers are coding data and observe the overlap of interpretations. Structured reflexivity practices include memos, dialogic engagement practices, research journals, and mapping strategies, aimed at producing complex studies. Mixed methods research combines qualitative and quantitative methods.

Magdalena Bielenia-Grajewska

See also Cognitive Neuroscience; Concurrent Validity; Content Validity Ratio; External Validity; Internal Validity; Predictive Validity; Triangulation; Validity; Validity Coefficients; Validity Generalization; Validity, History of

Further Readings

Levine, T. R. (2011). Quantitative social science methods of inquiry. In M. L. Knapp & J. A. Daly (Eds.), The SAGE handbook of interpersonal communication (pp. 25–57). Thousand Oaks, CA: Sage.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. Harvard Educational Review, 62(3), 279–301.

McBurney, D. H., & White, T. L. (2013). Research methods. Belmont, CA: Wadsworth Cengage Learning.

Newman, I., & Benz, C. R. (1998). Qualitative-quantitative research methodology: Exploring the interactive continuum. Carbondale: Southern Illinois University Press.

Ravitch, S. M., & Mittenfelner, C. N. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: Sage.

Matthew S. Johnson Matthew S. Johnson Johnson, Matthew S.

Thurstone Scaling

Thurstone scaling

1698

1701

Thurstone Scaling

One of the earliest data collection methods for the measurement of attitudes is Thurstone's law of comparative judgment (LCJ). The LCJ, which has its roots in psychophysical scaling, first develops a large number of attitude statements or stimuli and then uses information from judges to place the stimuli along a unidimensional continuum. There are three methods for collecting data from the judges: the method of equal-appearing intervals, the method of paired comparisons, and the method of successive intervals. After the scale positions of the stimuli have been determined, the stimuli are presented to research subjects or respondents whose attitudes are then measured. The average position of the items endorsed by a particular respondent serves as the estimate of the respondent's latent attitude. After providing background information, this entry further explores Thurstone's LCJ and discusses alternative methods.

Background

A large part of social psychology focuses on the study of attitudes. Attitude research covers a wide variety of topics, such as racism, sexism, attitudes about school, religion, politics, among others. Unlike weight, height, or age, attitudes are not easily quantified, complicating the social research questions at hand. Moreover, there is a great deal of confusion and ambiguity attached to the concept of attitude. In fact, few social psychologists even agree on the definition of *attitude*. In his early work, in 1928, Louis Thurstone defines attitude as "the sum total of a man's inclination and feelings, prejudices and bias, preconceived notions, ideas, fears, threats, and convictions about any specified topic" (p. 531).

Attitudes are simply one of many hypothetical constructs used in the social sciences. They are unobservable and therefore cannot be measured directly. If attitudes could be observed directly, then studying the relationships between attitudes and other variables of interest would be straightforward. Although attitudes cannot be observed directly, hundreds of methods for quantifying attitudes have been suggested in the psychological measurement literature.

One of the earliest methods for the measurement of attitudes was suggested by Thurstone in his 1928 paper, “Attitudes Can Be Measured.” As in Thurstone’s originally proposed methods, most attitude measurement methods assume that the range of attitudes falls on a single bipolar continuum or scale. This continuum can represent attitudes such as the liberal, conservative political scale, levels of prejudice, or how important a student thinks studying is, and countless others.

Assuming the attitude of interest lies on a unidimensional scale, the goal of attitude scaling is to estimate the position of individuals on the continuum. Scaling starts by constructing or selecting a set of items or stimuli that represent varying levels of the attitude. Items can be physical objects, but are usually a set of opinion statements, the verbal expressions of attitude. “All college students should be required to take a statistics course” is an example of an opinion statement that might be used to study attitudes about the importance of studying statistics. Once a set of items that the researcher believes and measures the attitude, or attribute of interest, is constructed, the locations of the stimuli and individuals’ attitudes along the unidimensional continuum are estimated. Thurstone’s LCJ is one of the earliest methods to accomplish this goal.

Thurstone’s LCJ

Thurstone’s LCJ built off of previous research in psychophysical scaling, whereby subjects are presented with physical stimuli and asked to rank them along some scale, such as size, loudness, or tone. Although psychophysical scaling was concerned with scaling stimuli that could be physically measured, Thurstone was interested in developing methods to scale unobservable constructs such as attitudes.

The LCJ proposes scaling attitudes by examining opinions expressing different attitudes toward a topic. The method begins by developing a large number of these opinion statements, or stimuli; the stimuli should express varying levels of attitudes toward the topic of interest. For example, to study attitudes about the teaching of statistics, a researcher might use the following opinion statements about the field of statistics:

1. Statistics is the most important subject to study in school.
2. Everybody should be required to take at least one statistics course.
3. Statistics should be an elective course and should not be required.
4. Statistics is not useful for most professions.
5. Statistics concepts should be taught all throughout school starting in kindergarten.
6. Knowledge of statistical methods is not very important for most people.

There is a wide range of opinions. For example, Statements 1, 2, and 5 express positive sentiments about the teaching of statistics. In contrast, Statements 3, 4, and 6 express indifferent to negative sentiments about statistics education.

Once a large set of items is constructed, the items are presented to a group of individuals acting as judges in order to scale the items. After item locations are estimated in this first stage of data collection, the respondent locations are determined in a second, separate data collection stage. The two stages for Thurstone scaling are summarized below.

Scaling the Items

Thurstone suggested three methods for collecting judgment data from the judges about the relative locations of the items along the scale: the method of paired comparison, the method of equal-appearing intervals, and the method of successive intervals.

The Method of Paired Comparisons

Suppose that there are j stimuli that the researcher would like to scale on the unidimensional continuum. The method of paired comparisons pairs each stimulus with all of the other stimuli to create a total of jj-12 stimuli pairs. The judges are then asked to examine each pair of stimuli and select one of the two as the one located higher on the scale. For example, consider the 6 items on the importance of studying statistics described earlier. A judge would examine the pair of items formed from the first two statements and then decide whether

Statement 1 or Statement 2 represented a more positive opinion about studying statistics. In this case, one might expect that most judges would consider Statement 1 to reflect a more positive view about statistics than Statement 2.

In order to find the location of each stimulus, the Thurston method assumes that when a judge is presented with a pair of stimuli, each stimulus produces a psychological sensation in the judge. If the psychological sensation produced by Item j is larger than the sensation produced by Item k for a particular judge, that judge would select Item j as being located above Item k. Thurstone’s Case V model assumes that the distribution of the sensations is normally distributed with means equal to the scale location of the stimulus and with variances being equal across stimuli. The method further assumes that the sensations are uncorrelated both within and between the jj-12 stimuli pairs. Under this setup, the probability that a judge would rate Item j above Item k is equal to Φ βj−βk, where Φ is the standard normal cumulative distribution function, and βj and βk are the locations of Items j and k, respectively; this formulation assumes the common variance of the psychological sensations, which can be arbitrarily set to a value, is equal to 12. To obtain estimates of the stimulus locations, one can use standard probit regression methods.

The Bradley-Terry-Luce method for paired comparisons is similar; however, it is formulated by assuming that the psychological sensations are logistic distributed rather than normally distributed. The resulting probability of choosing Item j over Item k is equal to Ψβj − βk, where Ψ is the *expit*, or inverse logit function expċ1 + exp. In this case, the parameters are estimated using standard methods for logistic regression.

The Method of Equal-Appearing Intervals

The drawback of the method of paired comparisons for locating items is that it requires a huge number of comparisons. For example, if the researcher started with 100 items, the method of paired-comparisons would require each judge to make 4,950 comparisons. Redesigning the data collection method (e.g., using a balanced incomplete block design) reduces the number of comparisons any one judge needs to make; however, the method still tends to be rather tedious.

Because the method of paired comparisons requires so many judgments, Thurstone and E. J. Chave developed the method of equal-appearing intervals. In this method, the judges are asked to separate the items into some fixed number of rating intervals according to where the judge believes the items are located on of rating intervals according to where the judge believes the items are located on the latent continuum. Assuming the rating intervals are of equal width, the intervals are assigned consecutive scores (e.g., 1–11), and the scale value assigned to each stimulus is estimated by the median score received by the item.

To select a subset of the items from the original pool to create the final battery of items, the method of equal-appearing intervals suggests selecting items that are nearly uniformly distributed across the scale. The interquartile ranges of the ratings assigned to the stimuli are also examined, with items with low interquartile ranges preferred over items with large interquartile ranges.

The Method of Successive Intervals

Thurstone considered the method of equal-appearing intervals as a way to approximate the method of paired comparisons. Realizing that the method of equal-appearing intervals and the method of paired comparisons did not yield perfectly linear results, he developed the method of successive intervals. This method also asks judges to sort the items into some number of interval categories, just as in the method of equal-appearing intervals. However, the intervals are not assumed to be of equal width.

As in the method of paired comparisons, the method of successive intervals assumes that each stimulus produces a psychophysical sensation in each judge that is normally distributed centered at the scale location βj, but with possibly different variances, denoted σj2. The method further assumes that there are thresholds defining the boundaries for each successive interval, denoted τ1, …, τJ−1. If the sensation observed by a judge is between the thresholds τk−1 and τk, the judge would place the stimulus in interval k (τ0 = −∞, τJ = ∞). This formulation produces a multinomial probit model for the judges’ scoring of the stimuli. Standard methods for probit regression are used to estimate the stimuli locations βj and the variances σj2. As in the other methods, stimuli are selected to cover the entire range of the scale, and stimuli with low variances are preferred over items with large variances.

Scaling Respondents

Once the survey items have been located on the latent continuum according to one of the three procedures discussed in the preceding paragraphs and the final set of items has been selected, the items are used to measure the attitudes of a set of respondents. Respondents are asked to examine each stimulus separately and to record whether they endorse (e.g., like/dislike) the stimulus. The method assumes that respondents will endorse only those stimuli that are located near the respondent, which implies that the probability that an individual endorses an item is a unimodal function of the attitude location. The average or median location of the items endorsed by the respondent serves as the estimated location of the respondent on the attitude scale.

Alternatives

Although Thurstone’s methods were popular for the 10–20 years after they were originally developed, a number of advances have made them much less popular today. Although the original Thurstone methods required two stages to locate stimuli and respondents on the attitude scale, a number of alternatives allow the stimuli and respondents to be scaled simultaneously. Likert scaling, Guttman scaling, the Rasch model, and item response theory models are one-step alternative methods that are appropriate when the stimuli are monotone and unidimensional (i.e., respondents with more positive attitudes are more likely to endorse all items). Clyde Coombs’s deterministic unfolding method and its probabilistic alternatives are appropriate for scaling unidimensional attitudes and stimuli, when the stimuli are unimodal; for example, if respondents were asked whether they endorse a particular politician, they could choose not to endorse the politician, either because the politician is too liberal or because the politician is too conservative. Multidimensional scaling techniques can be used to examine multidimensional attitudes.

Matthew S. Johnson

See also Attitude Scaling; Guttman Scaling; Item Response Theory; Likert Scaling; Logistic Regression; Multidimensional Scaling; Probit Transformation; Psychometrics; Rasch Model; Scales; Semantic Differential Scaling

Further Readings

Bradley, R. A. (1976). Science, statistics, and paired comparisons. Biometrika, 32(2), 213–219.

Coombs, C. H. (1964). A theory of data. New York, NY: Wiley.

Jansen, P. G. W. (1984). Relationships between the Thurstone, Coombs, and Rasch approaches to item scaling. Applied Psychological Measurement, 8, 373–383.

Saffir, M. A. (1937). A comparative study of scales constructed by three psychophysical methods. Psychometrika, 2(3), 179–198.

Thurstone, L. L. (1927). A law of comparative judgement. Psychological Review, 34(4), 278–286.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33(4), 529–554.

Thurstone, L. L., & Chave, E. J. (1929). The measurement of attitude. Chicago, IL: University of Chicago Press.

W. Holmes Finch W. Holmes Finch Finch, W. Holmes

Time Series Analysis Time series analysis

1701

1704

Time Series Analysis

In some applications, researchers collect longitudinal data for one or more subjects over a (usually) extended period of time. In the field of economics, for example, the federal government of the United States measures the nation’s gross domestic product every month, creating a long longitudinal record of gross domestic product over time. Similarly, meteorologists record measurements on temperature and rainfall, monitoring every day at stations around the world. Again, the resulting data set contains a great many measurements taken over a long period of time for each of these stations. Psychologists may collect such time series data in the form of diary entries in which participants are asked to record the number of times that they have certain thoughts or engage in specific behaviors over the course of each day, over many weeks or months. Educational researchers might look at student assessment scores across many years. The resulting observations represent a time series data set.

In all of these examples, the common trait is the recording of data values longitudinally for an extended period of time. The statistical methodologies designed to deal with these longitudinal records are known collectively as time series analysis. Time series analysis differs from more traditional repeated measures data in that the number of data points is much larger in the former than the latter, which typically involves only a few measurements for each individual. However, some of the same issues that are created by the collection of repeated measures data are also present in time series. For example, one of the core assumptions underlying many statistical analyses is the independence of the specific data points used. However, a signal quality of time series data is the presence of correlation in measurements taken over time (referred to as autocorrelation). In other words, a measurement made at time t is likely to be correlated with a measurement made at the immediately preceding time, $t - 1$.

Indeed, it is entirely possible that the measurement at time t is autocorrelated with measurements at $t - 2$, $t - 3$, and so on, although the magnitude of this relationship would be expected to decline over time.

Given this lack of independence in data points, standard analyses such as regression are not appropriate to use with time series data because this serial correlation structure leads to biased estimation, particularly for model parameter standard errors. This fact has given rise to an entire family of procedures designed to correctly model the autocorrelation that is present in time series, and indeed to use it for forecasting future measurements, as well as correct other modeling procedures such as regression. In the following paragraphs, discuss some of the basic time series models that are available for researchers and data analysts to use are discussed. However, it should be noted that these represent only a small number of all of the possible models that can be used with time series data.

Common Time Series Models

The field of time series is replete with a wide array of models for use in specific situations and to answer specific kinds of questions. It is well beyond the purview of this brief entry to describe all of these. However, there are three common models that serve as the backbone of much time series modeling. Each of these will be discussed briefly herein. The first time series analysis that we will examine is the autoregressive (AR) model, which can be written as:

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \ldots + b_i y_{t-i} + e_t \, ,$$

where y_t = measurement of response variable at time t; $y_{t\,-\,\mathrm{i}}$ = measurement of response variable at time $t - i$; b_0 = intercept; b_i = autocorrelation parameter for time $t - i$; e_t = random error (white noise) term for time t.

The order of this AR function is equal to the largest value of i such that the coefficient b_i is not 0. In other words, the order of the AR(p) function is equal to the number of previous time points that have an impact on the value of y_t. If only the value of y at the previous time point is related to the current value of y, then we would have an AR(1) process. If the last 3 times impact the value at the current time, then we have an AR(3) process.

The second major model that serves as a foundation to time series modeling is the moving average (MA) model. The MA(*q*) model is expressed as:

$$y_t = \mu + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_i e_{t-i} + e_t ,$$

where μ = overall mean of the time series; $e_{t - \mathrm{i}}$ = random error (white noise) term for time *i*; e_t = random error (white noise) term for time *t*; θ_i = model parameters linking the white noise from time $t - i$ to the response at time *t*.

The primary difference between the MA and AR models is that in the latter, we assume that the value of the current measurement in the series is directly correlated with prior measurements of the same variable in the series, whereas in the former, we posit that the current value of *y* is related to random errors from previous time points rather than the measurements themselves. These prior error terms are known as random shocks. The order of the MA model is equal to the number of prior random shocks that are found to be directly related to y_t. Thus, a model for which the last four random shocks all have nonzero values of θ_i would be termed an MA(4) model.

Finally, we can combine the AR and MA models together to form the ARMA(*p*,*q*) model.

$$\begin{aligned} y_t = {} & b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \ldots + b_i y_{t-i} + \\ & \mu + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_i e_{t-i} + e_t . \end{aligned}$$

All terms are as defined previously. This model incorporates elements of both AR and MA processes. It is possible that the values of *p* and *q* are not equal, meaning that the prior times’ impacts on the outcome vary depending on whether we consider the measurement itself or the error associated with those measurements. Thus, a time series in which the three previous measurements are autocorrelated with the current value, and where the random shocks from the prior two measurements are also related to the current value, would be referred to as an ARMA(3,2) process.

Selection of Appropriate Models

The selection of the optimal model for a time series is done using a combination of tools, including the autocorrelation function (ACF) and partial autocorrelation function (PACF), as well as penalized measures of unexplained variance. The ACF is simply the set of correlations between the target measurement, y, at each time, t, with measurements of y at all previous times in the series. The PACF is the set of correlations between y at time t with each previous value of y, *controlling for* the measurements of y at all previous times. For example, the PACF for y_t and y_{t-1} reflects the relationship between these two measurements, controlling for y_{t-2}, y_{t-3}, and so on. In contrast, the ACF does not control for these earlier measurements or their correlation with y_t, thereby reflecting the raw relationship between y_t and y_{t-1} in our example. Researchers typically refer to graphical representations of the ACF and PACF to diagnose the appropriate lag for the AR and/or MA functions.

In addition to the ACF and PACF, information functions, such as the Akaike information criterion (AIC) and Schwarz’s Bayesian information criterion (BIC), can be used to identify the optimal model to be selected. Information indices reflect the amount of variance in the data that is not explained by the model, with the addition of a penalty for model complexity. Therefore, larger values of the AIC or BIC reflect a relatively worse fitting model. If two models explain the same amount of variation in the data, the one with fewer terms (i.e., the simpler model) will have a lower AIC or BIC, making it the preferred model. Researchers can make use of the AIC and BIC to select the optimal model by fitting several, based on evidence from the ACF and PACF, and then comparing the fit indices. The model with the lowest AIC/BIC would then be selected, given that it is the most parsimonious that also explains a reasonable amount of variation in y_t. Using this approach, the researcher would be able to compare a wide variety of model forms, with differing numbers of lags for the AR and MA processes.

Extensions of Time Series Models

Time series models based on the AR, MA, or ARMA processes can be extended to account for seasonal patterns in the time series and can include covariates that predict y_t. These covariates can be static (measured at a single point in time) or time covarying. Furthermore, functions of other statistics, such as the variance, can also be incorporated into time series analyses leading to the ARCH family of models. For these models, it is assumed that the variance of the current error term (e_t) is a function of the prior error terms. Variations of these techniques include both AR and MA processes in the estimation of the error term (GARCH), the addition of nonlinear model terms (NGARCH), and continuous time-generating functions (COGARCH). As with the ARMA models, these error terms can be fit using AR or MA processes, and model selection is carried out in much the same fashion as described earlier for the ARMA family of models.

It is also possible to relate two time series to one another using the cross correlation function (CCF). The CCF is simply a measure of the correlation between each value of one series (y) with each value of the other series (x). Thus, using the CCF, we can obtain the correlation between, say, y_t and x_{t-1} in order to determine whether one series appears to presage or predict the other at an earlier time. Such CCF analyses can be further developed in the form of regression models where the values of one function can be used to predict the values of the other using traditional regression approaches.

A primary use of time series models is the forecasting of future values. In particular, economists use them in this fashion to provide estimates of economic growth, labor market participation, and the like. The advantage that such models have in forecasting is that they account for the long chain of relationships over the entire course of the time series rather than employing only a single point in time as is common with panel data. In addition, time series model can incorporate seasonality into the forecasts, as well as additional covariates, making it a very powerful and flexible tool for researchers.

W. Holmes Finch

See also Repeated Measures Designs; Survival Analysis

Further Readings

Box, E. P., & Jenkins, G. M. (2015). Time series analysis: Forecasting and control. Hoboken, NJ: Wiley.

Mills, T. C. (2011). The foundations of modern time series analysis. New York, NY: Palgrave Macmillan.

Montgomery, D. C., & Jenning, C. L. (2016). Introduction to time series analysis and forecasting. Hoboken, NJ: Wiley.

Terasvirta, T., Tjostheim, D., & Granger, C. W. J. (2010). Modeling nonlinear economic time series. Oxford, UK: Oxford University Press.

TIMSS

1704

TIMSS

See Trends in International Mathematics and Science Study

Bonnie Cramond Bonnie Cramond Cramond, Bonnie

Torrance Tests of Creative Thinking Torrance tests of creative thinking

1704

1706

Torrance Tests of Creative Thinking

Although E. Paul Torrance developed several assessments of creative thinking over his many years of research, the two tests of creative thinking commonly referred to as the Torrance Tests are actually entitled *Thinking Creatively with Pictures and Thinking Creatively with Words*. These two assessments, originally created in the 1950s, are the most used measures of creativity worldwide, have had the most research conducted with and about them, and have been translated into over 40 languages. Published by Scholastic Testing Company, these measures were designed for administration to children and adults aged 4 years and older. Originally intended for use in classrooms, they can be administered to groups or individuals with just pencils and in less than an hour. Scoring of the tests requires some training but can be achieved with high reliability with relatively little time and effort as compared with training to score individual intelligence tests. Alternatively, the tests can be sent to the publisher for scoring. The results produce a composite score and subtests scores, which can be converted to national percentiles by age and grade. This entry reviews the development, components and scoring, and uses of the Torrance Tests.

Development

Torrance began the development of the tests in the 1950s with the help of graduate assistants based on intense study of creativity definitions, characteristics of creative people, creative behaviors, and extant efforts at measuring creativity. Then, he decided on the criteria that each task had to fit his definition of creativity as a natural, everyday process as well as be:

1. suitable for students from kindergarten through graduate and professional

school,

2. easy enough for the young or disabled to make a creative response, yet difficult enough to challenge the most able,
3. unbiased with regard to gender and race,
4. open ended to allow for responses from different experiential backgrounds,
5. enjoyable and motivating.

Torrance and his assistants then chose tasks from existing measures and created new ones, which he personally screened as they were administered to various groups. This resultant battery, called the Minnesota Tests of Creative Thinking, was administered to students in elementary and high school in the Fall and Winter of 1958–1959. In 1966, when Torrance moved to the University of Georgia, the tests were retitled as Torrance Tests of Creative Thinking (TTCT).

Components and Scoring

The complete battery of the TTCT includes a verbal test consisting of six activities and a figural test consisting of three activities. Each test has parallel forms A and B. All raw scores can be converted to standard scores and national percentiles by age and/or grade.

The Verbal Tests

The verbal test, Thinking Creatively with Words, begins with a thought-provoking picture. Respondents write their responses, although young children or those with a disability can dictate their responses. For the first three activities, the respondent is required to ask questions about the picture, guess causes of the events in the picture, and then predict consequences from the picture. These activities are designed to elicit curiosity, speculation, and hypotheses beyond what is depicted.

The fourth and fifth activities, Product Improvement and Unusual Uses, require the respondent to think of improvements for a toy and imagine all of the possible uses for a common object. Torrance noted that economically disadvantaged individuals seem to have an advantage in responding to these two activities, probably because of their experiences with “making do” with what they have.

Finally, the sixth activity, Just Suppose, presents respondents with an unusual and unlikely situation from which they are to conjecture about possible and unlikely situation from which they are to conjecture about possible consequences. This is to assess individuals’ tolerance and playfulness with unusual ideas.

These six activities are each scored for (a) *fluency*, the number of relevant ideas; (b) *flexibility*, the variety of the ideas; and (c) *originality*, the unusualness of the ideas. The test results give a score for each of these measures and an overall composite score.

The Figural Tests

The figural TTCT is comprised of three activities that require a drawn response with a title for each. Drawings without titles are acceptable, and someone can write the titles for those who are not able to do so.

For the first activity, respondents are asked to add details to a solid shape in order to create an interesting picture that tells a complete story. Respondents are encouraged to try to think of something that no one else could think of and add details.

For the second activity, respondents are presented with 10 incomplete simple line drawings, and for the third activity, they are presented with pages of repeated figures of the same kind to which they must respond by adding details to create pictures. They are given the same directions to encourage originality and elaboration, but for these activities, they are also encouraged to think of as many ideas as they can, thereby promoting fluency.

In addition to fluency and originality, the figural tests are scored for elaboration, the degree of detail in the drawings; resistance to premature closure, measuring open-mindedness through the ability to delay closure of the incomplete figures; and abstractness of titles, measuring the ability to synthesize and represent ideas beyond the concrete. The figural tests also have 13 criterion-referenced abilities that are measured for appearance anywhere on the test, such as humor, emotional expression, fantasy, and boundary breaking. The figural tests report a score for each of the norm-referenced measures as well as a composite score. They also report a creativity index score that includes the criterion-referenced abilities.

Uses

The TTCT have been used to identify creativity in students for admission to special programs, to evaluate the effectiveness of interventions designed to increase creativity, and to measure individuals' creativity for research purposes. In 1959, results from longitudinal studies with children tested have shown good predictive validity with adult creative behaviors after 22-, 40-, and 50-year time spans.

Bonnie Cramond

See also Aptitude Tests; Creativity; Culturally Responsive Evaluation; Predictive Validity

Further Readings

Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium. Gifted Child Quarterly, 49, 283–291.

Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2011). Torrance Tests of Creative Thinking as predictors of personal and public achievement: A fifty year follow-up. Creativity Research Journal, 22(4), 361–368.

Torrance, E. P. (1981). Predicting the creativity of elementary school children (1958–80)—and the teacher who "made a difference." Gifted Child Quarterly, 25, 55–62.

Torrance, E. P., & Safter, T. H. (1999). Making the creative leap beyond. New York, NY: The Creative Education Foundation Press.

Johnny Saldaña Johnny Saldaña Saldaña, Johnny

Transcription

Transcription

1706

1707

Transcription

Transcription is the process of converting audio and/or video recording soundtracks into written forms for data documentation and analysis. Recordings can consist of naturally occurring talk in social settings, individual or focus group interviews with study participants, and media sources. Transcription can also occur simultaneously with talk, but for research purposes, audio recordings are most often used. High-quality recording equipment should be utilized; the better the audio, the easier it is to transcribe.

An audio recording is first heard several times by a transcriber to gain a holistic overview and to familiarize with the contents. Next, the recording is transcribed using text editing software. The text draft is then reviewed along with the recording and corrected, as needed, for transcription accuracy. Final or required formatting is applied to the finished document, then the recording is archived or erased, according to any preestablished research protocols.

As of 2016, technology exists that automatically converts speech to text, yet complete software accuracy is still not realized. Researchers can employ specialized digital tools to assist with the oft perceived drudgery of the task. But manual transcription—that is, repeated listening and stopping and restarting the recording to document talk data—is still the researcher's common way of working.

Verbatim transcription documents every spoken word, utterance, and vocal tone as interpreted by the transcriber: "Yeah, it was like a, what do you call it? Like, um, a rave! Yeah, that's it." Aside from punctuation, rich text and parenthetical notes can supplement the transcription: “Like, um, a RAVE! (*nods head vigorously*) Yeahhh, that’s it.” The Jefferson notation system uses standardized symbols to document speech intonations in print: “a ↑r*a*ve!”

Verbatim transcription is particularly important for conversation and discourse analyses in the social sciences. Children’s and adolescents’ speech should also be transcribed verbatim for language and developmental research. It is ill-advised to “clean up” or truncate authentic speech because this compromises the data’s fidelity. But not everything on a recording has to be transcribed; some prefer to focus only on salient passages related to the research questions of interest.

It is highly recommended that the researcher who recorded the fieldwork conversation or conducted the interview also transcribes the recording. The researcher will be intimately familiar with the field site contexts and will retain selected memories of the conversational interaction, thus generating a possibly more expedient transcription process and a more accurate document. Plus, the act of transcribing gives the researchers cognitive ownership of the data because they must listen to the recording several times and document literally every word spoken. Analytic reflections and insights may also occur during the process, which can be documented within the transcript or logged in a separate file.

Some researchers delegate transcribing to assistants or to commercial, professional transcription services. This provides the researcher more time for other necessary tasks but requires funds to pay others for their work. Also, transcription accuracy can be questionable if the researcher does not verify the finished document with the original recording. And, the researcher possibly loses more intimate familiarity with the data by assigning transcription tasks to others.

Johnny Saldaña

See also Data; Field Notes; Focus Groups; Interviews

Further Readings

Bischoping, K., & Gazso, A. (2016). Analyzing talk in the social sciences: Narrative, conversation & discourse strategies. London, UK: Sage.

Lisi Wang Lisi Wang Wang, Lisi

Andrew C. Butler Andrew C. Butler Butler, Andrew C.

Transfer

Transfer

1707

1710

Transfer

Transfer can be defined as the extent to which the knowledge or skills learned in one context affect performance or learning in another context. Within the field of education, the concept of transfer is often referred to as *transfer of learning* or *transfer of knowledge*. However, other fields use slightly different terms to describe this concept, such as *transfer of training,* which is often used in industrial and organizational psychology. Transfer of prior learning to a new context is considered to be *positive* if it facilitates performance or learning in this other context and *negative* if it hinders performance or learning. For example, learning a new language might be facilitated if the grammar of a person's native language is similar to the new language, resulting in positive transfer. However, if the grammar is very different, learning of the new language may be hindered, which would constitute negative transfer.

Transfer is an important topic within education because a primary goal of education is to help students acquire knowledge and skills that they can use in a broad variety of future contexts. This entry provides an overview of transfer as it applies to educational research and practice. First, a framework for understanding transfer is described that includes multiple dimensions along which transfer can occur. Second, a brief history of the early research on transfer is provided. Third, modern perspectives on transfer are discussed. Finally, implications for educational practice are considered.

A Framework for Understanding and Interpreting

Transfer

The transfer literature is rife with contradictory findings and conclusions. However, these disagreements often stem from inconsistencies in the use of modifiers for the term *transfer*. For example, a common distinction used in the literature is between *near* and *far* transfer. Near transfer refers to a situation in which the context of original learning and the new context are similar, whereas far transfer refers to a situation in which the contexts are very different. Although this distinction seems clear in abstract form, such a simple definition of near and far transfer becomes problematic when different researchers try to apply it to different sets of contexts. The result is the inconsistent use of terminology within the literature, where what constitutes near transfer in one study might be considered far transfer in another study. Other distinctions about the nature of transfer have been equally problematic in practice, such as *parallel* versus *vertical* transfer (i.e., whether original learning is on the same level or subordinate to new learning) and *specific* versus *general* practice (i.e., the degree of specificity in the relationship between original learning and new learning).

In an effort to resolve the confusion regarding the use of terminology and make sense of the findings in the literature, Susan Barnett and Stephen Ceci proposed a framework that conceptualizes the process of transfer as a set of dimensions, each characterizing a continuum along which the new context can differ from the context of original learning. The dimensions proposed in the framework are the following: *knowledge domain* (the domain to which the knowledge or skill belongs; e.g., physics, art, and sociology), *physical context* (the environment in which learning takes place, e.g., classroom, playground, and laboratory), *temporal context* (the amount of time that elapses between original learning and new learning; e.g., minutes, days, and years), *functional context* (the perspective or mind-set with which the individual views the situation; e.g., academic test, informal social interaction, and leisure activity), *social context* (whether the learner is alone or learning with others; e.g., alone, in a pair, and in a small group), and *modality* (the sensory modality and structure of the learning activity; e.g., visual, auditory, and written). Each dimension is represented as a continuum along which the degree of transfer can vary from near to far. For example, near transfer along the dimension of knowledge domain might consist of learning about physics and then applying this knowledge to another topic within physics or applying it to another science like biology. In contrast, far transfer might consist of learning about physics and then applying this knowledge to sociology or art. Importantly, two contexts can vary in terms of one or more dimensions, thus providing an additional way to conceptualize near versus far transfer.

Early History of Research on Transfer

The concept of transfer has long been a topic of interest within education. The history of empirical research on transfer dates to around the turn of the 20th century, when researchers began to evaluate a popular idea in education called the *doctrine of formal discipline*. Formal discipline conceptualized the mind as a muscle that could be strengthened by exercise. Such exercise was assumed to develop general thinking skills that would transfer broadly, so the specific type of exercise did not matter as long as it was sufficiently rigorous. This theory was used to support pedagogical practices (e.g., rote memorization of large quantities of information) and a focus on particular subjects (e.g., rhetoric, Latin, and mathematics) that were thought to be particularly good for strengthening the mind.

Given the importance of the doctrine of formal discipline to the practice of education in the United States and Britain, numerous researchers sought to test it. However, they approached the question in different ways, used different methods, and found different results (a portent of how the transfer literature would develop in the coming decades). For example, in one classic study, children were either trained on the law of refraction or not, and then they were given a task in which they had to use a dart to hit an underwater target. The children who were informed of the principle were more likely to hit the target, leading to the conclusion that instruction that emphasizes broad principles instead of specific details could promote general transfer. In another classic study, children were given instruction and practice on estimating the area of certain geometric shapes and then asked to estimate the area of different shapes. The children largely failed to apply their prior learning and performed poorly on the new estimation task, yielding the conclusion that transfer only occurs when the two tasks are highly similar.

One of the first and most influential psychological theories of transfer was the *theory of identical elements* proposed by Edward Thorndike. The basic idea is that transfer is most likely to occur when the elements or features of the learning task in one context match those of a task in the new context. That is, the greater the similarity between contexts, the more likely it is that transfer will occur.

Thorndike’s theory and research was influential in sparking research into the conditions of learning that promoted transfer. During the first half of the 20th century, much of this research was conducted within the field of verbal learning, the human-focused branch of behaviorism. The basic paradigm used for this research was learning pairs of words (e.g., dog-table). Participants in these experiments would learn a set of word pairs and then later learn another set of word pairs. Of interest was how the relationship between the two sets of word pairs affected transfer, which was operationalized in terms of how easy it was to learn the second set of word pairs. When learning of the first set of word pairs facilitated learning of the second set of pairs, positive transfer was observed. In contrast, when learning of the first set of word pairs interfered with the learning of the second set of pairs, negative transfer was observed.

Modern Perspectives on Transfer

Since the decline of interest in the behaviorist approach to learning in the 1950s, the concept of transfer has been investigated within a rapidly growing and diverse set of perspectives. Indeed, transfer is an essential concept within any perspective or theory that attempts to explain how learning occurs. More broadly, transfer has been a topic of interest for almost every field in education, including assessment, special education, curriculum and instruction, and teacher training. Given the difficulty of representing how all of these perspectives and fields conceptualize the process of transfer, two approaches are described to illustrate this diversity.

One approach to understanding transfer that has been especially fruitful is the *cognitive perspective*. Since the 1960s, researchers who study human cognition have investigated the mental processes that support and produce transfer. Research on transfer is central to many areas of cognition, including creativity, critical thinking, and problem solving. One area that has produced much progress in understanding transfer is the study of analogical reasoning. Analogical reasoning is thinking that involves examining the similarities between an original learning context and new context (e.g., a previously solved problem and a new problem) in order to determine the knowledge or procedure needed in the new context (e.g., how to solve the new problem). Analogical transfer involves recognizing similarities in content and structure between the two contexts, remembering learning from the original context, and mapping that learning to the new context. Thus, within the cognitive perspective, analogical transfer can be thought of as a process of decontextualization of learning. Critical to the success of analogical transfer is the ability to distinguish between structural and surface features. Structural features refer to the underlying similarities in the knowledge or procedure required between the two contexts that enable prior learning to be useful in the new context (e.g., the method for solving the two problems must be the same). In contrast, surface features refer to aspects of the two contexts that are not causally related to the transfer of prior learning (e.g., unimportant details in either of the descriptions of the two problems). Although the two contexts must share structural features in order for transfer to be possible, they may differ substantially in terms of surface features. The degree of similarity or dissimilarity of the surface features between two contexts is often important to determining whether people will recognize that an analogy exists between the two contexts. Similarity in surface features without matching structural features often leads people to incorrectly think an analogical relationship exists between two contexts when it does not.

In contrast to the cognitive perspective in which transfer involves decontextualization, the *situated learning perspective* emphasizes the importance of context in learning and, more generally, determining whether transfer will occur. One idea that learning is contextually bound is central to the situated learning perspective. That is, it is difficult, and often impossible, to divorce learning from the context in which it occurs. Learning occurs when individuals interact with the contexts in which they are situated (e.g., social, cultural), and this pattern of interaction determines how they construct their knowledge. Given the contextual nature of learning within the situated learning perspective, the potential for transfer is constrained to other contexts that are highly similar. As a result of the limited ability to decontextualize learning, the situated learning perspective emphasizes the importance of engaging learners within authentic learning contexts that are highly similar to the contexts to which learning will be transferred.

Implications for Educational Practice

Although each perspective on learning offers different recommendations for how transfer can be promoted in educational practice, these recommendations are best viewed as complementary rather than contradictory. For example, the cognitive and situated learning perspectives differ in that they focus on decontextualization of learning within an individual and an individual's learning through interactions with context, respectively. However, both these through interactions with context, respectively. However, both these perspectives can provide useful recommendations for education practice. The cognitive perspective would suggest that learning should be facilitated in a way that emphasizes a focus on structural features rather than surface features within a context. One way to accomplish such learning is to engage students in learning across a variety of contexts that share underlying structural features but differ in surface details. In contrast, the situated learning perspective would suggest that students should engage in learning within contexts that are similar to the contexts to which they will be expected to transfer their learning. In reality, these recommendations are compatible in that both place an emphasis on the similarity between the context of original learning and the new context to which learning will be transferred and helping students to understand the similarities and differences between these two contexts. Finally, more generally, research also suggests that teaching students metacognitive strategies and encouraging student motivation will also promote transfer. Metacognitive strategies (e.g., planning and monitoring) and motivation can benefit transfer by increasing effort, persistence, and cognitive engagement as the process unfolds.

Lisi Wang and Andrew C. Butler

See also Bloom's Taxonomy; Causal Inference; Creativity; Critical Thinking; Generalizability; Learning Theories; Outcomes; Problem Solving

Further Readings

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128, 612–637.

Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), Transfer on trial: Intelligence, cognition, and instruction (pp. 1–24). Westport, CT: Ablex Publishing.

Haskell, R. E. (2000). Transfer of learning: Cognition and instruction. New York, NY: Academic Press.

McGeoch, J. A. (1942). The psychology of human learning (pp. 394–452). New York, NY: Longmont, Green, and Co.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. Psychological Bulletin, 115, 381–400.

Jennifer L. Jewiss Jewiss Jewiss

Transformative Paradigm

Transformative paradigm

1710

1711

Transformative Paradigm

The transformative paradigm is rooted in the recognition that injustice and inequality are pervasive and the belief that research and evaluation are important tools for addressing these societal ills. As articulated by Donna Mertens, a leading transformative research and evaluation scholar, this paradigm maintains that research and evaluation can and should play an explicit role in identifying and alleviating discrimination and marginalization based on factors such as race, ethnicity, religion, gender, sexual orientation, socioeconomic status, age, and disability. The transformative paradigm draws from and provides an overarching means of categorizing an array of theoretical perspectives focusing on the concerns of distinct populations addressed by feminist, indigenous, postcolonial, queer, disability rights, and critical race scholars. Researchers and evaluators operating within this paradigm examine power dynamics and systems that privilege certain groups over others. They investigate policies and practices that perpetuate inequities in settings such as schools, communities, and social programs. Transformative approaches extend beyond knowledge generation and take an activist stance in promoting social justice.

The transformative paradigm views knowledge as a social construction shaped by the knower’s individual experiences, personal characteristics, and community affiliations. As a result, researchers and evaluators as well as study participants are called to reflect on their own beliefs, consider how beliefs are shaped by one’s identity and life experiences, and critically examine how such beliefs may influence one’s perspectives on the study topic and methods. This paradigm acknowledges that privileged groups, including researchers and evaluators, are typically afforded greater say in constructing knowledge sanctioned by academic institutions, government agencies, and other official entities. Consequently, institutions, government agencies, and other official entities. Consequently, transformative researchers and evaluators encourage traditionally marginalized groups to play a central role throughout the study process to ensure the findings are inclusive and represent the perspectives of all relevant groups.

In contrast to scholarly traditions that position the researcher at a distance from participants in an attempt to establish objectivity, the transformative paradigm values the development of trusting relationships and collaboration with participants. Transformative researchers and evaluators acknowledge their need to learn from community members and view participants as essential partners. This paradigm aims to give voice to local knowledge held by a diverse array of participants. Particular attention is paid to the members of marginalized groups traditionally excluded from research and evaluation efforts or viewed merely as study “subjects.” Within the transformative paradigm, members of marginalized groups are seen as having their own individual and community strengths and legitimate knowledge systems. Study processes intend to bolster strengths and foster resilience within the community. The researcher or evaluator is responsible for facilitating accessible opportunities for participants to engage in core tasks, such as defining research and evaluation questions, analyzing data, and interpreting findings. This paradigm allows for the use of culturally appropriate quantitative, qualitative, and mixed methods. Qualitative methods, such as individual interviews and focus groups, often play a prominent role in transformative studies, given the need to engage in dialogue with a cross section of participants to develop in-depth understanding of the range of perspectives within the community. Of utmost importance is that researchers and evaluators ensure study findings are used to advance social justice.

Jennifer L. Jewiss

See also Cultural Competence; Culturally Responsive Evaluation; Democratic Evaluation; Empowerment Evaluation; Feminist Evaluation; Minority Issues in Testing; Multicultural Validity

Further Readings

Mertens, D. M. (2009). Transformative research and evaluation. New York, NY: Guilford.

Amber Rowland Amber Rowland Rowland, Amber

Treatment Integrity

Treatment integrity

1711

1714

Treatment Integrity

Treatment integrity is the degree to which an intervention is implemented as intended. Fidelity, including accuracy and consistency to the independent variable or intervention, is important so that changes in the dependent variable can be attributed to the intervention. Treatment integrity helps ensure that the treatment was carried out the way it was designed. This entry describes the importance of treatment integrity, how it can be measured, influences on treatment integrity, and how it can be improved.

Many authors have noted how treatment integrity is complex and multidimensional, thus difficult to adhere to, measure, and report. Social science and specifically teacher-implemented interventions have a high risk of treatment integrity lapses due to the nature and variability of classrooms. Lack of adherence to treatment integrity during implementation of a teaching intervention can negatively influence student performance. When an intervention is not implemented as intended and data start revealing lack of growth or improvement, how does a teacher decide the reason behind the lack of success? Is the intervention a “bad” intervention or is the lack of treatment integrity to blame?

Evidence-based practices have been refined and researched and are intended to be implemented in certain ways. Reporting how well the interventionist adhered to the intended treatment helps consumers of the research understand that there is enough evidence to support that the use of a particular intervention will lead to expected student outcomes in a new setting.

Insufficient attention to treatment integrity in research dissemination greatly limits confidence that findings represent an effective intervention. Indeed, the social sciences have historically struggled with reporting on the relationship between how well the independent variable (intervention) was implemented as intended and the resulting influence it had on the dependent variable (e.g., student learning). Social science research in general tends to report on changes to the dependent variable as the ultimate mark of effectiveness in research without reporting on the fidelity of the implementation. For instance, a study may examine a protocol for middle school students to learn a strategy for increasing organization and productivity in writing. If there are seven components necessary for successful implementation, it is important to note the presence of each component before reporting on changes to student writing. Frequently, the independent variable is described and results in student writing (dependent variable) are reported; but without an understanding of how well the strategy was implemented, it is unclear whether the changes in student writing are to be attributed to the intervention or other variables, such as an increase in attention to, or frequency of, overall writing. Purposeful steps must be taken, throughout a research study to ensure treatment integrity.

Measurement of Treatment Integrity

The optimal ways to measure treatment integrity include direct observations and analysis of student products. A less reliable, but frequently used method of measurement includes self-report surveys, activity logs, or interview data. Although self-report methods are more convenient, there is greater opportunity for misunderstanding, misrepresentation, and skewed data because teachers may be aware of researcher intent and/or administrator expectations and be inclined, even if unintentionally, to report what they believe is expected instead of what actually happened. Even slight adjustments in reporting threaten the integrity of data, so it is helpful to include direct observation or analysis of products in conjunction with self-report data.

For direct observation, checklists and time sampling can be used to determine how accurately and/or frequently the independent variable is implemented. For instance, researchers may be working with teachers on behavior modification for students who struggle with keeping their hands to themselves when walking around the classroom. An independent observer would score a correct trial when the child successfully walks around the room without touching anyone else and the teacher reinforces the successful trial with a token. Observing the number of correct trials divided by the number of correct plus incorrect trials and correct trials divided by the number of correct plus incorrect trials and multiplying by 100 to yield a percentage of integrity can measure the integrity of the implementation. For example, for 1 hour-long observation, the student got up and walked around the room 10 times. Of those 10 trials, the student successfully avoided touching peers 8 times and was reinforced by the teacher with a token 7 times. So the integrity of the intervention was 70%.

For analysis of products, a rubric or checklist can be used to determine whether the necessary components are included in a submitted product. For instance, using the aforementioned writing example, students may have produced interactive graphic organizers using a writing strategy to organize their thoughts before drafting an essay. The writing strategy may have four components, so a checklist would have the four components listed in a column, with a “yes” or “no” column. The researcher would indicate the presence or absence of each of the four components in the submitted product by marking the corresponding column on the checklist.

For self-report data, surveys, logs, or interviews can be used. For any form of self-report data, similar tools can be used for educators to score themselves on the inclusion of necessary implementation components. An educator can use the same rubrics or checklists used by an observer to indicate the perceived treatment integrity. Indeed, comparing the treatment integrity results between what was observed during an implementation trial and how an educator self-reported the implementation can serve as helpful conversation starters between a researcher, instructional coach, and/or an educator. Self-report data may be inflated and can provide valuable insight if an educator is struggling to see adequate changes in student performance. Clarification of implementation components is a good use of self-report data.

Influences on Treatment Integrity

In 2009, Frank Gresham described a continuum of possible treatment integrity scenarios from low to high. The first scenario occurs when the intervention (independent variable) is implemented as it was intended and the dependent variable shows expected outcomes. Using the previous example of students walking around in class and keeping their hands to themselves, if the students perform the desirable behavior and the teacher offers the reinforcing token for every successful trial, while in turn, the students maintain the positive behavior over time, then the treatment integrity is high and the intervention achieves the expected change in the dependent variable. In a second scenario, when a student performs the undesirable behavior of touching his peers every time he walks through the class and the teacher reprimands the student with harsh, embarrassing words, the student may change his behavior, but it was not due to the successful implementation of the intended intervention. Because the teacher changed the treatment to a more punitive, yet successful tact, the treatment integrity would be low, despite the fact that the student achieved the desired behavior change. In a third scenario, the teacher may have implemented the token-based intervention as intended, but the inherent problem with the student was not a desire to touch peers, but rather, the student was seeking attention from the adults in the room, so the increase in attention to the behavior maintained the undesired behavior. This is an example of misunderstanding the root cause of the problem, thus implementing the wrong treatment (independent variable) to generate change in the dependent variable. The final scenario would be seen if the teacher neglects to implement the reinforcing token system when the student performs the desired behavior, thus over time, the behavior slips back to being undesired. This scenario represents low integrity of the independent variable, thus no change in the dependent variable.

Treatment integrity can be influenced by several factors including the environment, the expertise and motivation of the interventionist, complexity of the intervention (e.g., number of steps in an intervention), professional learning and planning surrounding the intervention, feedback given to, and self-assessment by, the interventionist, how far an interventionist drifts away from the intended intervention, and the perceived and actual effectiveness of the intervention over time. In addition, researchers can influence treatment integrity by the extent to which they accurately score observations to match the predetermined standard.

Improving Treatment Integrity

To improve treatment integrity, several steps can be taken before, during, and after implementation trials. Preemptively, educators need a clear understanding of the intervention with opportunities to gain clarity and practice. They also need time to plan and prepare for the intervention. Depending on the complexity of steps or protocols required, materials needed, and time required, educators need the opportunity to co-plan, contextualize for their setting and students, and visualize successful implementation. Working with researchers, coaches, and/or peers can help ensure that educators have the information, understanding, and peers can help ensure that educators have the information, understanding, and planning necessary to be poised for success.

During implementation, educators can be empowered through the use of a rubric or checklist on which all components of an intervention are listed. The educator can then self-assess the presence or absence of each component using a Likert-type scale ranging from 0 (*not present*) to 2 (*present*). Periodic self-assessment can help a teacher avoid lapses in treatment integrity.

Once educators have begun implementation, they benefit from ongoing feedback to ensure treatment integrity. Specific performance feedback helps an educator understand where the implementation is straying from the intended course. Specific feedback could involve a researcher, instructional coach, or peer observing the implementation and rating the presence or absence of necessary components using a rubric or checklist. Follow-up conversations can revolve around the observation data and/or analysis of student products, stating the strengths and areas for improvement and responding to questions or concerns from the participant.

In addition, researchers or coaches can support treatment integrity by modeling intended use of the implementation to educators. In turn, the educator can observe classrooms or video examples of successful implementation and use the rubric or checklist to practice rating treatment integrity levels.

Future Directions

The importance of treatment integrity in the implementation of research or evidence-based practices in education has direct consequences for educators and students. Fidelity to the independent variable helps ensure that the students are experiencing the best dosage for their learning needs and that the teacher is getting the best understanding possible of what will or will not work for individual learners.

The importance of ensuring and reporting treatment integrity in social science research is vital not only to maintaining credibility of individual research studies and implementation projects but also for ensuring that consumers of the research are receiving reliable recommendations for generalizing practices and strategies to other educational contexts. It also helps strengthen the quality and credibility of social science research, as a whole, which has direct or indirect influence on law and policy decisions.

Amber Rowland

See also Experimental Designs; Feasibility; Validity

Further Readings

Goense, P. B., Boendermaker, L., & van Yperen, T. (2016). Support systems for treatment integrity. Research on Social Work Practice, 26(1), 69–73.

Gresham, F. M. (2009). Evolution of the treatment integrity concept: Current status and future directions. School Psychology Review, 38(4), 533–540.

Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S., & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980–1990. School Psychology Review, 22, 254–272.

Jenkins, S. R., Hirst, J. M., & DiGennaro Reed, F. D. (2015). The effects of discrete-trial training commission errors on learner outcomes: An extension. Journal of Behavioral Education, 24(2), 196–209.

Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential—but often forgotten—component of school-based interventions. Preventing School Failure, 48(3), 36–43.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the *Journal of Applied Behavior Analysis*: 1991–2005. Journal of Applied Behavior Analysis, 40, 659–672.

Noell, G. H. (2008). Research examining the relationships among consultation process, treatment integrity, and outcomes. In W. Erchul & S. Sheridan (Eds.), Handbook of research in school consultation: Empirical foundations for the field (pp. 315–334). Mahwah, NJ: Erlbaum.

Sanetti, L. M., & Kratochwill, T. R. (2009). Treatment integrity assessment in the schools: An evaluation of the treatment integrity planning protocol. School Psychology Quarterly, 24(1), 24–35.

Ina V. S. Mullis Ina V. S. Mullis Mullis, Ina V. S.

Michael O. Martin Michael O. Martin Martin, Michael O.

Trends in International Mathematics and Science Study Trends in international mathematics and science study

1714

1717

Trends in International Mathematics and Science Study

Trends in International Mathematics and Science Study (TIMSS) is an international assessment of mathematics and science at the fourth and eighth grades that has been conducted every 4 years since 1995. In 2015, the International Association for the Evaluation of Educational Achievement (IEA) and IEA’s TIMSS & PIRLS International Study Center at Boston College conducted TIMSS 2015 at fourth and eighth grades and TIMSS Advanced 2015 for students in the final year of secondary school enrolled in special STEM programs or tracks. Both TIMSS 2015 and TIMSS Advanced 2015 provided 20-year trend measures for countries that participated in the first TIMSS assessments in 1995.

TIMSS 2015 and TIMSS Advanced 2015 continue the long history of international assessments in mathematics and science conducted by IEA. IEA is an independent international cooperative of national research institution and government agencies that has been conducting studies of cross-national achievement since 1959. IEA pioneered international comparative assessments of educational assessments in the 1960s to gain a deeper understanding of the effects of policies across countries’ different systems of education. IEA has a Secretariat headquartered in Amsterdam and a Data Processing Center in Hamburg.

To work with the international team and coordinate within-country activities, each participating country designates an individual to be the TIMSS National

Research Coordinator (NRC). The NRCs have the challenging task of implementing TIMSS in their countries in accordance with the TIMSS guidelines and procedures. In addition, the NRCs provide feedback and contributions throughout the development of the TIMSS assessment. The quality of the TIMSS assessment and data depends on the work of the NRCs and their colleagues in carrying out the complex sampling, data collection, and scoring tasks involved.

This entry discusses the TIMSS 2015 and TIMSS Advanced 2015, including the countries that participated, quality assurance, and results, and then looks ahead to the next assessment in the TIMSS series: TIMSS 2019.

TIMSS 2015

The TIMSS 2015 mathematics and science assessments are based on comprehensive frameworks developed collaboratively with the participating countries. For each curriculum area at each grade, the frameworks are organized around two dimensions: a content dimension specifying the content to be assessed and a cognitive dimension specifying the thinking processes to be assessed. The TIMSS assessments contain nearly 800 assessment items, about 200 per grade for each curriculum area. The majority of TIMSS items assess students’ applying and reasoning skills.

New for TIMSS 2015, a home questionnaire was completed by fourth-grade students’ parents or caregivers, in addition to the questionnaires routinely given at both fourth and eighth grades to students, teachers, school principals, and curriculum specialists. The questionnaire data primarily are reported in the form of indices created using item response theory scaling methods, and results are presented for three regions of the scales (most to least desirable). When possible, the scales were developed in parallel to provide comparisons between mathematics and science as well as between the fourth and eighth grades.

TIMSS has the goal of helping countries make informed decisions about how to improve teaching and learning in mathematics and science. With its strong curricular focus and emphasis on policy relevant information about the home, school, and classroom contexts for learning, TIMSS is a valuable tool that countries can use to evaluate achievement goals and standards and monitor students’ achievement trends in an international context. The *TIMSS 2015 Encyclopedia* complements the quantitative information in the international reports with a chapter on each country summarizing mathematics and science curricula, instructional practices, and teacher education requirements.

Countries Participating in TIMSS 2015

In 2015, 57 countries, including some distinct educational systems within countries that have always participated separately throughout IEA’s long history (e.g., the Dutch-Speaking part of Belgium and Hong Kong Special Administrative Region [SAR] of the People’s Republic of China), and seven benchmarking entities (regional jurisdictions of countries such as states or provinces) participated in TIMSS. In total, more than 580,000 students participated in TIMSS 2015.

Countries and benchmarking participants could elect to participate in the fourth-grade assessment, the eighth-grade assessment, or both. Also, countries where students were expected to find the TIMSS assessments too difficult at the fourth grade could participate in the newly developed TIMSS numeracy assessment, a less difficult version of the fourth-grade mathematics assessment. Fifty countries and the seven benchmarking participants administered the fourth-grade assessments. Of those, seven countries and one benchmarking entity participated in the numeracy assessment: Bahrain, Indonesia, Iran, Kuwait, Jordan, Morocco, South Africa, and Buenos Aires. They gave both the fourth-grade assessments in mathematics and science as well as the numeracy assessment, except Jordan and South Africa, which participated in the numeracy assessment only. Thirty-eight countries and the seven benchmarking participants administered the eighth-grade mathematics and science assessments. Norway chose to assess fifth and ninth grades to obtain better comparisons with Sweden and Finland (but also collected benchmark data at fourth and eighth grades). Botswana and South Africa assessed ninth grade to better match their curricula and to maintain trend measurement.

In each grade, nationally representative samples of approximately 4,000 students from 150 to 200 schools participated in TIMSS 2015. Including the mathematics, numeracy, and science assessments and questionnaires, more than 312,000 students, 250,000 parents, 20,000 teachers, and 10,000 schools participated in the fourth-grade assessments and a further 270,000 students, 31,000 teachers, and 8,000 schools in the eighth-grade assessments.

TIMSS Advanced 2015

TIMSS Advanced 2015

With the current emphasis on college and career readiness and increasing global competitiveness in STEM fields, in 2015, TIMSS Advanced once again was joined with TIMSS. First conducted in 1995 and then again in 2008, TIMSS Advanced is the only international assessment that provides essential information about students' achievement in advanced mathematics and physics. It assesses students in their final year of secondary school (often 12th grade) who are engaged in advanced mathematics and physics studies that prepare them to enter STEM programs in higher education.

TIMSS Advanced 2015 was offered together with TIMSS to provide 20 years of trends at three important points in students' schooling (fourth grade, eighth grade, and final grade) and provide information on how the foundations established in primary school can influence students' educational career through lower secondary and impact achievement in students' final year of secondary school.

Quality Assurance

TIMSS 2015 aimed to attend to the quality and comparability of the data through careful planning and documentation, cooperation among participating countries, standardized procedures, and rigorous attention to quality control throughout the assessment process. The assessments were administered to nationally representative and well-documented probability samples of students in each country. Staff from Statistics Canada and the IEA Data Processing and Research Center worked with the NRCs on all phases of sampling activities to ensure compliance with sampling and participation requirements, with the few exceptions from compliance annotated in the data exhibits. The IEA Secretariat worked with the TIMSS & PIRLS International Study Center to manage an extensive series of verification checks to ensure the comparability of translations of the assessment items and questionnaires and to conduct an international quality assurance program of school visits to monitor and report on the administration of the assessment. The IEA Data Processing and Research Center staff and the NRCs also collaborated to organize data collection operations and to check all data for accuracy and consistency within and across countries.

TIMSS 2015 Results

The TIMSS 2015 results are presented separately for mathematics and science and within each curricular area separately for fourth grade and eighth grade. Essentially, there are four reports—mathematics at fourth grade and eighth grade as well as science at fourth grade and eighth grade. Each of the four reports contains 10 chapters providing overviews in the form of infographics and numerous exhibits summarizing students' achievement distributions, performance at the TIMSS International Benchmarks, achievement trends over time, and achievement in relation to students' home, school, and classroom educational contexts for learning mathematics and science.

The international results for TIMSS 2015 and TIMSS Advanced 2015 can be accessed at the TIMSS & PIRLS website. This website includes links to the following:

- *TIMSS 2015 assessment frameworks* presents the mathematics and science assessment frameworks that detail the major content and cognitive domains to be assessed at the fourth and eighth grades, the types of learning situations and factors that will be investigated via the questionnaire data, and an overview of the assessment design.
- *TIMSS 2015 encyclopedia: Education policy and curriculum in mathematics and science* describes national contexts for mathematics and science teaching and learning. It contains selected data about the countries' curricula together with a chapter on each participating country, summarizing the structure of its education system, the mathematics and science curricula and instruction in primary and secondary grades, the teacher education requirements, and the types of examinations and assessments employed.
- *Methods and procedures in TIMSS 2015* describes the methods and procedures used to develop, implement, and analyze the results from the TIMSS 2015 assessments.

TIMSS 2019

Marking 24 years of trend data since 1995, the seventh TIMSS assessment will be in 2019 and will include more than 60 countries. To keep up to date and relevant, TIMSS evolves with each assessment cycle. For 2019, TIMSS is converting to a digital format (eTIMSS); however, the assessments also will still be available in paper format.

eTIMSS will continue all the benefits of TIMSS, enabling countries to measure how effective they are in teaching mathematics and science. eTIMSS will use a tablet or computer format but also maintain continuity with previous TIMSS assessments to preserve trend measurement.

Newly created assessment items comprise 40% of each TIMSS cycle. The items newly developed for eTIMSS 2019 will assess areas of the TIMSS frameworks that have been difficult to measure using the traditional paper-and-pencil approach. The tasks call for applying and integrating content knowledge and cognitive capabilities in problem situations that simulate real-world contexts and laboratory experiments. In particular, problem-solving and inquiry tasks require students to solve a problem or follow a scientific line of inquiry.

For example, in mathematics, fourth-grade students can interact with geometric shapes and patterns to demonstrate their understanding of fractions and symmetry or work with a robot to examine relationships and functions. At the eighth grade, students help design a storage building by calculating dimension and areas or explore how to maximize profit for a clothing store. In science, fourth-grade students can solve a mystery based on classification of animals or investigate magnetic properties while assembling a toy train. Eighth-grade students can conduct an inquiry about why a ship sank or plan a plant growth experiment and see the results.

The eTIMSS digital format aims to improve measurement by including complex tasks and tracking the paths that students use in working out their solutions. Also, eTIMSS will provide interactive tasks that are colorful, animated, and dynamic, delivering an assessment experience that can motivate students. Finally, the eTIMSS systems will increase operational efficiency for item development, translation and translation verification, and data entry and scoring, while reducing printing and shipping costs.

eTIMSS will consist of a series of interconnected software modules hosted on the IEA servers at the Data Processing Center to assist NRCs in developing and conducting assessments. For example, NRCs can use the item builder module to develop assessment items, the eTIMSS player module to administer the assessment, the online data monitor module to observe the progress of data collection, and the online scoring system module to score students’ written responses according to the eTIMSS scoring guides. Once the data collection is complete, the data will be sent to the TIMSS & PIRLS International Center at Boston College for review, analysis, scaling, and reporting.

Boston College for review, analysis, scaling, and reporting.

Ina V. S. Mullis and Michael O. Martin

See also National Assessment of Educational Progress; Organisation for Economic Cooperation and Development; Programme for International Student Assessment; Progress in International Reading Literacy Study

Further Readings

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2017). Diagnosing competency mastery in science: An application of GDM to TIMSS 2011 data. Applied Measurement in Education, 30, 27–38.

Lim, H., & Sireci, S. G. (2017). Linking TIMSS and NAEP assessments to evaluate international trends in achievement. Education Policy Analysis Archives, 25, 11.

Mullis, I. V., Martin, M. O., & Hooper, M. (2017). Measuring changing educational contexts in a changing world: Evolution of the TIMSS and PIRLS questionnaires. In Cognitive abilities and educational outcomes (pp. 207–222). Springer International Publishing.

Websites

TIMSS & PIRLS International Center at Boston College: timss.bc.edu

Judith L. Green Judith L. Green Green, Judith L.

Monaliza M. Chian Monaliza M. Chian Chian, Monaliza M.

Triangulation

Triangulation

1717

1720

Triangulation

Since the late 1950s, educational researchers within and across different programs of research have developed strategies for exploring how and in what ways their findings for particular social phenomena are convergent, divergent, conflicting, or null through a process referred to as *triangulation*. Guided by their particular logic of inquiry, researchers across traditions engage in triangulation to make conceptually driven decisions about how to design, collect, analyze, interpret, and warrant claims about social, cultural, linguistic, psychological, and academic phenomena in education and other settings. In this entry, two telling cases are presented to make visible how triangulation, as a logic of inquiry, has been conceptualized by researchers within ongoing programs of research that differ in their goals, purposes, and theoretical groundings: multitrait/multimethod research processes and ethnographic and field-based qualitative research processes. These two telling cases are designed to make visible the ways in which data, theories, records, perspectives, methods, and/or levels of analytic scale are triangulated in the conduct of particular studies in different programs of research in education.

Telling Case 1: Multimethod and Multimeasure Research

In 1959, Donald T. Campbell and Donald W. Fiske introduced the concept of triangulation as critical for validating variables defined as constitutive of psychological traits of individuals through methods including pencil-and-paper tests, observations, and/or performance measures. They argued that researchers could confirm and/or disconfirm assumptions about the reality or validity of the phenomena being assessed by using a multitrait or multimethod approach that they called *triangulation*. Since that time, this argument has been expanded to include multimethod and multimeasure approaches to assessing or measuring educational traits or phenomena. Triangulation is undertaken to ensure that the result of the study is not dependent on characteristics of a single measure or of a measurement method.

Triangulation of constructs and/or traits is undertaken by constructing a statistical matrix consisting of a table of correlations in which the relationship within and across variables or constructs by methods is examined. This table provides a basis for facilitating and/or assessing the interpretation of convergent and discriminant validity of actions, which are assumed to reflect the traits or phenomena being assessed or measured. This process focuses on construct validity, by confirming the degree to which two measures of constructs that theoretically should be related are in fact related (convergent validity). Discriminant validity provides a basis for confirming that a particular test of a concept is not highly correlated with other tests designed to measure theoretically different concepts; that is, the two measures are unrelated.

In 2012, Robert Coe provided a summary of multiple forms of triangulation that are used to assess a broad range of quantitative forms of validity, including internal validity (causal relationship definitions), external validity (population and ecological), construct validity (causal), measurement forms of validity (e.g., face, content, criterion related, predictive, concurrent, and systemic), and construct validity (measurement—convergent, divergent, and factorial). In this program of research, triangulation is undertaken to validate constructs assessed by measurement instruments as well as the reliability of particular measurements and to construct warrants to confirm or disconfirm the validity of particular measures, instruments, evaluation processes, or relationships among variables. The logic in use used by the researcher to construct the claims based on this triangulation process provides a level of transparency for assessing the warrants of the interpretations and the conclusions drawn.

Telling Case 2: Ethnographic and Field-Based Qualitative Research

Telling Case 2 focuses on the role and nature of triangulation central to ethnographic and field-based educational research traditions grounded in advances since the 1960s in the social sciences. These traditions are influenced by philosophical turns (e.g., social, linguistic, and interactional) guiding conceptualizations of the nature of social reality. Central to these turns are conceptual arguments about the social construction of reality; that is, the ways that members of particular social groups, in particular social spaces, interactionally formulate and construct common knowledge, norms and expectations, roles and relationships, social identities, power relationships, and rights and obligations, among other social constructions that define what counts as members' knowledge and actions in the everyday life in particular social groups.

Triangulation as a logic of inquiry within field-based and ethnographic research is undertaken within and across times and events, through a range of collection and analysis processes and methods: formal and informal interviewing, participant observation, artifact collection, video and audio recording, social and geographic mapping, and searches of archival/historical records. Such field methods are grounded in particular theoretical perspectives (e.g., anthropological theories of culture; sociological theories of social order and social accomplishment of everyday life; and linguistic/sociolinguistic/discourse theories of communication). These multiple collection and analysis processes are designed to minimize limits to certainty that what is observed and recorded is the phenomenon as experienced by members of the social group in classrooms and other social settings.

Martyn Hammersley and Paul Atkinson, building on initial arguments about what constitutes triangulation proposed by sociologist Norman Denzin, conceptualized for education research the following forms of triangulation and actions or foci that today continue to guide ethnographic and qualitative field-based research:

- *Data triangulation* involves time, space, and persons
- *Investigator triangulation* involves multiple researchers in an investigation
- *Theory triangulation* involves using more than one theoretical scheme in the interpretation of the phenomenon
- *Methodological triangulation* involves using more than one method to gather data, such as interviews, observations, questionnaires, and documents.

From this perspective, triangulation is an ongoing and complex process that seeks to confirm the warrants or claims about particular phenomena studied. It also seeks to confirm that the phenomenon recorded or observed by different investigators is the same phenomenon and to explore what the difference in observations makes in terms of what can be known through the particular observation and analyses. When observers do not agree, or record different phenomena, or wonder what is happening, as anthropologist Michael Agar argues, rich points (anchors) are constructed. Such rich points support investigation of the roots and pathways that led to the observed differences or point of challenge for the researcher.

In field-based and ethnographic studies, triangulation is part of an ongoing logic in use throughout a study in order to build warrants for the accounts of how the researcher’s decisions led to empirical evidence of the constitutive ways in which knowledge is socially constructed across times and events collectively and individually by participants. By triangulating theories, methods, data sources, and investigator actions and observations, the field-based/ethnographic researcher lays a foundation for uncovering unanticipated findings about the social construction of life in particular educational and social settings. Given the complex and multifaceted nature of triangulation processes, the field-based researcher not only reports the outcomes of these processes but also includes the basis of each form of triangulation and its relationship to the developing account being constructed. By tracing actors (individually and collectively) across times, events, and disciplinary areas within an educational context, the researcher makes transparent the ongoing process of triangulating theories, methods, data, investigator observations, and analysis processes and the iterative, recursive, and abductive nature of field-based research.

The ongoing triangulation processes enable the field-based researcher/ethnographer to identify the boundaries of units, the relationships among units of analysis, and the chains of actions and reasoning necessary to construct warranted accounts of the educational and social phenomena under study. By reporting the decision-making processes, the field-based researcher makes transparent the empirical basis of the ways that triangulation supports the construction of warranted accounts of such phenomena as the construction of multiple social identities, academic processes, and epistemic knowledge within and across disciplines (e.g., science, engineering and mathematics, literacy, history, and the arts), among other educational phenomena.

Judith L. Green and Monaliza M. Chian

See also Convergence; Discriminant Function Analysis; Ethnography; Qualitative Data Analysis; Validity

Further Readings

Agar, M. (2006). Culture: Can you take it anywhere? International Journal of Qualitative Methods, 5(2), 1–16.

American Education Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. Educational Researcher, 35(6), 33–40. Retrieved from http://www.aera.net/Publications/Standards-for-Research-Conduct

Bloome, D., & Egan-Robertson, A. (1993). The social construction of intertextuality in classroom reading and writing lessons. Reading Research Quarterly, 28(4), 305–333.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105.

Coe, R. (2012). Conducting your research. In J. Arthur, W. Waring, R. Coe, & L. V. Hedges (Eds.), Research methods & methodologies in education. (pp. 41–52). London, UK: Sage.

Gee, J. P., & Green, J. L. (1998). Discourse analysis, learning, and social practice: A methodological study. Review of research in education, 23, 119–169.

Hammersley, M., & Atkinson, P. (1991). Ethnography: Principles in practice (3rd ed.). New York, NY: Routledge.

Heap, J. L. (1995). The status of claims in “qualitative” research. Curriculum

Inquiry, 25(3), 271–292.

Krathwohl, D. R. (1993). Methods of educational and social science research: An integrated approach. London, UK: Longmans.

Mitchell, C. J. (1984). Typicality and the case study. In P. F. Ellen (Ed), Ethnographic research: A guide to general conduct (pp. 238–241). New York, NY: Academic.

Patton, M. Q. (2002). Qualitative research and evaluation methods. Thousand Oaks, CA: Sage.

Strike, K. A. (1989). Liberal justice and the Marxist critique of education: A study of conflicting research programs. New York, NY: Routledge.

Walford, G. (Ed.). (2008). How to do educational ethnography. London, UK: Tufnell.

Robert J. Sternberg Robert J. Sternberg Sternberg, Robert J.

Triarchic Theory of Intelligence

Triarchic theory of intelligence

1720

1723

Triarchic Theory of Intelligence

The triarchic theory of (successful) intelligence explains in an integrative way the relationship between intelligence and (a) the internal world of the individual, or the mental mechanisms that underlie intelligent behavior; (b) experience, or the mediating role of the individuals' passage through life between their internal and external worlds; and (c) the external world of the individual, or the use of these mental mechanisms in everyday life in order to attain an intelligent fit to the environment. The theory has three subtheories, one corresponding to each of the three relationships mentioned in the preceding sentence.

Definition of Successful Intelligence

According to the proposed theory, *successful intelligence* is the use of an integrated set of skills needed to attain success in life; however, individuals define it within their sociocultural context. People are successfully intelligent by virtue of recognizing their strengths and making the most of them, at the same time that they recognize their weaknesses and find ways to correct or compensate for them. Successfully intelligent people adapt to, shape, and select environments through finding a balance in their use of analytical, creative, practical, and wisdom-based skills. This section considers each element of the theory in turn.

According to the first element, there is no one definition of success that works for everyone. Education should be geared toward the goals of each individual rather than toward one predefined goal that may be relevant to some students but not to many others.

The second element asserts that there are different paths to success, no matter what goal one chooses. For most of us, there are at least a few things we do well, and our successful intelligence is dependent in large part upon making these things “work for us.” At the same time, we need to acknowledge our weaknesses and find ways either to improve upon them or to compensate for them.

The third element asserts that success in life is achieved through some balance of adapting to existing environments, shaping those environments, and selecting new environments. There may be times when our attempts to adapt and to shape the environment lead us nowhere—when we simply cannot find a way to make the environment work for us. In these cases, we leave the old environment and select a new environment. Sometimes the smart thing is to know when to get out.

Finally, we balance three kinds of skills in order to achieve these ends: analytical skills, creative skills, practical skills, and, in the augmented version of the theory, wisdom-based skills. We need creative skills to generate ideas, analytical skills to determine whether they are good ideas, practical skills to implement the ideas and to convince others of the value of our ideas, and wisdom-based skills to help achieve a common good that goes beyond just our own self-interest.

Most people who are successfully intelligent are not equally endowed with these diverse skills, but they find ways of making the three skills work harmoniously together. People exercise their analytical skills when they analyze, compare and contrast, judge, critique, and evaluate. People exercise their creative skills when they create, invent, discover, design, imagine, and suppose. People exercise their practical skills when they put into practice, apply, utilize, implement, and persuade. People exercise their wisdom-based skills when they utilize their knowledge and their other skills to serve a common good, by balancing their own with others’ and higher order interests over the long and short terms, through the infusion of positive ethical values.

Traditional teacher-made as well as standardized tests, in assessing intelligence, tend to focus on analytical skills as well as knowledge base. Unfortunately, these tests tend to ignore the other skills that are important to intelligence, namely, the creative, practical, and wisdom-based skills.

Intelligence and the Internal World of the Individual

In the triarchic theory of successful intelligence, there are three basic kinds of information-processing components, referred to as *metacomponents*, *performance components,* and *knowledge acquisition components*.

Metacomponents

Metacomponents are higher order, executive processes used to plan what one is going to do, to monitor it while one is doing it, and evaluate it after it is done. These metacomponents include recognizing the existence of a problem, deciding on the nature of the problem confronting one, selecting a set of lower order processes to solve the problem, selecting a strategy into which to combine these components, selecting a mental representation on which the components and strategy can act, allocating one's mental resources, monitoring one's problem solving as it is happening, and evaluating one's problem solving after it is done. Let us consider some examples of these higher order processes.

Performance Components

Performance components are lower order processes that execute the instructions of the metacomponents. These lower order components solve the problems according to the plans laid out by the metacomponents. Although the number of metacomponents used in the performance of various tasks is relatively limited, the number of performance components is probably quite large. Many of these performance components are relatively specific to narrow ranges of tasks. Examples of performance components are inferring the relations between two elements and applying that relation to another element.

Knowledge Acquisition Components

Knowledge acquisition components are used to learn how to do what the metacomponents and performance components eventually do. Three knowledge acquisition components appear to be central in intellectual functioning: selective encoding, selective combination, and selective comparison. Selective encoding involves sifting out relevant from irrelevant information. Selective combination involves combining selectively encoded information in such a way as to form an integrated, plausible whole. Selective comparison involves discovering a nonobvious relationship between new information and already acquired information.

Intelligence and Experience

Components of information processing are always applied to tasks and situations with which one has some level of prior experience (even if it is minimal experience). Hence, these internal mechanisms are closely tied to one’s experience. According to the experiential subtheory, the components are not equally good measures of intelligence at all levels of experience. Assessing intelligence requires one to consider not only components but also the level of experience at which they are applied.

According to the experiential subtheory, intelligence is best measured at those regions of the experiential continuum involving tasks or situations that are either relatively novel on the one hand or in the process of becoming automatized on the other. Totally, novel tasks and situations provide poor measures of intelligence: They just would not make sense to people, as when one gives calculus problems to 10-year-olds.

Ability to Deal With Novelty

Intelligent people can deal well with relatively novel tasks and situations. Confronted with a situation that they have never seen before, such as living for the first time in a foreign country, they can adapt despite the differences in the environment from what they are used to.

Ability to Automatize Information Processing

Automatization occurs for many tasks, such as driving a car or reading. Initially, one has to concentrate exclusively on the road when driving. Eventually, one can concentrate on the road, listen to music, and carry on a conversation. Driving has become automatic.

The ability to deal with novelty and the ability to automatize information processing are interrelated, as shown in the example of the automatization of reading described in this section. If one is well able to automatize, one has more resources left over for dealing with novelty. Similarly, if one is well able to deal with novelty, one has more resources left over for automatization. Thus, performances at the various levels of the experiential continuum are related to one another.

Intelligence and the External World of the Individual

According to the contextual subtheory, intelligent thought is directed toward one or more of three behavioral goals: *adaptation to an environment*, *shaping of an environment*, or *selection of an environment*. These three goals may be viewed as the functions toward which intelligence is directed.

Adaptation

Most intelligent thought is directed toward the attempt to adapt to one’s environment.

Different contextual milieus may result in the development of different mental abilities. For example, Puluwat navigators must develop their adaptive large-scale spatial abilities for dealing with cognitive maps to a degree that far exceeds the adaptive requirements of contemporary Western societies. Similarly, Australian Aboriginal children probably develop their visual–spatial memories to a greater degree than do Australian children of European descent. The latter are more likely to apply verbal strategies to spatial memory tasks than are the Aboriginal children, who employ spatial strategies. This greater development is presumed to be due to the greater need the Aboriginal children have for using spatial skills in their everyday lives. In contrast, members of Western societies probably develop their abilities for thinking abstractly to a greater degree than do members of societies in which concepts are rarely dealt with outside their concrete manifestations in the objects of the everyday environment.

Shaping

Shaping of the environment is often used as a backup strategy when adaptation fails. If one is unable to change oneself to fit the environment, one may attempt to change the environment to fit oneself. For example, repeated attempts to adjust to the demands of one’s romantic partner may eventually lead to attempts to get the partner to adjust to oneself. But shaping is not always used in lieu of adaptation. In some cases, shaping may be used before adaptation is ever tried, as in the case of the individual who attempts to shape a romantic partner with little or no effort to shape himself or herself so as to suit the partner’s wants or needs better.

Selection

Selection involves renunciation of one environment in favor of another. In terms of the rough hierarchy established so far, selection is sometimes used when both adaptation and shaping fail. Sometimes one attempts to shape an environment only after attempts to leave it have failed. Other times, one may decide almost instantly that an environment is simply wrong and feel that one need not or should not even try to fit into or to change it. For example, every now and then new graduate students may realize almost immediately that they came to graduate school for the wrong reasons or who find that graduate school is nothing at all like the continuation of undergraduate school they expected. In such cases, the intelligent thing to do may be to leave the environment as soon as possible, to pursue activities more in line with one's goals in life.

Final Thoughts

The triarchic theory of successful intelligence is a way of understanding intelligence that views intelligence in a broader way than do conventional theories of intelligence. The theory emphasizes the role of individuals in defining what is important to them in life and to achieve their own personal goals. Individuals achieve these goals by capitalizing on strengths and compensating for or correcting weaknesses, in order to adapt to, shape, and select environments. They operate on these environments through a combination of analytical, creative, practical, and wisdom-based skills.

Robert J. Sternberg

See also Cognitive Development, Theory of; Critical Thinking; Intelligence Tests

Further Readings

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. New York, NY: Cambridge University Press.

Sternberg, R. J. (1997). Successful intelligence. New York, NY: Plume.

Sternberg, R. J. (2003). Wisdom, intelligence, and creativity synthesized. New York, NY: Cambridge University Press.

Samantha B. Goldstein Samantha B. Goldstein Goldstein, Samantha B.

Marc H. Bornstein Marc H. Bornstein Bornstein, Marc H.

Triple-Blind Studies Triple-Blind studies

Triple-Blind Studies

A *triple-blind study* is one in which participants, researchers, and analysts are unaware of whether the participant received the treatment or the placebo in a random assignment trial. Although a double-blind design leaves only the participants and researchers unaware of the treatment assignment, a triple-blind study additionally keeps the team analyzing the data from knowing which group’s data—placebo or treatment—it is evaluating. This design allows objectivity in the data collection phase and in the data analysis phase—analysts are able to evaluate data without any bias.

Application

Barbara Henker and colleagues compared double-and triple-blind designs to examine a medication for hyperactivity in boys. They randomly assigned boys to one of two groups: the treatment group, which received the medication, and the placebo group, which received a placebo. They also used blind and nonblind raters/analysts. Results showed positive effects of the treatment, meaning the boys who received the medication displayed a significantly larger change in behavior than the boys in the placebo group. The double-blind design allowed the researchers to conclude that the medication itself was effective, as participants’ ignorance to their assignment ruled out any behavior that could be attributed to knowledge of receiving the medication.

The results of Henker and colleagues’ study also indicated that there was no difference in ratings between the double-blind raters and the triple-blind raters. Using a triple-blind design gave legitimacy to the results of the double-blind raters. The conclusions drawn from the double-blind design remained because even raters uninformed of group identity evaluated significant differences in behavior change between the treatment and placebo groups. A research team could also use this design to explore the legitimacy of school interventions for children with various learning disabilities.

Implications

Blinding in a study is important for its credibility, and each level of blinding adds more objectivity. Single blinding allows the researchers to make assumptions about the effect of a treatment without possible bias from the participants' knowledge of their treatment status. Double-blind studies prevent any bias in the researchers' behavior, causing confounding effects in the participants. Finally, triple-blind studies provide an even higher level of objectivity. They keep the participants' treatment status unknown all the way through to the data analysis phase. Triple-blind studies add a degree of credibility to a study that is not often achieved with other designs.

Samantha B. Goldstein and Marc H. Bornstein

See also Double-Blind Design; Experimental Designs; Placebo Effect; Random Assignment

Further Readings

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (2010). Blindness. In L. M. Friedman, C. D. Furberg, & D. L. DeMets (Eds.), Fundamentals of clinical trials (pp. 119–132). New York, NY: Springer.

Henker, B., Whalen, C. K., & Collins, B. E. (1979). Double-blind and triple-blind assessments of medication and placebo responses in hyperactive children. Journal of Abnormal Child Psychology, 7(1), 1–13.

Pocock, S. J. (1983). Clinical trials: A practical approach. Chichester, UK: Wiley.

Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomized trials: Hiding who got what. Lancet, 359(9307), 696–700.

Robert L. Brennan Robert L. Brennan Brennan, Robert L.

Won-Chan Lee Won-Chan Lee Lee, Won-Chan

True Score

True score

1724

1728

True Score

In conventional discourse, *true score* almost always has the Platonic connotation of “in the eye of God” truth. That is, there is no acknowledgment of the possibility of error of any type. This notion of true score may have some philosophical value, but it has no scientific utility. All measurements (i.e., scores) in scientific disciplines are *observed* under certain *conditions of measurement*, with the implicit acknowledgment that such measurements can differ under other conditions. There are two broad classes of perspectives on true score: expected-value perspectives and model-trait perspectives. The expected-value perspectives include classical test theory (CTT) and generalizability (G) theory, both of which view true score as the expected value of observed scores over replications of a measurement procedure. The only model-trait perspective considered here is item response theory (IRT), in which a person parameter θ plays a role similar to that of true score.

CTT

CTT asserts that observed scores for a person X_p can be split into two parts: a true score (T_p) that is specific to the person and error scores (E_p):

$$X_p = T_p + E_p.$$

Although this equation is algebraically simple, it is complicated by the fact that neither the person’s true score nor the errors are directly observable. Indeed, T_p is undefined. Typically, this problem is circumvented by assuming that the expected value (E) of the errors is 0; that is,

$$\boldsymbol{E}(E_P) = 0.$$

Under this assumption,

$$\begin{aligned}\boldsymbol{E}(X_P) &= \boldsymbol{E}(T_P + E_P) = \boldsymbol{E}(T_P) + \boldsymbol{E}(E_P)\\ &= \boldsymbol{E}(T_P) = T_P,\end{aligned}$$

because $\boldsymbol{T}_p$ is a constant specific to the person. Although it is common practice to refer to a *definition* of true score, the aforementioned development shows that in CTT, true score is typically better viewed as a quantity *derived* from the model in Equation 1 and the assumption in Equation 2.

Another central assumption in CTT is that true scores and error scores are uncorrelated,

$$\sigma_{TE} = 0.$$

Under this assumption, the variance (over persons) of observed scores is simply:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

True scores play a crucial role in reliability, which is defined canonically as the squared correlation between observed and true scores, . Given the assumption that $\sigma_{TE} = 0$, it is easy to show that , where the last formula is the most frequently used basis for estimating reliability.

Returning to Equation 3, there is one very important unanswered question, that is, what constitutes the replications over which the expectations in Equation 3 are taken? In CTT, the traditional answer to this question is expectations are taken over forms of a test that have equal *observed* score means, variances, and covariances—called classically parallel forms. This is only one of many possible answers, however, and each different definition of replications can (and usually does) lead to different results. This is the principal reason why there are so many different formulas for estimating quantities such as reliability.

It is particularly important to note that there is no “right” or “best” definition of replications. It follows that there is no universally right or best characterization of true score. Rather, an investigator must *choose* how the true score shall be viewed. So, for example, an investigator who chooses to use coefficient α to estimate reliability is assuming (knowingly or unknowingly) that replications consist of forms that are essentially τ equivalent, which is not quite the same as classically parallel.

In effect, Equation 3 states that observed scores are unbiased estimates of true scores. In most contexts, therefore, observed scores are viewed as the best estimates of true scores. Alternatively, if it is assumed that the regression of true scores on observed scores is linear (which is an assumption that is not required for most results in CTT), *regressed-score estimates* of true score (typically designated) can be obtained. Such estimates are biased and have a Bayesian interpretation, but they do not alter the rank ordering of examinees. The principal advantage of over X_p is that the standard error of estimation (based on) is smaller than the standard error of measurement (based on X_p). However, for high-scoring examinees, , and for low-scoring examinees, .

Most applications of CTT attend to means and variances, only, of observed, true, and error scores. The principal exception is *strong* true-score models that model the entire joint distributions of observed, true, and error scores. The simplest of these models is the β-binomial model in which it is assumed that true scores have a β distribution, and errors (conditional on true scores) have a binomial distribution. Under certain linearity assumptions, it follows that observed scores have a negative hypergeometric distribution. If the actual distribution of observed scores is approximately negative hypergeometric, then the model can be assumed to hold, and true scores can be assumed to be distributed as β, which is a rather flexible distribution. The advantage of a strong true-score model is that just about any parameter of interest in measurement can be estimated.

G Theory

The simplicity of CTT is both a strength and a weakness. For example, Equation 1 has only one error term, but a serious consideration of most real-life measurement procedures typically reveals that there are at least several potential sources of random error in observed measurements, all of which are undifferentiated in the *E* term in the CTT model. By contrast, using G theory, different sources of random error can be separately modeled and estimated.

Consider, for example, a reading test that consists of four passages (*T*), with 10 different items (*I*) nested within passages. We refer to passages and items as “facets.” In the notation of G theory, the model for decomposing the observed *mean* score over all 40 items is:

$$X_{pIT} = \mu + \upsilon_p + \upsilon_T + \upsilon_{I:T} + \upsilon_{pT} + \upsilon_{pI:T},$$

where the colon designates nesting and the ν terms are score effects. Importantly, is the universe (of generalization) score for person *p*, which is analogous to true score in CTT. The universe score μ_p is interpreted as the expected value of the observed scores X_{pIT} over *randomly parallel* forms. When both passages and items are assumed to be random facets, as they are here, each randomly parallel form consists of a *different* set of four passages with different sets of 10 items.

The total variance of the observed mean scores in Equation 6 is:

$$\sigma^2\left(X_{pIT}\right) = \sigma^2(p) + \sigma^2(T) + \sigma^2(I:T) + \\ \sigma^2(pT) + \sigma^2(pI:T),$$

where $\sigma^2(p)$ is universe score variance, which is the analogue of true score variance in CTT. The variance components $\sigma^2(pT)$ and $\sigma^2(pI{:}T)$ constitute relative error variance, $\sigma^2(\delta)$, which is the analogue of error variance in CTT. The remaining two variance components are necessarily zero under the assumption of classically parallel forms but not under the assumption of randomly parallel forms. Rather, $\sigma^2(T)$ and $\sigma^2(I{:}T)$ contribute to absolute error variance, $\sigma^2(\Delta)$, in G theory.

Strictly speaking, this discussion relates to univariate G theory for a random effects model. The word *univariate* implies that there is only one universe (or true) score for each person. The phrase *random effects model* excludes the possibility that one or more facets are fixed. *Fixed* refers to facets that have the same conditions for all forms. Univariate G theory can accommodate fixed facets, but doing so is sometimes awkward. By contrast, multivariate G theory considers fixed facets explicitly and, in particular, disentangles multiple true scores that might contribute to a composite true score. An example is considered next.

Expanding the previous example, suppose each form of a reading test consists of four passages: two history (*H*) passages and two science (*S*) passages, but the passages themselves differ across forms. This means that passage types (history and science) are fixed, but passages are random. The multivariate G theory representation of this example mirrors Equations 6 and 7, with observed scores, score effects, and variance components replaced by 2 × 2 matrices.

For example, the universe score variance–covariance matrix for persons is:

$$\sum_{p}=\begin{bmatrix}\sigma_H^2(p) & \sigma_{HS}(p)\\ \sigma_{HS}(p) & \sigma_S^2(p)\end{bmatrix},$$

where and are the universe score variances for history and science, respectively, and is the covariance between the two universe scores, μ_{pH} and μ_{pS}. A composite (*C*) of these two universe scores can be defined as , where w_H and w_S are weights chosen by the investigator. Composite universe score variance is:

$$\sigma_C^2(p)=W_H^2\sigma_H^2(p)+W_S^2\sigma_S^2(p)+2\sigma_{HS}(p).$$

IRT

As noted earlier, the only model-trait perspective considered here is IRT. Discussion is further restricted to the unidimensional one-, two-, and three-parameter logistic (i.e., 1PL, 2PL, and 3PL) models.

In an IRT model, conceptually the notion of true score is linked to the person ability (or proficiency) parameter θ. The 1PL, 2PL, and 3PL models differ with respect to the number of item parameters in the model. So, the most obvious difference between expected value models (i.e., CTT and G theory) and IRT models is grain size; that is, expected value models are defined with respect to forms of a test, whereas IRT models are defined with respect to items. Less obviously, the various IRT models effectively employ a different conception of true score.

The 3PL model involves difficulty (*b*), discrimination (*a*), and pseudoguessing (*c*) item parameters. The 2PL model involves *b* and *a*, only. Finally, the 1PL model involves *b* only. Philosophical disagreements in the basis for choosing among these models are so strong that no amount of data can resolve these disagreements. In a sense, the different proponents have different conceptions of what sources of variability should be modeled, which means they have different notions of how θ should be defined, because θ is defined through selection of a model. Parallelism of forms is not explicitly defined in IRT, although conceptually in IRT forms are assumed to be "strictly" parallel in the sense that they have the same item parameters.

The person parameter that is modeled in IRT is θ, which has conceptual similarities with true score, but there are still nontrivial differences between θ and true score. Perhaps the most obvious difference is that θ always has a range of $-\infty$ to $+\infty$, whereas the range of true scores is bounded. For a *k*-item multiple-choice test, in CTT, true scores typically have a range of 0–*k*, whereas in G theory, the universe scores typically have a range of 0–1 (in the mean score metric).

A test characteristic curve (TCC) provides a nonlinear transformation of θ, such that the transformed values have a range of 0–*k* for a *k*-item multiple-choice test. A TCC transformed value of θ is sometimes referred to as a true score because the transformed value is the expected value of the hypothetical distribution of observed scores given θ and the item parameters. A TCC value is conceptually closer to a true score in CTT than is θ. Still, a TCC value is based on a hypothetical distribution of observed scores, not an actually observed distribution. So, in this sense, TCC values are more closely associated with strong true-score models in CTT. It is important to remember, however, that a TCC value is model-specific. So the TCC values will differ for the 1PL, 2PL, and 3PL models.

The discussion in this entry has focused on similarities and differences between true score in CTT and true score conceptions in IRT. Comparisons with universe scores in G theory (univariate or multivariate) are much more tenuous. The principal problem is that IRT currently has no obvious way to differentiate among multiple random facets. (Some consideration has been given to an adjustment to the TCC.) Furthermore, IRT cannot distinguish between fixed and random facets, although multivariate IRT has the potential to accommodate multiple fixed facets.

Robert L. Brennan and Won-Chan Lee

See also Classical Test Theory; Generalizability Theory; Item Response Theory; Standard Error of Measurement

Further Readings

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, 38, 285–317.

Brennan, R. L. (2001b). Generalizability theory. New York, NY: Springer-Verlag.

Brennan, R. L. (2011). Generalizability theory and classical test theory. Applied Measurement in Education, 24, 1–21.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105–146). New York, NY: American Council on Education and MacMillan.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 11–154). Westport, CT: American Council on Education/Praeger.

Stella Bollmann Stella Bollmann Bollmann, Stella

True–False Items True–False items

1727

1728

True–False Items

True–false (TF) tests are tests of multiple statements with each judged to be true or false. One statement in a TF test is called a TF item. Items may be administered separately or in sets with a common stem. If they are presented in sets of, for example, four or five, they are often named *multiple true/false* or *multiple-choice* (MC) tests because MCs have to be made while any number of items may be correct. However, commonly, the expression MC is used for items with one statement and multiple alternative answers, of which only one is correct.

Concerning multiple true–false items, another distinction can be made between *loosely linked* and *strongly linked* sets. In a loosely linked set, items are grouped together more for administrative than for content-related reasons. Although they have a common stem, a simple introduction may be like “Which of the following statements concerning probability distributions is true: … .” Loosely linked items are basically independent; this kind of grouping simplifies orientation of the examinees. The stem of strongly linked items is usually more specific as, for example, “Measures of central tendency are … .” The true items within one strongly linked set all belong to the same unit of a lecture or a textbook.

TF items may also be used for tests that do not measure ability but a trait. In these tests, items are not true or false, but they measure the latent trait of a person. But mostly, they are used for the measurement of an ability where only one answer is true.

MC and TF tests have been investigated in the early 20th century, and their diagnostic quality has been discussed controversially among researchers. Still, they are immensely popular among examiners. An important advantage is their high economy in both administration and analysis. Furthermore, in comparison high economy in both administration and analysis. Furthermore, in comparison to constructed response questions, whereby examinees are asked to give a short answer, the analysis of MC or TF questions is more objective. Still, there are some problematic matters that shall be discussed below.

One broadly discussed issue is how to score TF items. Two commonly applied techniques are *number right scoring* and *formula scoring*. In number right scoring, the score of an examinee simply equals the number of the examinee's correct responses. It is the most straightforward method. Because this technique might encourage examinees who do not know the correct answer to guess, other methods have been proposed. These more complex methods can be summarized under the expression formula scoring. One possible option in formula scoring is negative marking. Here, one mark is given for a correct response and one mark is deducted for each incorrect choice. The aim is to penalize for a wrong guess. Many studies have shown that most people benefit from guessing if no penalty for guessing (like negative marking) is applied.

It is a well-known effect that MC and TF tests encourage guessing, which is known to be a huge diagnostic issue. The problem of guessing in MC and TF tests is that there is a certain probability that examinees who do not know the right answer but give a (guessed) response have a certain probability to give the right response. This probability depends on the number of statements to choose from in relation to the number of true statements. If, for example, in an MC test, only one of the four statements is true, then the probability for a true guess is 0.25 for the whole set of statements. If each of the statements may be either true or false as it is in TF tests, the probability for a true guess is 0.5 for each statement. Thus, the probability that the whole set of statements is guessed correctly is $0.5^4 = 0.0625$. As can be seen, to achieve the same reduction in probability for a correct guess, fewer items are required in TF tests compared to MC tests.

One issue in ability and in personality tests is that many studies have shown that scores may be influenced by variables other than the one intended to be measured. This means, the tests are not unidimensional anymore, which is an important property of all tests. One example is the influence of acquiescence, the tendency to agree to an item rather than to disagree. Using TF items instead of MC items diminishes the effect of response styles to some extent, but their effect cannot be entirely erased. This was shown by obtaining separate scores on the true items and on the false items of a test and correlating these two scores. The correlation was shown to be near 0. Apparently, these two forms of items (true and false) do not measure the same trait, which is caused by the effect of acquiescence. Some examinees tend to respond “true” more often than others, so that their score on true items is higher and the score on false items is lower than that of the other examinees. Also, this effect can be observed when using other expressions as *like* versus *dislike* or *agree* versus *disagree*. Examinees tend to use one more than the other, and most examinees respond “yes” more than “no.” These individual differences between examinees preferring yes and those preferring no have been shown to be reliable by split-half and parallel tests with elapsed time methods. The fact that more people tend to agree than to disagree results in the effect that false items are more valid than true items. Of course, this only occurs when an examinee is guessing. Then, the examinee will be correct more often on true items than on false items because of acquiescence. Response styles have their greatest influence in ambiguous situations on items where the person is not sure how to answer or, in the case of measuring ability, guesses. Thus, response styles may be diminished by clear instructions and explicit wordings.

Stella Bollmann

See also Matching Items; Multiple-Choice Items

Further Readings

Frey, B. B. (2013). Modern classroom assessment. Thousand Oaks, CA: Sage.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. Teaching and Teacher Education, 21(4), 357–364.

Popham, W. J., & Popham, J. W. (2005). Classroom assessment: What teachers need to know. Pearson.

Demetri L. Morgan Demetri L. Morgan Morgan, Demetri L.

Sharon M. Ravitch Sharon M. Ravitch Ravitch, Sharon M.

Trustworthiness

Trustworthiness

1728

1731

Trustworthiness

The term *trustworthiness* refers to an overarching concept used in qualitative research to convey the procedures researchers employ to ensure the quality, rigor, and credibility of a study while (re)establishing congruence of the epistemological and ontological underpinnings of the researcher with the design, implementation, and articulations of a research study. Hence, trustworthiness is both an aim and a practice. The trustworthiness section of a study typically asserts why the findings and implications can be viewed as acceptable and of worth to the reader by making the methodology and methods that undergird the study transparent. Transparency in the approach, implementation, and evaluation of a study enables consumers of the research to take important details into account when assessing the study’s value and utility. Thus, trustworthiness is relevant to educational research, measurement, and evaluation because the related procedures are a key task qualitative researchers must respond to in the execution of a research project. Despite this, the role of trustworthiness in qualitative research is an unsettled paradigmatic debate because of the concept’s overlap with the positivist notion of “validity,” the cornucopia of approaches involved when seeking trustworthiness, and the lack of standardization on how to best judge the effectiveness of trustworthiness across different fields and disciplines. This entry first overviews the epistemological and ontological roots of trustworthiness and then describes common procedures for addressing trustworthiness concerns.

Epistemological and Ontological Schism in

Qualitative Research

Evaluating the quality of scientific research has largely been rooted in an implicit assumption that there is a single “truth” or one reality that is experienced similarly by everyone. The purpose of the scientific process then is to objectively observe, measure, and report the dimensions and properties of any given phenomenon. This positivist worldview maintains that what the scientific method reveals in a given context or culture is generalizable to other contexts and cultures because there is only one reality. Accordingly, quality research in this paradigm is concerned with how valid and reliable the data, procedures, and analysis of a study are in revealing this objective truth. Validity in this interpretation is concerned with how well a study meets the established requirements of the scientific method, which have been agreed upon to be important steps in uncovering observable realities. Reliability, on the other hand, conveys quality by ensuring that the outcomes of a study are repeatable and replicable. To this day, this scientific viewpoint pervades the physical sciences; however, in the late 1970s, anthropologists, sociologists, and qualitative educational researchers began to question whether these assumptions and steps were proper approaches to research concerning multiple truths or realities that were context-specific.

The questions from this early cadre of researchers grew into a chorus of critiques of the hidden and embedded approaches to research in general and qualitative research in particular that emanated from the positivist paradigm. The belief was that these approaches were complicit in marginalizing the experiences and voices of cultures and peoples who did not have the resources, power, or space to assert their own narratives into the cannon of formalized knowledge about human beings and the social world. This new wave of thinking critiqued the prevailing positivist paradigm, injecting the notion of multiple truths and realities into the scientific method not only as an important and worthy endeavor but as an epistemological and ontological stance. In agreement, some scholars informed by critical social theory believed that the quality of research should be assessed by the political power it manifests for minoritized and oppressed peoples. Another subsection of critical scholars dismissed the need for any criteria to judge the quality of research as reductionist and denying the complexity that exists in the world. Still another group tried to reconcile these postpositivist critiques with an explicit focus on subjectivity; broad and flexible criteria for the evaluation of quality research; and sensitivity to histories, context, and the positionality of the researcher. The concept of trustworthiness emerged from the thinking and writing of scholars in the third group as a way to effectively address the epistemological and ontological concerns of research while attending to the issues of research quality.

In 1985, Yvonna Lincoln and Egon Guba are credited with establishing the first iteration of trustworthiness in qualitative research. Their initial idea was concerned with evolving the four questions that evaluators and consumers of research typically raise. Truth value, or how a researcher “can establish confidence in the truth of the findings of a particular inquiry,” was refashioned as credibility. Applicability, or how a researcher “can determine the extent to which the findings of a study have applicability in other contexts,” was reframed as transferability. Consistency, or how a researcher can determine “whether the findings of a study would be repeated if the study were replicated with similar [participants],” became dependability. Finally, the question of neutrality, or how a researcher establishes “the degree to which the findings of a study are determined by the [participants]” and not the “biases, motivations, interests, or perspectives” of the researcher, was adapted as confirmability (Lincoln & Guba, 1985, p. 290). Variations of credibility, transferability, dependability, and confirmability have subsequently become the core tenants of trustworthiness.

Common Trustworthiness Strategies

Although there are numerous procedures that qualitative researchers can employ to address the various tenets of trustworthiness, this section focuses on two of the more widely used strategies: triangulation and participant validation (also known as member reflections, member validation, respondent validation, verification, and member checks). Triangulation in qualitative research relates to trustworthiness because it is concerned with using multiple indicators throughout a research project to convey the dependability, credibility, and likely transferability of a study. The underlying philosophy of triangulation is to use multiple strategies to cancel out the weaknesses of any one method. There are four common approaches to triangulation. Researcher triangulation involves the engagement of multiple researchers in a study that brings together their unique insights during the inquiry. Data triangulation includes seeking out two or more forms of data from diverse sources to build more comprehensive interpretations of a phenomenon. Theory triangulation denotes the need to approach a research study with various frameworks, sensitizing a researcher to the contexts or dynamics that may be of relevance. Finally, methodological triangulation is typically seen as the use of both qualitative and quantitative research approaches typically seen as the use of both qualitative and quantitative research approaches. However, qualitative researchers can also employ methodological triangulation by pairing different qualitative methods together (e.g., grounded theory and case study).

Another prominent trustworthiness procedure related to credibility and confirmability is participant validation. Seeking participant validation is the systematic process of engaging the study participants with the data, findings, and/or analysis of a project both to ascertain if researchers accurately reflected their lived experiences and to garner new data that may spur richer insights, a fuller understanding of context and how it mediates experiences and events, and deeper analysis. Engaging in participant validation can occur at any point in a research project; it is one critical way to deal with, account for, and make explicit data that do not coincide with emergent themes or categories in a study. The systematic search for disconfirming evidence or what some refer to as “outliers” and for contradictions through participant validation also adds to the trustworthiness of a study.

Although there are myriad strategies available to a researcher seeking trustworthiness, it is vital to understand and respond to the unique audiences and disciplinary expectations that also exert influence on how the quality of research is judged. Consequently, while the criteria for seeking trustworthiness remain flexible, serious attention and concern should be given in the research design, implementation, and articulation phases of a research project to weave in strategies that address the credibility, transferability, dependability, and confirmability of a study.

Demetri L. Morgan and Sharon M. Ravitch

See also Ethical Issues in Educational Research; Member Check; Naturalistic Inquiry; Pilot Studies; Qualitative Data Analysis; Qualitative Research Methods; Representativeness; Triangulation; Validity

Further Readings

Bochner, A. P. (2000). Criteria against ourselves. Qualitative Inquiry, 6(2), 266–272.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Thousand Oaks, CA:

Sage.

Mathison, S. (1988). Why triangulate? Educational Researcher, 17(2), 13–17.

Milner, H. R. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. Educational Researcher, 36(7), 388–400.

Patton, M. Q. (2015). Qualitative research & evaluation methods: Integrating theory and practice (4th ed.). Thousand Oaks, CA: Sage.

Pratt, M. (2009). From the editors: For the lack of a boilerplate: Tips on writing up (and reviewing) qualitative research. Academy of Management Journal, 52(5), 856–862.

Ravitch, S. M., & Mittenfelner C. N. (2016). Qualitative research: Bridging the conceptual, theoretical, and methodological. Thousand Oaks, CA: Sage.

Seale, C. (1999). The quality of qualitative research. Thousand Oaks, CA: Sage.

Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. Qualitative Inquiry, 16(10), 837–851.

Sara A. Florence Sara A. Florence Florence, Sara A.

Jennifer Davidtz Jennifer Davidtz Davidtz, Jennifer

Twin Studies

Twin studies

1731

1733

Twin Studies

The study of twins in scientific research allows researchers to attempt to disentangle the effects of genetic and environmental effects on biology and psychology. Twin studies have provided valuable insight into detecting and treating various diseases and psychological disorders, as they reveal the importance of genetic and environmental influences on traits, phenotypes, and disorders. Twin research is a key tool in the field of behavioral genetics, and twin studies are part of the broader methodology used in behavior genetics, which uses genetically informative data to track a variety of traits ranging from personal behavior to the presentation of severe mental illness such as schizophrenia. In attempting to describe, understand, and explain how learning takes place throughout a person's life through educational research, twin studies provide an ideal framework to distinguish the effects of genetics from that of the environment.

Monozygotic (MZ, or identical) twins share 100% of their genetic material and as such are genetically identical, whereas dizygotic (DZ, or fraternal) twins share, on average, 50% of their genes, which is the same as a normal sibling relationship. Because MZ twins are genetically identical, most of their differences on certain traits (e.g., height, intelligence, depression) are due to environmental experiences that vary between the twins. Comparing the phenotypic expression of MZ twins and DZ twins on a specific trait can provide insight into the degree of genetic and environmental influence of that trait. If MZ twins are more similar on a specific trait than DZ twins, then this provides evidence that this trait is largely influenced by genes. However, if MZ and DZ evidence that this trait is largely influenced by genes. However, if MZ and DZ twins share a trait to an equal extent, then it is likely that the environment influences the trait more than genetic factors.

Twins also share many aspects of their environment, including their uterine environment, parenting style, education, socioeconomic status, and community because they are born into the same family. However, twin studies are still useful to study trait presentation in twins when there are unique environmental differences between them, such as an event or occurrence that has only affected one twin, like a head injury or birth defect. The presence of a given trait in only one identical twin (called discordance) provides powerful insight into environmental effects on that trait.

To maximize the available data for twin studies, large, worldwide registers of data on twins and their relatives have been established as resources. These registers no longer focus on the assessment of a single phenotype but collect a wide range of traits and environmental factors in twins and their family members. These registers make it possible to conduct analyses of many different variables in relatives, such as using multivariate analysis or including covariates when assessing for the interaction between genotype and environment in influencing a certain trait. This entry reviews the history and methodology of twin studies.

History

Twins have been of interest to scholars, researchers, and artists since early civilization, and they have been proposed as a “natural experiment” in empirical research as early as 415 CE. Sir Francis Galton is usually credited with pioneering the use of twins to study the role of genes and environment on human development and behavior with his 1875 article *The History of Twins*. Galton’s article represents the first detailed attempt to use twins to estimate the relative powers of nature and nurture; however, he did not propose the distinction between MZ and DZ twins when assessing for these differences. In 1924, dermatologist Hermann Werner Seimens introduced the systematic analysis of similarity between MZ and DZ twins. When studying skin moles, Seimens correlated mole counts on one twin with mole counts on the other twin and compared this correlation in MZ and DZ pairs of twins. The correlation for mole count in MZ twins was double that of DZ twins, which indicated the importance of genetic factors in variation in mole count. Seimens’s discovery introduced the idea that any heritable disease will be more concordant in identical twins than in nonidentical twins and concordance will be even lower in nonsiblings.

The role of genetics in determining intelligence as measured by IQ scores has often been addressed through twin studies and examining the correlation of IQs between twins. Such studies have found that between about 40% and 75% of the variance in IQ is attributable to genes, with the remaining percentage accounted for by the environment in which one was raised.

Methodology

The classical twin study explained earlier compares phenotypic resemblance of MZ and DZ twins. The known differences in genetic similarity between MZ and DZ twins combined with the assumption of equal environments create the basis for the twin design for exploring the effects of genetic and environmental effects on a phenotype. By comparing the phenotypic expression of a specific trait in MZ versus DZ twins, an estimate can be made on the extent to which genetic variation determines the phenotypic variation of that trait. If MZ twins are found to resemble each other more on a certain trait than do DZ twins, then the heritability (h^2) of the trait can be estimated from twice the difference between the MZ and DZ correlations. The proportion of the variance that is due to a shared environment is the difference between the total twin correlation and the part that is explained by heritability. That is, $r_{\mathrm{MZ}} – h^2$ in MZ or $r_{\mathrm{DZ}} – h^2/2$ in DZ twins, where r_{MZ} is the correlation between MZ twins and r_{DZ} is the correlation between DZ twins.

Beginning in the 1970s, research on behavior genetics improved from the classic twin study and transitioned into using structural equation modeling or covariance modeling. This procedure uses computationally complex methods to model genetic and environmental effects as the contribution of unmeasured (latent) variables to the potentially multivariate phenotypic differences between individuals. Structural equation modeling can accommodate the analysis of covariant factors (such as gender) on heritability estimates to infer the relative importance of these unmeasured latent factors.

There are many additional designs of twin studies to measure different traits among MZ and DZ twins. Multivariate analyses examine more than one phenotype per person to test for the potential for correlated traits. Co-twin control research studies MZ twins who are perfectly matched for genes and control research studies MZ twins who are perfectly matched for genes and family background and introduce a new factor that differs between twins. Many studies are also conducted that include genotyping twins at candidate loci or marker loci to test the variation of genes between twins and their family members. The classic MZ–DZ design can also be extended to include the testing of parents, siblings, spouses, and offspring of both MZ and DZ twins.

Sara A. Florence and Jennifer Davidtz

See also Correlation; Intelligence Tests; Quasi-Experimental Designs

Further Readings

Boomsma, D., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond. Nature Reviews Genetics, 3(11), 872–882.

Galton, F. (1876). The history of twins, as a criterion of the relative powers of nature and nurture. The Journal of the Anthropological Institute of Great Britain and Ireland, 5, 391–406.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness. Behavior Genetics, 23, 21–28.

Neale, M. C., & Cardon, L. R. (1992). Methodology for genetic studies of twins and families. Dordrecht, the Netherlands: Kluwer Academic Press.

Pam, A., Kemker, S. S., Ross, C. A., & Golden, R. (1996). The “equal environments assumption” in MZ-DZ twin comparisons: An untenable premise of psychiatric genetics? Acta Geneticae Medicae et Gemellologiae, 45(3), 349–360.

Rende, R. D., Plomin, R., & Vandenberg, S. G. (1990). Who discovered the twin method? Behavior Genetics, 20(2), 277–285.

Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. Briefings in Bioinformatics, 3(2), 119–133.

Siemens, H. W. (1924). Die Zwillingspathologie: Ihre Bedeutung, ihre Methodik, ihre bisherigen Ergebnisse {Twin pathology: Its importance, its methodology, its previous results}. Berlin, Germany: Verlag von Julius Springer.

Vogler, G. P. (1993). Methodology for genetic studies of twins and families. Behavior Genetics, 23(1), 107–108.

Winerman, L. (2004, April). A second look at twin studies. Monitor on Psychology, 35(4). Retrieved from http://www.apa.org/monitor/

Jeffrey R. Harring Jeffrey R. Harring Harring, Jeffrey R.

Tessa Johnson Tessa Johnson Johnson, Tessa

Two-Way Analysis of Variance Two-Way analysis of variance

1733

1737

Two-Way Analysis of Variance

Two-way analysis of variance (ANOVA) is a statistical technique used to analyze data from a study in which a researcher wishes to examine both the separate and the combined effects of two categorical independent variables, called factors, on a continuous dependent (or outcome) variable. While the ideas of ANOVA as a statistical approach date back more than two centuries, it was not until the seminal work of R. A. Fisher in the 1920s on analyzing data from complex experiments that two-way ANOVA became a popular, reliable procedure used by practitioners and methodologists alike. This entry first describes the data analytic context for ANOVA and the logic behind its implementation. Two-way ANOVA is then introduced and several key analytic elements are discussed in the context of a real data example.

The Logic of ANOVA

In its simplest form, a one-way ANOVA assesses whether mean differences exist on a single outcome variable across levels of a single factor. Historically, ANOVA was utilized for analyzing experimental data where the independent or grouping variable was manipulated by the researcher. For example, a random sample of subjects desiring to lose weight may be randomly assigned to a dieting group, an exercise group, a dieting and exercise group, and a control group (for which there is no intervention). The mean weight loss computed for each group is compared to every other group to see which treatment was the most effective weight loss regimen. Although ANOVA was initially grounded using data obtained through experimentation, it is applicable to data stemming from quasi-experimental and observational studies as well, where some or all of the factors are not manipulated and groups are intact.

Interestingly, the means of the outcome variable across levels of the factor in ANOVA are not directly compared but rather the magnitudes of their differences are evaluated by partitioning, then comparing, different sources of variability in the outcome. The overall variation in scores on the outcome can be partitioned into two components—variation of individual values around their group means and variation of the group means around the overall mean. These two sources of variation are frequently referred to as variability in within groups and between groups, respectively. If the within-group variation is small compared to the between-group variation, this suggests that the population means are different. Mean differences of levels of a factor are formally tested using a test of significance based on the *F* distribution, which tests the null hypothesis (H_0) that the means of the *J* groups are equal:

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_J.$$

More formally, the *F* test is used to compare the equality of two variances—the variance of scores within groups and the variance of means between groups. These variance estimates, called *mean squares*, are computed as the *sum of squares* divided by their respective degrees of freedom:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{\sum_{j=1}^{J} n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2}{J-1},$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2}{N-J}.$$

The *F* test statistic is calculated as the ratio of these mean squares or variances.

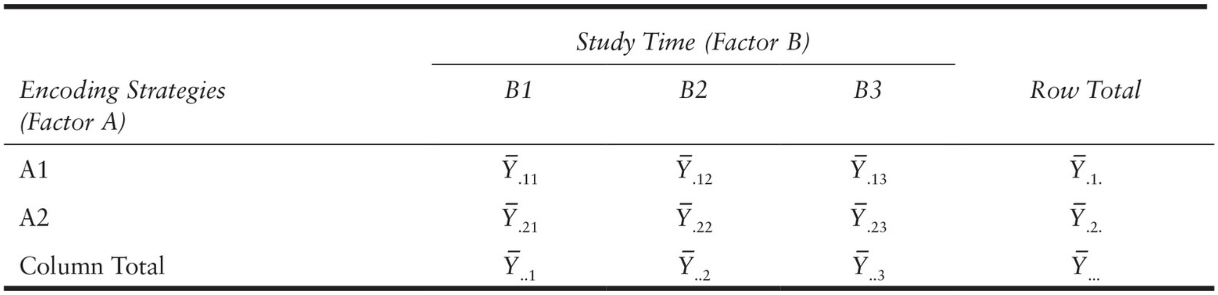
$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}.$$

is an estimate of the population variance, , based upon the deviation of scores about the group means. It is not influenced by mean differences among the groups. is also an estimate of the population variance if the null hypothesis is true. It is based upon the deviations of group means about the grand mean. Because its value is impacted by any group mean differences that exist in the population, it is only an estimate of the same population variance if those group effects are assumed to be zero, that is, if the null hypothesis is true. Under the null hypothesis, these two mean squares are thought to be estimating the same population value, and thus, their ratio should be approximately 1. If there were true group mean differences, would be sensitive to them, but would not. Therefore, a large computed *F* test statistic suggests that group mean differences, in fact, do exist in the population and the null hypothesis should be rejected.

Two-Way ANOVA Designs

The primary difference between a one-way ANOVA and a two-way ANOVA is that data for the latter come from a factorial design in which separate levels of each of two factors (e.g., Factor A and Factor B) are selected and all combinations are formed. Of major interest is whether the effect of Factor B on the outcome measure differs for individuals defined by different levels of Factor A. Consider, for example, a study of the effects of encoding strategies (Factor A) and study time (Factor B) on an outcome variable, number of words recalled. Using a population of adolescents of a certain age (say, 14-to 15-years-old), a random sample of adolescents are randomly assigned to one of the six treatment groups defined by completely crossing two levels of encoding strategies (A1 = memorization, A2 = story and imagery mnemonics) with three levels of study time in minutes (B1 = 30, B2 = 60, and B3 = 180; i.e., A1B1, A1B2, …, A2B3). Table 1 provides the basic design of a 2 × 3 ANOVA, where is the overall or “grand” mean of the Y_{ijk} scores for individual *i* in encoding strategy *j* and study time *k*. The marginal totals for rows and columns denote the mean number of words recalled for levels of Factor A and Factor B. Each cell in a factorial ANOVA is often referred to as a treatment condition or treatment cell.

One analytic approach for these data would be to execute a one-way ANOVA on the six experimental conditions to see which treatment combination elicited the greatest average number of recalled words. This strategy, treating a two-way (or any higher order design) ANOVA as a one-way ANOVA, is problematic, however. Typically, in a two-way ANOVA, investigators are interested in understanding the unique effects of individual factors (called *main effects*) and any combined effect of the factors (called the *interaction*). This information is not available in a one-way analysis and can lead to ambiguity in the results. For example, if differences in average number of words recalled were uncovered between the A1B1 and A2B2 treatment conditions, there is uncertainty as to the *cause* of the effect though only two means are being compared. It is impossible with this analytic approach to determine whether the mean difference in number of words recalled is due to a difference in encoding strategy, a difference in the amount of time dedicated to studying, or both. Analyzing this experimental data using a two-way ANOVA not only allows for the disentanglement of these effects but better aligns with the research hypotheses of interest.



	Study Time (Factor B)			
Encoding Strategies (Factor A)	*B1*	*B2*	*B3*	*Row Total*
A1	$\bar{Y}_{.11}$	$\bar{Y}_{.12}$	$\bar{Y}_{.13}$	$\bar{Y}_{.1.}$
A2	$\bar{Y}_{.21}$	$\bar{Y}_{.22}$	$\bar{Y}_{.23}$	$\bar{Y}_{.2.}$
Column Total	$\bar{Y}_{..1}$	$\bar{Y}_{..2}$	$\bar{Y}_{..3}$	$\bar{Y}_{...}$

Three null hypotheses are of interest in a two-way ANOVA. Continuing with the word recall example, these correspond to (1) testing the main effect of encoding strategies (averaged across study time, is there a mean difference between individuals who used memorization as opposed to imagery?), (2) testing the main effect of study time (averaged across encoding strategies, is there a mean difference between the three study times?), and (3) testing the interaction between encoding strategy and study time (is word recall better for individuals using a particular combination of encoding and study time above and beyond the unique main effects?).

Hypothesis Testing

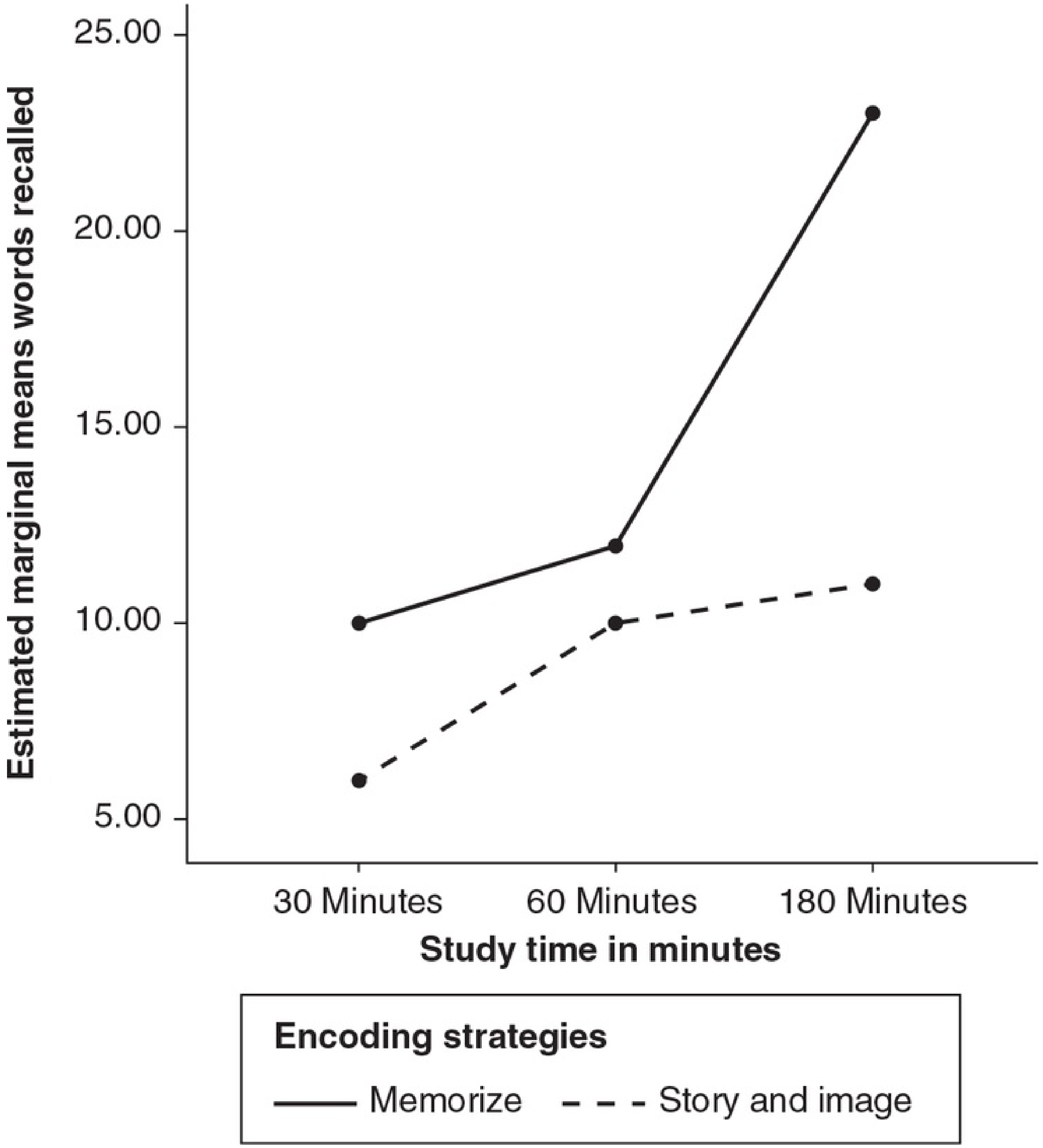
To carry out the two-way ANOVA, the total variability in word recall scores is decomposed into between-factor variability and within-factor (error) variability. However, in contrast with the decomposition in a one-way ANOVA, the between-factor variability itself is further decomposed into variability due to Factor A (encoding strategy), variability due to Factor B (study time), and variability due to the interaction of Factors A and B. Mean squares for each effect are then computed as the appropriate sum of squares divided by its corresponding degrees of freedom. *F* test statistics for each effect are then calculated as the ratio of the mean square of that effect to the mean square of within (mean square error). Statistical significance of each effect is then adjudicated by comparing the computed *F* test statistic to its theoretic sampling distribution following an *F* distribution: .

Post Hoc Analysis

A significant main effect result from the analysis suggests that mean differences exist among levels of the factor; however, isolating exactly which levels are different must still be assessed. In the event that a factor has only two levels, like the encoding strategy factor, then a significant *F* test is sufficient to convey that the two levels are significantly different from each other. If, however, a factor has three or more levels, like the study time factor, following up the omnibus *F* test with a post hoc test such as Tukey’s honest significance test or Scheffe test is common practice. A particular post hoc testing procedure is often chosen among numerous alternatives because it (1) provides adequate Type I error control for carrying out multiple comparisons, (2) provides acceptable power to detect significant differences if they indeed exist, and (3) accommodates the types of comparisons (i.e., pairwise or complex contrasts) that are substantively interesting.

A significant omnibus interaction effect from a two-way design in which at least one factor is three or more levels also indicates the need for follow-up post hoc tests. An interaction plot like that depicted in Figure 1 can nicely illustrate the relation between the two factors (in this case encoding strategy and study time) on the outcome (number of words recalled) and can suggest which mean differences should be further tested.

Figure 1 Interaction plot of encoding strategies and study time on number of words recalled.



Two types of post hoc procedures are common for teasing apart the interaction effect: simple effects and tetrad contrasts. A simple effect is a mean difference in one of the factors within a particular level of the second factor. Based on Figure 1, an investigator may be interested in testing if a mean difference exists between encoding strategies within the 180-minute study time condition. The null hypothesis could be tested. A tetrad contrast is one degree of freedom test involving four cell means. Figure 1 suggests that the mean difference between

180 and 30 minutes of study time is different for the two encoding strategies. The null hypothesis for this contrast would be:

$$H_0 : \mu_{180,\mathrm{img}} - \mu_{30,\mathrm{img}} = \mu_{180,\mathrm{mem}} - \mu_{30,\mathrm{mem}}.$$

Of course, there are many ways to define mean differences as simple effects or tetrad contrasts. Thus, any post hoc testing procedure to be employed must exert a sufficient level of Type I error control on performing multiple tests or comparisons.

Design Considerations and Effect Size

Two-way ANOVA designs can be balanced or unbalanced. A balanced design most often obtained in an experiment is one that has an equal number of subjects in each of the groups. An unbalanced design in which group sample sizes are not equal is indicative of intact groups from an observational study. One advantage of using a balanced design in a two-way ANOVA is that the sum of squares of the effects comprising the are additive and nonoverlapping:

$$\begin{aligned} SS_{\mathrm{total}} &= SS_{\mathrm{between}} + SS_{\mathrm{within}} \\ &= SS_A + SS_B + SS_{AB} + SS_{\mathrm{within}}. \end{aligned}$$

Importantly, this means that the effects are orthogonal and can be interpreted unambiguously. Main and interaction effects can be interpreted in many different ways, but one intuitive approach is to compute the proportion of variance in the outcome that is attributable to each effect. Different measures of association that are regularly used in practice include , , ω^2, and the intraclass correlation, ρ, serves as effect size measures when reporting the results of an analysis. η^2 is computed as and is intuitively interpreted as the proportion of variance in the outcome explained by between-group differences. Due to the balance of the design, this total effect size can be partitioned into separate effects sizes for each factor and the interaction.

$$\begin{aligned} \eta^2 = \frac{SS_{\mathrm{between}}}{SS_{\mathrm{total}}} &= \frac{SS_A}{SS_{\mathrm{total}}} + \frac{SS_B}{SS_{\mathrm{total}}} + \frac{SS_{AB}}{SS_{\mathrm{total}}} \\ &= \eta_A^2 + \eta_B^2 + \eta_{AB}^2. \end{aligned}$$

The primary issue with an unbalanced factorial design is the effects of the factors, and their interaction becomes correlated or nonorthogonal. As a result, the variance components for the main effects are either too large or too small depending on the particular imbalance, which signify that the estimates of the main effects need to be corrected. This problem has long been recognized and corrective procedures have been proposed that primarily involve alternative ways to calculate the sum of squares for each effect. Current statistical software calculates these alternative *SS* with ease; however, the decision as to which *SS* is appropriate for a particular analytic scenario ultimately rests on the shoulders of the researcher.

Conclusions

Like its one-way counterpart, two-way ANOVA focuses on group mean differences where the primary inferential goals are to investigate and test the separate and combined effects of two factors on an outcome variable. The use of two-way ANOVA in conjunction with a factorial design makes it possible to accomplish this within a single study. The same principle can be extended to higher order factorial ANOVA designs that are defined by three or more factors.

Jeffrey R. Harring and Tessa Johnson

See also Analysis of Variance; Effect Size; Experimental Designs; *F* Distribution; Interaction; Post Hoc Analysis; Quasi-Experimental Designs

Further Readings

Gelman, A. (2005). Analysis of variance? Why it is more important than ever. The Annals of Statistics, 33, 1–53.

Kirk, R. E. (2013). Experimental design: Procedures for behavioral sciences (4th ed.). Thousand Oaks, CA: Sage.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use type II instead of type III sums of squares. Statistics and Computing, 13, 163–167.

Lomax, R. G., & Hahs-Vaughn, D. (2013). An introduction to statistical concepts (3rd ed.). New York, NY: Routledge.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments for analyzing data: A model comparison perspective (2nd ed.). New York, NY: Psychology Press.

Roberts, M., & Russo, R. (2014). A student’s guide to analysis of variance. New York, NY: Routledge.

Mary L. McHugh Mary L. McHugh McHugh, Mary L.

Two-Way Chi-Square

Two-Way chi-square

1737

1740

Two-Way Chi-Square

The two-way chi-square is one of a number of tests of goodness of fit between actual values of two nominal or ordinal variables and the values that would be expected if the variables were unrelated to each other or *independent*. The chi-square is also called the Pearson chi-square and the chi-square test of independence. The symbol for the chi-square test is a lower case Greek letter chi (χ) with a 2 as superscript as follows: χ^2. The χ^2 is a significance test and should be accompanied by an appropriate strength test, which should be selected on the basis of the number of rows and columns in a table of observed values. Specifically, the most commonly used strength tests include the π coefficient (for 2 × 2 tables), Cramer’s V (for tables > 2 × 2 but < 5 × 5 or greater tables), or contingency coefficient (for tables equal to or greater than 5 × 5).

The χ^2 is one of the most useful statistics for testing hypotheses when the variables are nominal as often happens in clinical research. Unlike most statistics, the χ^2 can not only provide information on the significance of any observed differences but also provide detailed information on exactly which categories account for any differences found. Thus, the amount and detail of information this statistic can provide renders it one of the most useful tools in the researcher’s array of available statistical tools.

The χ^2 should be used when the researcher wants to know if two variables are related, the variables are measured at the nominal level, and the data represent counts of the number of subjects in each category (or level) of the variable. The subjects in the study in which the χ^2 is used must be independent of each other. Specifically, data from paired or related subjects must be analyzed using other statistics because the χ^2 is not appropriate for such data. Given that the χ^2 is a nonparametric (or distribution free) test, it may be used with collapsed interval/ratio data. This is most commonly done when the assumptions of the parametric statistics usually employed with such data are violated and the researcher must step down to a nonparametric test. For example, if the measure is room temperature in Celsius degrees, but the data are nonnormally distributed, the data may be divided into categories such as “uncomfortably cold,” “comfortable,” and “uncomfortably warm.” Although the χ^2 may be used for 2 × 2 table data, the Fisher’s exact probability test is more appropriate for that type of table. In addition, χ^2 can be used with ordinal data, although other statistics may be more appropriate. This entry reviews the background, presents the assumptions, describes the calculation, and discusses the interpretation of two-way χ^2; the entry concludes with an educational example to illustrate the χ^2 in practice.

Background

Karl Pearson, one of the mathematicians involved in the development of the theory of general linear models, developed χ^2. In 1900, Pearson published his article introducing the χ^2.

π is a correlation statistic, and as such, it measures the *strength* of an association between the two variables. Correlation statistics provide 4 items of information: First, they answer the question, “Do these two variables covary?” That is, does one variable change when the other changes? Second, when two variables covary, these statistics describe the direction of the association, which can be positive or negative. A positive correlation means as one variable increases, the other also increases. A negative correlation means that as one variable increases, the other decreases. Third, correlations describe the strength of the association. Strength in this context means how closely do the two variables change together? In a perfect correlation, for every one level of rise in one variable, the other variable would change exactly one level; it would either rise (positive correlation) or fall (negative correlation) that one level. The π value can range from 0 to +1.0. (Given that the calculation requires the square root of a number, the result cannot be negative with the standard formula. Some other methods of calculation can return a negative number.) Fourth, the significance of the obtained π value can be determined if hand calculated, and the statistical programs that produce π will provide a significance level.

Assumptions

The χ^2 is a nonparametric statistic and thus has fewer assumptions than parametric tests. However, it is important to use it only with data that fit the few assumptions that the χ^2 test demands. Violation of any of the assumptions results in the potential for faulty interpretation of the results and a higher likelihood of a Type I error in which the claim of significance is made for results that are truly not significant. The χ^2 has four key assumptions:

1. Subjects must be randomly selected from the population of interest.
2. Every subject must be independently selected, meaning no subject's selection can be dependent upon or related to the selection of any other subject.
3. Data represent counts of subjects in each category (and as with any nominal measure, each category must be mutually exclusive of every other category).
4. There must be expected values of 5 or greater in at least 80% of the cells in the χ^2 table, and all cells must have expected values of at least 1.

Calculation

The χ^2 is easily hand calculated using the following formula:

$$\sum \frac{(O-E)^2}{E},$$

where O = observed cases and E = expected cases.

To obtain the expected cases, the following formula is used:

$$E = R_m \times C_m \div n,$$

where R_m = row marginal (sum of all values in a row), C_m = column marginal (sum of all values in a column), and n = total sample size.

In addition, the researcher will need to calculate the degrees of freedom (*df*) in order to determine the significance of the obtained χ^2 value. The formula for calculating *df* is as follows:

$$df = (R-1)(C-1),$$

where *R* = the number of rows and *C* = the number of columns.

An important product of manually calculating the χ^2 is that one can view the *cell* χ^2 values. These values are summed to calculate the table χ^2. The cell χ^2 values represent each cell's contribution to the table χ^2 and thus how "deviant" that cell is from what would happen if there was no relationship between the two variables. Cells that have very low χ^2 values can be ignored because the actual number is close to "no relationship." Cells with high χ^2 values are the categories where the lack of independence between the two variables is most pronounced.

Interpretation

The χ^2 value must be compared to a table of χ^2 values to obtain the significance of the test. χ^2 values can be negative or positive, but the direction will depend upon how the variables are set up in the table. Values of 0 or near 0 will not be significant. Values that deviate from 0 must be compared to the table of significances to determine the result, and both the sample size and *df* (calculated based on the number of rows and number of columns) affect the significance of the obtained χ^2 value.

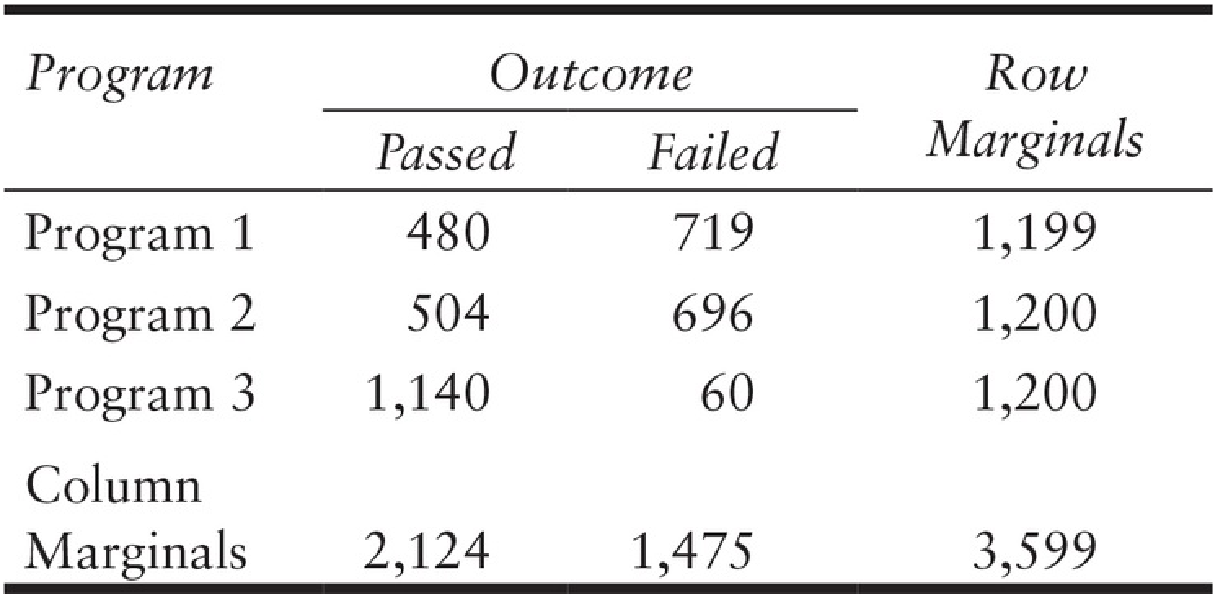
Computer programs that calculate the χ^2 provide the significance of the test as part of the output, so typically the researcher does not have to look up the significance of an obtained χ^2 value. It should be noted there are several sites on the Internet that will calculate the χ^2 and provide a significance level automatically. However, few automatically provide the table of expected values or cell values, and those values are very helpful to interpretation.

Example of the χ^2

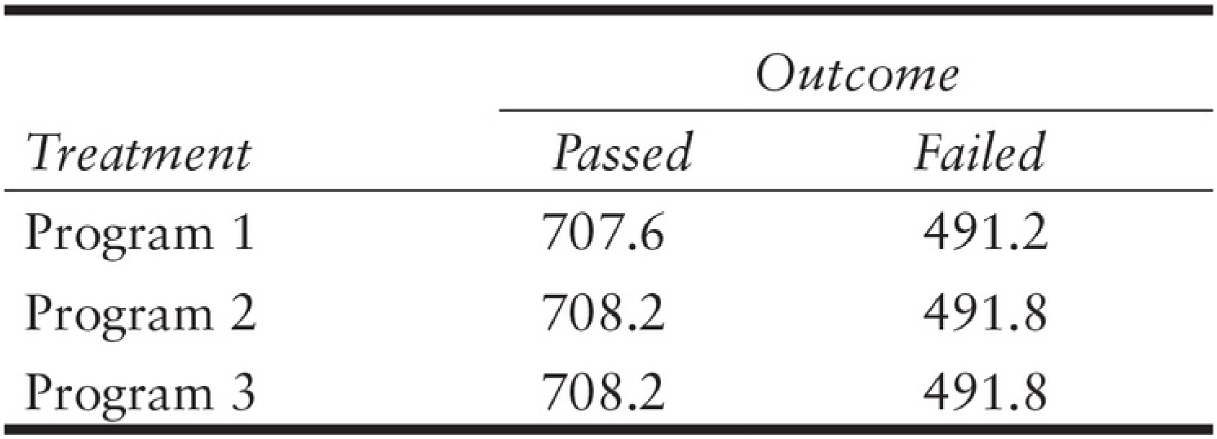
In this hypothetical example, a school system wants to know if a new reading program in first grade is more effective than the existing word-recognition program or the purely phonics-based approach used by many of its schools. The school system has been having a problem with third-grade students failing the reading-level achievement tests given at the end of third grade and has obtained permission to run a study over the next 5 years. The school system will randomly assign each school in the system to one of the three reading programs. (Teachers will have intensive training on each of the programs and will be provided with standardized work sheets, textbooks, and teaching aids for each of the three programs.) During the 5 years of the study, 3,599 children were enrolled in kindergarten and were in the same school they started in at the end of third grade. No other students were included in the study. Table 1 presents the results of students' reading tests at the end of third grade. The dependent variable is the student's performance on the reading achievement test, and the levels are "pass" or "fail."

To calculate the χ^2, first the marginal values must be calculated. In Table 1, the marginal values are italicized. To obtain the expected values for each cell, that cell's column marginal is multiplied by the row marginal and that product is divided by the sample size (*n*). Table 2 presents the cell expected values.

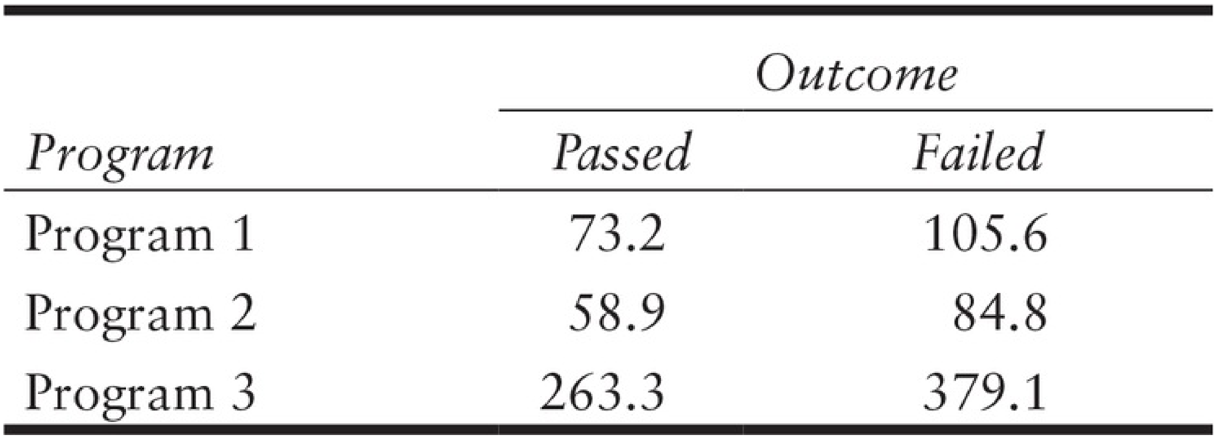
The cell representing students who passed (707.6) enrolled in Program 1 was calculated with the formula (rounded to the nearest 10th). Finally, the cell χ^2 values are presented in Table 3.



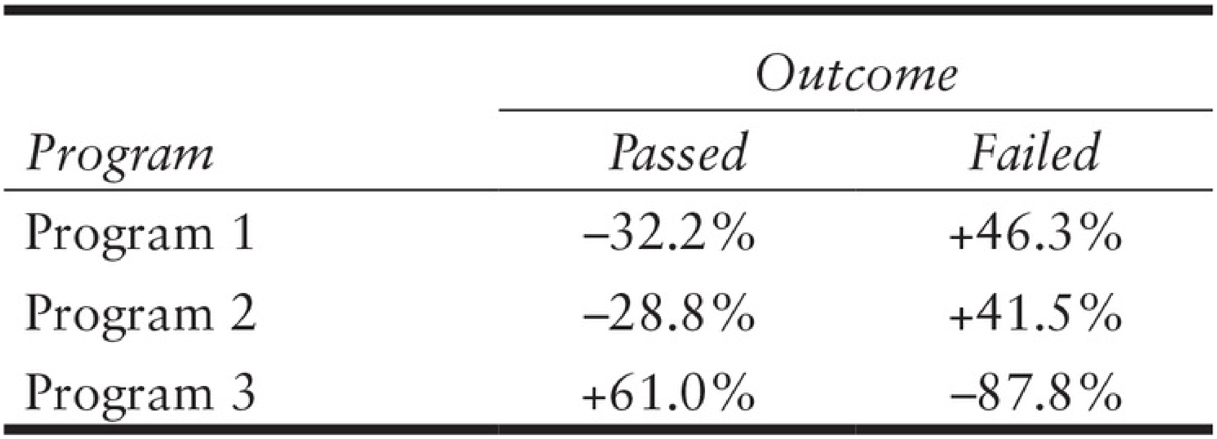
Program	*Outcome*		*Row Marginals*
	Passed	*Failed*	
Program 1	480	719	1,199
Program 2	504	696	1,200
Program 3	1,140	60	1,200
Column Marginals	2,124	1,475	3,599



Treatment	Outcome	
	Passed	*Failed*
Program 1	707.6	491.2
Program 2	708.2	491.8
Program 3	708.2	491.8



Program	Outcome	
	Passed	*Failed*
Program 1	73.2	105.6
Program 2	58.9	84.8
Program 3	263.3	379.1



Program	Outcome	
	Passed	*Failed*
Program 1	–32.2%	+46.3%
Program 2	–28.8%	+41.5%
Program 3	+61.0%	–87.8%

The sum of the cell χ^2 values is 964.9. Eqn107.eps. Looking up the value of 964.9 with two *df*, we find this result is significant at the $p < .0001$ level. The interpretation is that there is a significant difference among the three programs for reading outcomes, $\chi^2(2) = 964.9$, $p < .0001$. Inspection of the cell, χ^2 values reveals test failures were higher than expected and number of students passing the test were lower than expected for both Program 1 and Program 2. The greatest differences, however, were the much higher than expected passed tests and much lower than expected failed tests for Program 3. Thus, the best program for teaching reading such that the students pass the reading achievement test in third grade is Program 3.

In addition, the researcher can calculate the percentage deviation of each cell from its expected values. The higher the percentage deviation, the more that cell contributes to any differences found among the sample. The percentage deviation is calculated as (observed – expected)/expected × 100. The table of percentage deviations is presented in Table 4. Table 4 shows that the largest deviations are the much higher than expected percentage of students in Program 3 who passed the reading test and the much smaller percentage in this group (−87.8%) who failed the reading achievement test.

The χ^2 only tests significance, which merely reflects how likely this result is to be found in the full population of interest but does not provide an effect size. The appropriate effect size for this 2 × 3 table is the Cramer's *V* test, which provides an effect size of .5177 (rounded to .52) that represents a moderately strong effect size.

Mary L. McHugh

See also Chi-Square Test; Phi Correlation Coefficient

Further Readings

Norton, B. J. Karl Pearson and statistics: The social origins of scientific innovation. In Social Studies of Science (Vol. 8, No. 1), Theme Issue: Sociology of Mathematics (February 1978) (pp. 3–34). Thousand Oaks, CA: Sage.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302). Retrieved from http://www.tandfonline.com/doi/abs/10.1080/14786440009463897

VassarStats. (2016). Chi-square, Cramer's V, and Lambda online calculator. Author. Retrieved June 22, 2016, from http://vassarstats.net/newcs.html

Brenda Hannon Brenda Hannon Hannon, Brenda

Type I Error

Type I error

1740

1743

Type I Error

In the context of statistical hypothesis testing, a Type I error occurs when the null hypothesis is rejected when, in fact, the null hypothesis should have been accepted. More specifically, a researcher observed a significant difference between two experimental conditions and consequently rejected the null hypothesis when, in truth, the observed significant difference between the two experimental conditions did not occur because of the manipulation, rather it occurred because of random chance. Three everyday examples of Type I errors are when a medical test indicates that a patient has a disease when, in fact, the patient is actually disease-free; when a fire alarm indicates there is a fire when, in fact, there is no fire; and when a jury decides a person is guilty of a crime when, in fact, that person is innocent. Type I errors also occur in educational research; therefore, this entry considers various educational examples to further illustrate Type I errors.

Understanding Type I Errors

In statistical analysis, hypothesis testing is used to determine whether the variations among different groups can be attributed to a manipulation or random chance. In educational contexts, hypothesis testing generally includes two types of hypotheses: a null hypothesis and alternative hypothesis. A null hypothesis states that the phenomenon or manipulation under investigation produces no effect (i.e., or makes no difference). An alternative hypothesis (also referred to as the research hypothesis) states the opposite of the null hypothesis; that is, an alternative hypothesis states that the phenomenon or manipulation under investigation *does* produce an effect (i.e., it *does* make a difference). An example of a null hypothesis is “The new teaching strategy does not positively influence students’ learning outcomes.” An example of an alternative hypothesis is “The new teaching strategy does positively influence students’ learning outcomes.”

Because researchers want to correctly conclude that variations among different groups are attributed to a manipulation rather than random chance, researchers take precautions to avoid making a false claim. That is, researchers take precautions against making a Type I error. This precaution is setting a *level of significance*, which is really just a “rejection of the null hypothesis” decision threshold that is based on probability. For example, a researcher may set a level of significance to .05, meaning that the researcher is willing to say 5 times out of 100 that the variation among the groups is attributed to the manipulation when, in fact, the variation among the groups is attributed to random chance. In other words, a level of significance of .05 means that 5 times out of 100, the researcher will commit a Type I error (i.e., claiming that the variations among the different groups is attributed to a manipulation when, in fact, the variation is attributed to random chance).

Consider the following hypotheses:

> Null hypothesis: The new measure is not a better predictor of academic outcomes than is the old measure.
>
> Research hypothesis: The new measure is a better predictor of academic outcomes than is the old measure.

If the level of significance is set to .01 and the researchers reject the null hypothesis based on their significant statistical results, then the researcher has a 99 in 100 chance of correctly saying that the new measure is a better predictor of academic outcomes than is the old measure and a 1 in 100 chance of incorrectly saying that the new measure is not a better predictor of academic outcomes than is the old measure. In other words, the researcher has a 1 in 100 chance of making a Type I error (i.e., claiming that the new measure is indeed a better predictor of academic outcomes than the old measure when, in reality, the new measure is not a better predictor of academic outcomes than is the old measure).

Controlling the Rate of Type I Errors

Controlling the Rate of Type I Errors

The most common conventions in educational research is to set the level of significance to .05 (i.e., approximately 5 times out of 100 a researcher will claim that the variations among groups is attributed to the manipulation when it actually is a consequence of random chance) or .01 (i.e., approximately 1 time out of 100 a researcher will claim that the variations among groups is attributed to the manipulation when it actually is a consequence of random chance). However, this is only a convention. When making decisions about how to control for Type I errors, researchers should consider other factors such as the potential benefits and the potential risks of their research (i.e., the real-world impact of the research findings); the need for replication; and pairwise versus experiment-wise Type I error rates.

Potential Benefits and Potential Risks

The potential benefits and potential risks should be assessed for every study, whether the study is medical or educational. Potential benefits are interpreted as the amount or level of positive real-world impact that the significant findings will have, whereas potential risks are interpreted as the amount or level of negative real-world impact that the significant findings will have. Consider the following hypothetical situation:

> A researcher has developed a new cancer drug that has the potential to cure cancer but, based on animal testing, there is a 10 out of 100 chance of killing a patient. The researcher’s null hypothesis is that the new drug does not cure cancer. The researcher’s alternative hypothesis is that the new drug does cure cancer.

In such a situation, should the researcher set a .05 level of significance, thereby running a 5% chance of claiming that the drug works when it does not as well as running a 10% chance of killing a patient? Should the researcher set a lower level of significance, such as .01, thereby running a 1% chance of claiming that the drug works when it does not as well as running a 10% chance of killing a patient? Or, because the drug has the potential to cure cancer, should the researcher set a higher level of significance (i.e., .10 or higher), thereby running a high chance of claiming the drug works when it does not as well as running a 10% chance of killing a patient? And, what if this drug were for terminal

10% chance of killing a patient? And, what if this drug were for terminal patients who are expected to live only 6 more months and 5 of these 6 months are in agony? Clearly, in such instances, a great deal of thought about the real-world benefits and the real-world risks must be considered before a level of significance is set.

Replication

When deciding on how to control for Type I error rates, researchers should also consider whether the research has been replicated. That is, is the current study the first study to test the influence of a particular variable or is the current study attempting to replicate previous findings? Even when a level of significance is set to a very strict .01 (i.e., probability is less than 1%), there is still a probability that 1 out of 100 times a researcher will claim significant results when, in fact, the significant results are a consequence of random chance. If, however, other researchers are able to replicate the experiment multiple times and find similar results, the probability of 1 out of 100 times of having a false result may be harsh.

Pairwise Versus Experiment-wise Type I Error Rates

Another important consideration when considering Type I error rates is to determine a pairwise Type I error rate versus an experiment-wise Type I error rate. A pairwise Type I error rate is the error rate that is set for each statistical test that is completed in the same study, whereas an experiment-wise Type I error rate is the overall Type I error rate for the entire experiment. Consider the following two scenarios where each *t* test has a .05 level of significance:

> Scenario 1: A researcher completes a single *t* test in an experiment.
> Scenario 2: A researcher completes five *t* tests in an experiment.

In Scenario 1, the pairwise Type I error rate is .05 and because only one *t* test was completed the experiment-wise Type I error rate is .05. In contrast, in Scenario 2, the pairwise Type I error rate is still .05; however, the experiment-wise Type I error rate has increased to .25 because there were five *t* tests that each had a .05 Type I error rate (i.e., 5 tests × .05 = .25). What this .25 experiment-wise Type I error rate means is that there is a 25 out of 100 chance that the researcher will claim that one of the *t* tests is significant when, in fact, the significant result is a consequence of random chance.

To control the influences that multiple tests have on the experiment-wise Type I error rate, researchers usually set more strict standards (i.e., reduce the level of significance) for claiming significance for each test. The rationale behind setting more strict standards is by reducing the probability of a Type I error rate for each test the probability of the overall experiment-wise Type I error rate is also reduced. For example, if a researcher wishes to maintain an experiment-wise Type I error rate of .05 and has 10 tests to complete, then the α level for each test should be set to .005 (i.e., .05/10 = .005). This technique is also known as the Bonferroni correction. Many post hoc tests, which are used with tests of analysis of variance, also consider pairwise and experiment-wise Type I error rates.

The Relationship Between Type I and Type II Errors

Although the discussion so far has been based solely on Type I errors, Type I errors do not occur in isolation. Rather, discussions of Type I errors are frequently accompanied by discussions of Type II errors. Although a Type I error occurs when a researcher rejects a null hypothesis when the null hypothesis should have been accepted (determined by the level of significance), a Type II error occurs when a researcher accepts the null hypothesis when the null hypothesis should have been rejected. That is, the researcher claimed there was no significant result when, in fact, there was a significant effect. Examples of a Type II error include claiming a new learning strategy does not improve students’ learning outcomes when, in fact, the new learning strategy does improve students’ learning outcomes. Or, when a diagnostic measure indicates a child does not have autism when, in fact, the child does have autism. Thus, in many respects, a Type II error is the opposite of a Type I error; a Type II error occurs when researchers fail to reject a null hypothesis when they should have rejected it, whereas a Type I error occurs when researchers reject a null hypothesis when they should not have rejected it.

Additional Examples of Type I Errors

Recall that a Type I error occurs when researchers reject the null hypothesis because they believe that the observed significant result is a consequence of their manipulation when, in fact, the significant result is not a consequence of the researcher’s manipulation. Rather the significant result is a consequence of random chance.

Example 1

> Null hypothesis: Daily music training does not increase the general intelligence level of preschoolers.
> Research hypothesis: Daily music training does increase the general intelligence level of preschoolers.

In this particular example, researchers will be committing a Type I error if they reject the null hypothesis (thereby claiming that daily music training *does* increase general intelligence) when, in fact, they should have accepted the null hypothesis (because the significant result was a consequence of random chance, not the music manipulation).

Example 2

> Null hypothesis: Level of phonemic awareness is not related to the reading abilities of beginning readers.
> Research hypothesis: Level of phonemic awareness is positively related to the reading abilities of beginning readers.

In this particular example, researchers will be committing a Type I error if they reject the null hypothesis (thereby claiming that phonemic awareness *is* related to the reading abilities of beginning readers) when, in fact, they should have accepted the null hypothesis (because the significant result was a consequence of random chance, not level of phonemic awareness).

Brenda Hannon

See also Power; Significance; Type II Error; Type III Error

Further Readings

Garside, G. R., & Mack, C. (1976). Actual type 1 error probabilities for various tests in the homogeneity case of the 2 × 2 contingency table. The American Statistician, 30(1), 18–21.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. Epidemiology, 1(1), 43–46.

Xiaofeng Steven Liu Xiaofeng Steven Liu Liu, Xiaofeng Steven

Type II Error

Type II error

1743

1743

Type II Error

Type II error refers to the probability of not rejecting a false null hypothesis in hypothesis testing; it is denoted by a Greek symbol β. For instance, a hypothesis test is set up to examine the presence of bias in a new standardized test. The null hypothesis states that there is no bias. If there is indeed bias in the test, not rejecting the null hypothesis means failure to confirm the suspected bias.

The concept of Type II error was conceived by Jerzy Neyman and Egon Pearson who developed a mathematical framework later known as Neyman–Pearson lemma to quantify Type II error in hypothesis testing. They considered decision behavior in the significance test and theorized Type I error (false positive) and Type II error (false negative).

Type II error is commonly associated with false negative in decision making because hypothesis testing resembles dichotomous decision making—a positive or negative decision in the end. A no answer means not rejecting the null hypothesis. If the null hypothesis is false, not rejecting it will fail to confirm a researcher’s theory stated in the alternative hypothesis—this constitutes an error of the second kind or Type II error. A close analogue of Type II error can be found in a criminal trial that can render a verdict of guilty or not guilty. A guilty verdict corresponds to a positive decision and a not guilty verdict to a negative decision. When the suspect who indeed committed the crime is acquitted, the verdict of not guilty would be considered a travesty of justice. In this case, the decision is “false negative”—it lets go a real criminal.

Type II error is related to Type I error and statistical power. The latter represents the probability of rejecting a false null hypothesis. As rejecting a false null hypothesis is an event complementary to not rejecting a false null, statistical power can be expressed as one minus Type II error, that is, 1 − β. Thus, increasing statistical power will lower the Type II error in hypothesis testing. In addition, Type II error has an inverse relationship with Type I error. As Type I error goes down, Type II error goes up. The Type I error rate is traditionally limited to 5%, the significance level in hypothesis testing. If the significance level is lowered to 1%, it will be more difficult to reject the null hypothesis. Consequently, the Type II error rate will increase.

Xiaofeng Steven Liu

See also Significance; Type I Error; Type III Error

Further Readings

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Kraemer, H. C., & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Lehmann, E. L. (1993). The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association, 88, 1242–1249.

Charlotte Tate Charlotte Tate Tate, Charlotte

Type III Error

Type III error

1744

1746

Type III Error

Statisticians and researchers are most familiar with two types of decision errors that can occur in empirical research using null hypothesis statistical testing (NHST): the false positive (Type I error) and the false negative (Type II error). However, statisticians have also identified another class of epistemic decision errors that also have some (yet-to-be formalized) probability of occurring. In 1957, A. W. Kimball called this class of error as Type III error, whose pithy description was “giving the right answer to the wrong problem” (p. 134). What is more, this error was relevant to all of statistical thinking, not simply NHST decisions. Kimball details a variety of ways in which the right answer to the wrong problem can be given based on the match between theory, research methods, and statistical evaluation of either the theory or the methods. Kimball’s examples include (a) running a test for independent correlation coefficients on a matched samples design and (b) making statistical transformations incorrectly based on not understanding the experimental method used—both of which provide the right answer to a separate question not being asked in the design of each example study.

Other Definitions of Type III Error

Kimball was neither the first nor the only statistician to provide a definition of Type III error. In fact, Type III error has several definitions with none appearing to enjoy wide acceptance. Frederick Mosteller described Type III error with respect to the null hypothesis as *correctly rejecting the null hypothesis for the wrong reason*. Henry F. Kaiser argued that a Type III error is an incorrect decision of the direction of rejection in a two-tailed NHST. In addition, S. Schwartz and K. M. Carpenter defined Type III error as a discrepancy between the causal components of a theory and how they are operationalized. For instance, the rate of homelessness in any country is structural—based on policies related to housing and poverty—but if researchers focused on the demographics of homeless individuals, they would miss the structural component and believe that individual differences might contribute to or control the phenomenon. Similarly, Schwartz and Carpenter argued that obesity is a result of gene and environment interplays, but research that focuses on individual difference contributions to obesity (e.g., gender, age) will routinely miss or underplay the environmental factors.

The Commonality Among the Type III Error Definitions

Although disparately focused, each of the definitions of Type III error provided in the previous section can reasonably be subsumed under Kimball's axiom of *giving the right answer to the wrong problem*. The misspecification errors of Mosteller, Kaiser, and Schwartz and Carpenter can be viewed as correctly answering another question—just the wrong problem for the investigation at hand. In this way, Type III errors appear to be focused on how theories become operationalized for empirical research. Stated differently, Type III errors focus on methodological implementation errors.

The Need for a Fourth Type of Error

If one can argue for *giving the right answer to the wrong problem* as an epistemic decision error (i.e., Type III error), then it is worth asking: Can one demonstrate that *giving the wrong answer to the right problem* also exists as another kind of epistemic decision error? The answer appears to be yes insofar as there are existing definitions of Type IV error that focus on the incorrect interpretation of an interaction term in regression or analysis of variance.

Yet, incorrectly interpreting one kind of statistical effect might simply be an exemplar for the general phenomenon of misspecifying any statistical effect by violating the underlying assumptions of the test. In multiple regression, for instance, it is known that collinearity of the predictors reduces one's ability to interpret the influence of either predictor. Thus, if one has severe collinearity between predictors and still proceeds with analysis, this can be viewed as *giving the wrong answer to the right problem* insofar as the setup is correct (i.e., the multiple regression analysis is the right problem [as in correct tool]), but the wrong or incorrect answer was provided because the assumptions of the statistical technique were violated.

Recall that Kimball's definition of Type III error appears to subsume all errors that are not Type I or Type II, but this entry has tried to show that it is worthwhile to specify the location of the error. Is the location theory or method implementation (Type III) or statistical evaluation (Type IV) when considering the conclusions made by researchers? Depending on the location of the error, different strategies must be enacted to remove or resolve the error.

Moving Forward With Type III and IV Errors

The import of considering whether epistemic errors occur on the side of theory or method implementation (Type III error) or statistical evaluation (Type IV error) for empirical research is that attention can be focused on two sources of errors in addition to false positive (Type I error) and false negative (Type II error). In fact, with this organization of information, one can see that Type III or IV errors could lead to either Type I or II errors. For example, not noticing and not addressing outliers in a variance-based test (a Type IV error) can provide either a false positive effect (Type I error) or a false negative effect (Type II error)—depending on outlier locations across the comparison groups. Likewise, focusing on obesity as only gene based—when it is also influenced by the environment (a Type III error)—can produce either Type I or Type II errors on the genetic influence itself across studies. In the end, Type III and IV errors provide researchers and statisticians with relevant language and conceptual tools to understand the many sources of influence on statistical inference with or without relying on the NHST.

Charlotte Tate

See also Statistical Inference; Type I Error; Type II Error

Further Readings

Kaiser, H. F. (1960). Directional statistical decisions. Psychological Review, 67,

160–167.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. Journal of the American Statistical Association, 52(278), 133–142.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post-hoc comparisons for interaction terms and nested hypotheses in the analysis of variance designs: The elimination of type IV errors. American Educational Research Journal, 7, 397–421.

Mosteller, F. (1948). A *k*-sample slippage test for an extreme population. The Annals of Mathematical Statistics, 19, 58–65.

Schwartz, S., & Carpenter, K. M. (1999). The right answer for the wrong question: Consequences of Type III errors for public health research. American Journal of Public Health, 89, 1175–1180.

Umesh, U. N., Peterson, R. A., McCann-Nelson, M., & Vaidyananthan, R. (1996). Type IV error in marketing research: The investigations of ANOVA interactions. Journal of the Academy of Marketing Science, 24, 17–26.

U

Rachel Darley Gary Rachel Darley Gary Gary, Rachel Darley

UCINET

UCINET

1747

1751

UCINET

Originally developed by Linton Freeman, UCINET is a comprehensive social network analysis software program. Designed by Steve Borgatti, Martin Everett, and Linton Freeman, Version 6 (the current version as of 2017) is a menu-driven Windows-based program to be used even by non-tech-savvy researchers. Mostly used by social scientists in the field of sociology and management, UCINET also has growing users in the fields of education, health, and life sciences. In general, UCINET is used to analyze sociometric survey data; however, it is also used to analyze other one-mode and two-mode matrix data. UCINET features a number of metric routines used to describe positions of nodes, dyads, groups, and whole networks.

Routine functions of UCINET include algorithmic outputs unique to social network analysis methods such as measures of centrality (e.g., degree, betweenness, closeness, and eigenvector) and cohesion (e.g., density, fragmentation, and components), permutation-based statistical analysis (e.g., *t* tests, analysis of variance, and regression), as well as algebraic and multivariate statistical matrix analyses. Built in to UCINET is the visualization program NetDraw. Developed by Borgatti, NetDraw works in tandem with UCINET to draw diagrams or visualize social networks.

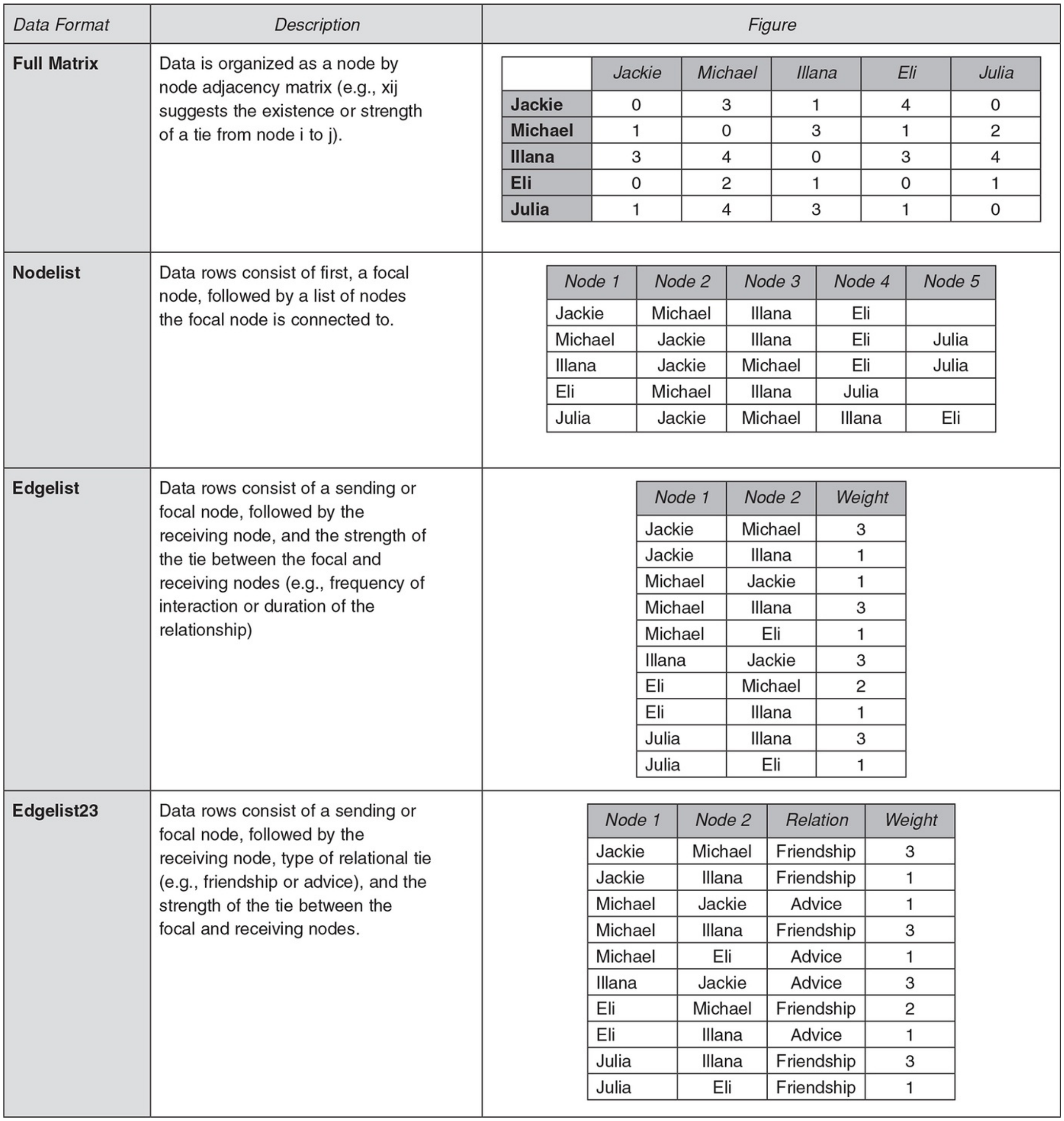
Major Routine Functions of UCINET

Typical users of UCINET are those familiar with the fundamental theories and methods associated with network analysis. Social capital and diffusion of innovation theories ground the majority of social network research. For example, social capital studies are geared toward the investigation of how social connections support or constrain opportunity while diffusion of innovation research is concerned with the exploration of how knowledge and resources spread. UCINET is useful to researchers interested in these theories as the program is designed for the mathematical and visual analysis of network data.

UCINET has the capacity to read and write a variety of differently formatted data and text file types, as in Pajek, Krackplot, Negopy, and the VNA format used by NetDraw. Unlike other social network analysis programs, UCINET can also import and export Microsoft Excel files. At maximum, the program can handle network data comprising 32,767 nodes or actors; however, many of the routine procedures become less efficient with around 5,000 to 100,000 nodes.

UCINET requires a data set for most routine functions. As a result, the program stores and describes almost all data as collections of one or more matrices. Examples of data formats are described in Figure 1.

Figure 1 UCINET data formats. This table provides descriptions and examples about UCINET acceptable data formats. Adapted from Borgatti (2014).



<table>
<tr><th>Data Format</th><th>Description</th><th>Figure</th></tr>
<tr><td><b>Full Matrix</b></td><td>Data is organized as a node by node adjacency matrix (e.g., xij suggests the existence or strength of a tie from node i to j).</td><td>
<table>
<tr><th></th><th>Jackie</th><th>Michael</th><th>Illana</th><th>Eli</th><th>Julia</th></tr>
<tr><td><b>Jackie</b></td><td>0</td><td>3</td><td>1</td><td>4</td><td>0</td></tr>
<tr><td><b>Michael</b></td><td>1</td><td>0</td><td>3</td><td>1</td><td>2</td></tr>
<tr><td><b>Illana</b></td><td>3</td><td>4</td><td>0</td><td>3</td><td>4</td></tr>
<tr><td><b>Eli</b></td><td>0</td><td>2</td><td>1</td><td>0</td><td>1</td></tr>
<tr><td><b>Julia</b></td><td>1</td><td>4</td><td>3</td><td>1</td><td>0</td></tr>
</table>
</td></tr>
<tr><td><b>Nodelist</b></td><td>Data rows consist of first, a focal node, followed by a list of nodes the focal node is connected to.</td><td>
<table>
<tr><th>Node 1</th><th>Node 2</th><th>Node 3</th><th>Node 4</th><th>Node 5</th></tr>
<tr><td>Jackie</td><td>Michael</td><td>Illana</td><td>Eli</td><td></td></tr>
<tr><td>Michael</td><td>Jackie</td><td>Illana</td><td>Eli</td><td>Julia</td></tr>
<tr><td>Illana</td><td>Jackie</td><td>Michael</td><td>Eli</td><td>Julia</td></tr>
<tr><td>Eli</td><td>Michael</td><td>Illana</td><td>Julia</td><td></td></tr>
<tr><td>Julia</td><td>Jackie</td><td>Michael</td><td>Illana</td><td>Eli</td></tr>
</table>
</td></tr>
<tr><td><b>Edgelist</b></td><td>Data rows consist of a sending or focal node, followed by the receiving node, and the strength of the tie between the focal and receiving nodes (e.g., frequency of interaction or duration of the relationship)</td><td>
<table>
<tr><th>Node 1</th><th>Node 2</th><th>Weight</th></tr>
<tr><td>Jackie</td><td>Michael</td><td>3</td></tr>
<tr><td>Jackie</td><td>Illana</td><td>1</td></tr>
<tr><td>Michael</td><td>Jackie</td><td>1</td></tr>
<tr><td>Michael</td><td>Illana</td><td>3</td></tr>
<tr><td>Michael</td><td>Eli</td><td>1</td></tr>
<tr><td>Illana</td><td>Jackie</td><td>3</td></tr>
<tr><td>Eli</td><td>Michael</td><td>2</td></tr>
<tr><td>Eli</td><td>Illana</td><td>1</td></tr>
<tr><td>Julia</td><td>Illana</td><td>3</td></tr>
<tr><td>Julia</td><td>Eli</td><td>1</td></tr>
</table>
</td></tr>
<tr><td><b>Edgelist23</b></td><td>Data rows consist of a sending or focal node, followed by the receiving node, type of relational tie (e.g., friendship or advice), and the strength of the tie between the focal and receiving nodes.</td><td>
<table>
<tr><th>Node 1</th><th>Node 2</th><th>Relation</th><th>Weight</th></tr>
<tr><td>Jackie</td><td>Michael</td><td>Friendship</td><td>3</td></tr>
<tr><td>Jackie</td><td>Illana</td><td>Friendship</td><td>1</td></tr>
<tr><td>Michael</td><td>Jackie</td><td>Advice</td><td>1</td></tr>
<tr><td>Michael</td><td>Illana</td><td>Friendship</td><td>3</td></tr>
<tr><td>Michael</td><td>Eli</td><td>Advice</td><td>1</td></tr>
<tr><td>Illana</td><td>Jackie</td><td>Advice</td><td>3</td></tr>
<tr><td>Eli</td><td>Michael</td><td>Friendship</td><td>2</td></tr>
<tr><td>Eli</td><td>Illana</td><td>Advice</td><td>1</td></tr>
<tr><td>Julia</td><td>Illana</td><td>Friendship</td><td>3</td></tr>
<tr><td>Julia</td><td>Eli</td><td>Friendship</td><td>1</td></tr>
</table>
</td></tr>
</table>

Breakdown of UCINET Menu Items and Major Routine Functions

UCINET 6.616 is made of the following menu items: file, data, transform, tools, network, visualize, options, and help. Most menu items are nested, comprising additional submenus and choices (e.g., network > cohesion measures > multiple cohesion measures). As with most menu-based programs, the file menu tab deals with routines specific to the files, folders, printing, and exiting of UCINET.

In this entry, the following sections describe and provide examples of the major routine functions associated with four of the menu items (data, transform, tools, and network), followed by a section on UCINET output, visualization (i.e., NetDraw), and a final section that reviews obtaining UCINET and program help features.

Data

The data menu item comprises routine functions specific to managing UCINET data sets such as data import, export, editing, and manipulation. In its simplest form, the method for data entry or import involves a cut and paste of network data from an Excel file into UCINET's data language editor and vice versa.

The data menu tab also includes routines to join rows, columns, or matrices, unpack multi-relational data sets into individual matrices and to identify affiliations by converting a two-mode matrix to a one-mode matrix (e.g., convert a two-mode matrix of teachers by teams to a one-mode matrix of teachers by teachers as connections are defined by faculty comembership on teams).

Transform

The transform menu item includes routine functions for transforming network and graph data into other types. There are routines to combine rows and/or columns, which, for example, allow users to create block densities that are later transformed into block models. Some other routines include options to symmetrize, dichotomize, recode, and standardize matrix data along with operations to create multigraphs via conversion of valued network data into a collection of binary (i.e., 1 or 0) adjacency matrices for each value (e.g., 1, 2, 3, and 4).

Tools

The tools menu item includes routines widely used by network analysts but are not classified as network procedures. Submenus contain multivariate statistical routines such as cluster analyses (e.g., hierarchical and cluster adequacy) and network scaling or decomposition functions (e.g., metric and nonmetric multidimensional scaling). Network analysts tend to frequently use the hypothesis testing routines as these provide valuable node-, dyadic-, and mixed-level statistical descriptive methods including QAP regression and correlation, analysis of variance, and *t* tests. For example, a node-level hypothesis that more peripheral faculty in the school network tend to have lower job satisfaction can be tested. Other routines for creating scatterplots, dendrodiagrams, and tree diagrams from network data are also located within the submenus of the tools tab.

Figure 2 Teacher × Team Two-Mode Matrix to Teacher × Teacher One-Mode Affiliation Matrix. This figure shows the conversion of a two-mode matrix of teachers by teams to a one-mode matrix of teachers by teachers as connections are defined by faculty comembership on teams. UCINET routine aata > affiliations (2-mode to 1-mode).

Teacher x Team	**TEAM**				
TEACHER	*History*	*Math*	*English*	*Science*	*World Language*
M. Smith	1				
J. Garcia				1	
E. Penze					1
Y. Riddle		1		1	
G. Miller			1		
P. Jones		1			

↓

Teacher x Team	***TEACHER***					
TEACHER	*M. Smith*	*J. Garcia*	*E. Penze*	*Y. Riddle*	*G. Miller*	*P. Jones*
M. Smith						
J. Garcia						
E. Penze						
Y. Riddle						
G. Miller						
P. Jones						

Network

The network menu item comprises routines highly specific to network analysis. Here, network analysts can find a comprehensive set of network-specific routines; however, this entry discusses only the major routines of cohesion, centrality, and ego-level metrics.

Cohesion

Cohesion refers to the extent to which network actors are tangled up or knitted together. Cohesion routines enable the examination and description of the structural features of a network from various angles. Measures of cohesion such as density, average degree, and components provide network analysts with valuable information for making predictions or assumptions about network characteristics and capacity. UCINET includes more sophisticated cohesion routines such as connectedness, geodesic distances, and homophily. For example, density refers to the number of existing ties between people divided by the number of total possible ties, whereas connectedness measures the proportion of pairs of mutually reachable nodes.

Centrality

Centrality refers to node-level positionality as situated within a network structure. Of all network metrics, centrality measures are conceivably the most widely used by social network analysts. As such, UCINET provides a large array of standard centrality routines such as degree, betweenness, and closeness while offering more advanced options within these routines. For example, not only does the closeness centrality routine compute the normalized sum of geodesic distances, but users are also provided with additional options such as choice regarding the sums of reciprocal distances and strategies to deal with unreachable nodes.

Ego-Level Metrics

Ego-level metric routines are concerned with both the structural features, such as density and ego betweenness, and the compositional features, referring to the specific attributional characteristics such as an individual’s gender and ethnicity. These routines permit users to extract ego-level data from the whole network

These routines permit users to extract ego-level data from the whole network data, which in turn allows for interpretable metrics at the node level.

UCINET Output

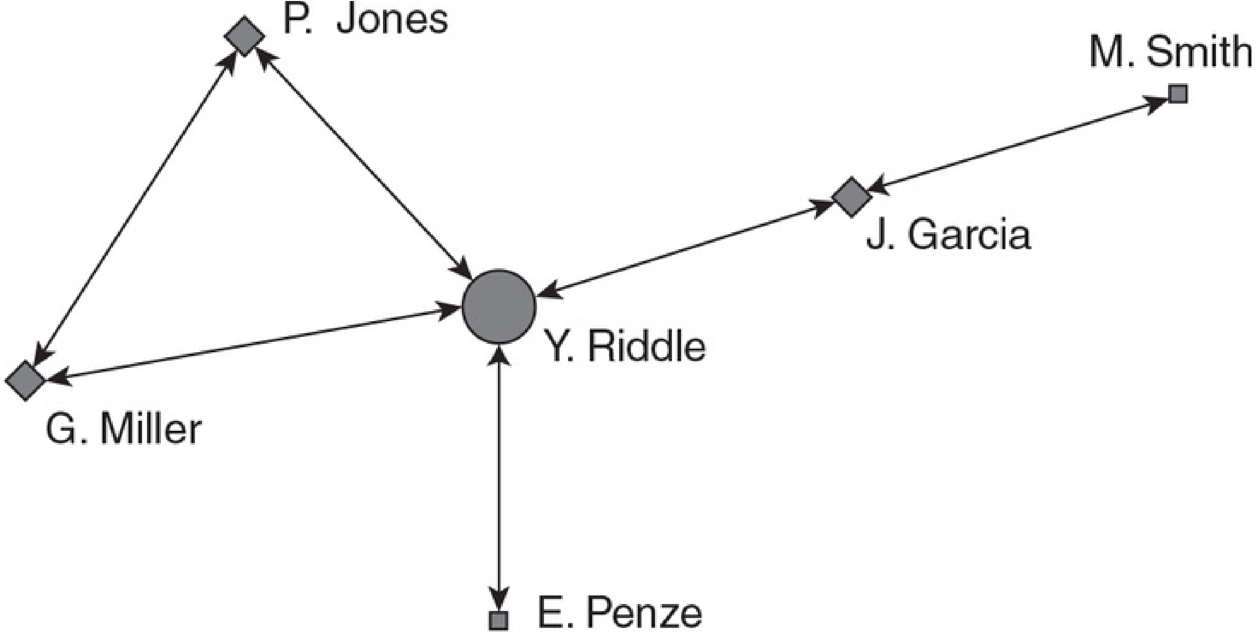
Routines run in UCINET generate two types of output: a text file and a data file. By default, when a routine is run, the text file displays results using Windows Notepad. During an active session, the text files are automatically saved by UCINET; however, once the session is closed, these text files are deleted. Users must choose to manually save text files if they seek to revisit them at later sessions. Data files also contain the same results of the routine run, yet unlike text files, data files are not deleted at the end of a UCINET session and can be revisited by the user at later sessions. When same routines are run, data files are overwritten. Data files can be viewed using the export feature of the data language editor and results subsequently copied and pasted into Excel. Similar to text files, users should save results of routines that generate graphical outputs (e.g., scatterplot, dendrogram, and tree diagram), if there is a desire to revisit these results at later sessions.

Visualization

The visualize menu item includes three options for visualizing networks (i.e., NetDraw, Pajek, and Draw). As previously mentioned, NetDraw is a built-in companion program used to visualize social networks and is the focus of this section.

By default setting, NetDraw creates graphs using a spring-embedded layout algorithm, but its program features include other graphical layout algorithms such as MDS scaling and principal components as well as optional layouts based on specific node attributes and attributes as coordinates. Users can also visually map node and tie characteristics via NetDraw’s drawing features. This includes features to assign color, size, and shape to node symbols based on specific attributes and line thickness, color, and style to show the tie strength. NetDraw also includes basic analytic routines such as measures of centrality, components, and isolates. In addition, NetDraw enables users to easily explore subsets of the network defined by attributional characteristics of nodes such as Grade 9 teachers as well as by relation type such as advice-seeking ties. Figure 3 provides an example of how these drawing and analytic features can be applied to enhance the visual analysis of networks.

Figure 3 NetDraw sociogram example. This figure shows a NetDraw sociogram depicting the collaboration network among nine sixth-grade teachers of a school. Node shape indicates years of teaching (square = 0–5; diamond = 6–10; and circle = 11–15), *color shows subject taught (lightest gray to black; history, math, English, science, world language)*, and size indicates degree centrality. Adapted from Borgatti (2002).



Network diagrams can be saved in a variety of formats including JPEG and Windows metafile. When saved as a metafile, the network diagrams can be further edited in other programs such as Microsoft PowerPoint and Microsoft Word. To save all network diagram data including matrix, attribute, and other relations created in NetDraw, users should save data as a VNA file, NetDraw’s specialized file format.

Obtaining UCINET and Help Features

UCINET users benefit from a thorough built-in help system where all routine functions are searchable, helping the system provide users with information about the purpose and description of routines, and the system is also available in a downloadable pdf format. UCINET is distributed by Analytic Technologies and is available for purchasable download.

A free, 60-day trial version is also available to those interested in trying out

UCINET before buying. In addition, distributors have a “quick start” guide also available in Analytic Technologies UCINET website. An online help forum is also available for users.

Finally, Borgatti, Everett, and Johnson’s 2013 book titled *Analyzing Social Networks* draws from UCINET to describe social network research design and implementation. This book has a companion website that further assists in explicating practical application of UCINET.

Rachel Darley Gary

See also Correlation; Data Visualization Methods; Descriptive Statistics; Matrices; Matrix Algebra; Quantitative Research Methods; Regression Toward the Mean; Social Network Analysis; Sociometric Assessment

Further Readings

Borgatti, S. P. (2002). NetDraw: Graph visualization software. Harvard, MA: Analytic Technologies.

Borgatti, S. P. (2014). UCINET. In R. Alhajj & J. Rokne (Eds.), Encyclopedia of social network and mining. New York, NY: Springer Science+Business Media.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). Analyzing social networks. Thousand Oaks, CA: SAGE. Retrieved from https://sites.google.com/site/analyzingsocialnetworks/

UCINET Software. Retrieved from http://www.analytictech.com/ucinet/

Iman Ghaderi Iman Ghaderi Ghaderi, Iman

Unitary View of Validity Unitary view of validity

1751

1756

Unitary View of Validity

Validity is a fundamental aspect of evaluation and testing, especially for high-stakes testing for promotion, graduation, and competency. The concept of validity has been the subject of inquiry and debate over the past century. Many authors have studied this complex concept and attempted to define its components. The discussions and controversies about the meaning of validity and its components have continued, and our understanding about validity has substantially evolved. The question in defining validity is whether a test measures the variable for which it was intended.

The early approach toward validity was criterion based, which assumes there is a definitive value for each variable of interest and the goal of assessment is to measure this variable as accurately as possible. Therefore, criterion-based validity is used to determine how well the test scores predict the criterion scores. The main challenge with this approach is that one has to come up with a well-defined and clearly valid criterion measure. In many situations, however, these criterion measures are not readily available. Moreover, the validity of a criterion measure can always be questioned.

In 1954, Paul Meehl and Robert Challman coined the term *construct validity,* which was later expanded by Lee Cronbach and Meehl. At that point in time, the concept of construct validity was an addition to the criterion and content models but not an alternative option. Over the next couple of decades, construct validity became a common approach to validity. Samuel Messick built upon this approach into what is now considered the unitary framework of validity.

Unitary Framework of Validity

The unitary concept of validity evolved from the traditional framework of validity, which consisted of three separate types of validity: content, criterion (including concurrent and predictive validity), and construct. Messick’s main argument for abandoning the traditional view was that the old framework was fragmented and incomplete, and it did not incorporate the value of score meaning and social consequences of score use into the definition of validity. He asserted any validity type in its traditional form, individually or combined with other types, could not address all upcoming questions and queries regarding the various aspects of validity when a test is used in different contexts or situations.

Messick proposed a unifying concept of validity that brings all these types under one umbrella and integrates content, criterion, and consequences to create a comprehensive theory of construct validity to address both score meaning and social values in test interpretation. In this view, validity is an evaluation of the degree to which the theoretical and empirical evidence as well as theoretical rationales obtained through the validation process support the score interpretations for the intended use. It combines scientific evidence with logical argument to justify test interpretation and use. Validation is *a work in progress*, and this continuous and open-ended process is based on the accumulation of evidence to support certain meanings that the user aims to associate with scores from an assessment or test. The construct validation of the test score is not validation of the construct itself and it does not attempt to define the construct. In other words, validity only applies to the scores or their interpretation in a specific context. Therefore, the commonly used terms *valid instrument* and *valid test* are inaccurate. In addition, because validity is a property of inferences, not instruments or tests, validity must be established for each intended interpretation.

In 1999, the unitary definition of validity was endorsed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Since then, this definition has been incorporated in the *Standards for Educational and Psychological Testing*.

Sources of Validity

In the unitary framework, construct validity is the only form of validity. In this approach, validity is a construct with various facets and the validation process requires identification of the relevant sources of validity for these facets. Hence, the phrase *types of validity* has been abandoned and replaced with *sources of validity*. These sources are content, response process, internal structure, relationships to other variables, and consequences of testing. A brief description of each source of validity, with relevant examples, is provided in Table 1.

Content

Content evidence is the relationship between the content of the test and the construct of interest. In order to identify evidence for content validity, a researcher with expertise in the domain of interest should create a blueprint that represents the targeted construct. There should be a logical and empirical relationship between the content of the test and the targeted construct. The expert consensus approach is the most common method, whereby local experts create a blueprint, using their own expertise and input from relevant literature. One can include independent content experts to review the blueprint or use a multicenter design or the Delphi model to incorporate the opinions of a larger pool of experts in different locations to improve the representativeness of the test.

Response Process

Response process is evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible. It entails analysis of responses and accuracy of scoring and reporting of results. The reasoning and thought processes of examinees or learners should be studied to reduce the likelihood of response error because the differences in response processes may result in variance that is irrelevant to the construct being measured. Therefore, instrument items and anchors describing points on the rating scale should be explicit and clear. Response process also includes accuracy of data collection and the process of data entry into a database. When assessment of performance is observational, a trained observer is required to increase the consistency of observations.

Internal Structure

This source of validity evidence relates to the statistical or psychometric characteristics of the test or assessment tool. It is usually referred to as *reliability* and includes reproducibility and generalizability of results. In the unitary framework, reliability is categorized under validity because lack of reliability equals lack of validity. If test scores are not reliable, it is almost impossible to interpret the meaning of the scores. The classic method of assessing reliability is based on the classical test theory. In the classic method, the measurement error is treated as an undifferentiated random variation. It calculates the impact of raters or subjects on scores separately and does not evaluate the interactions between various factors. To address this limitation, Cronbach introduced the generalizability (*G*) theory, which is a statistical method for evaluating reliability of behavioral measurements or test scores. It allows examination of multiple sources of measurement error in order to estimate the impact of each factor in variations of the assessment scores.

Relationship to Other Variables

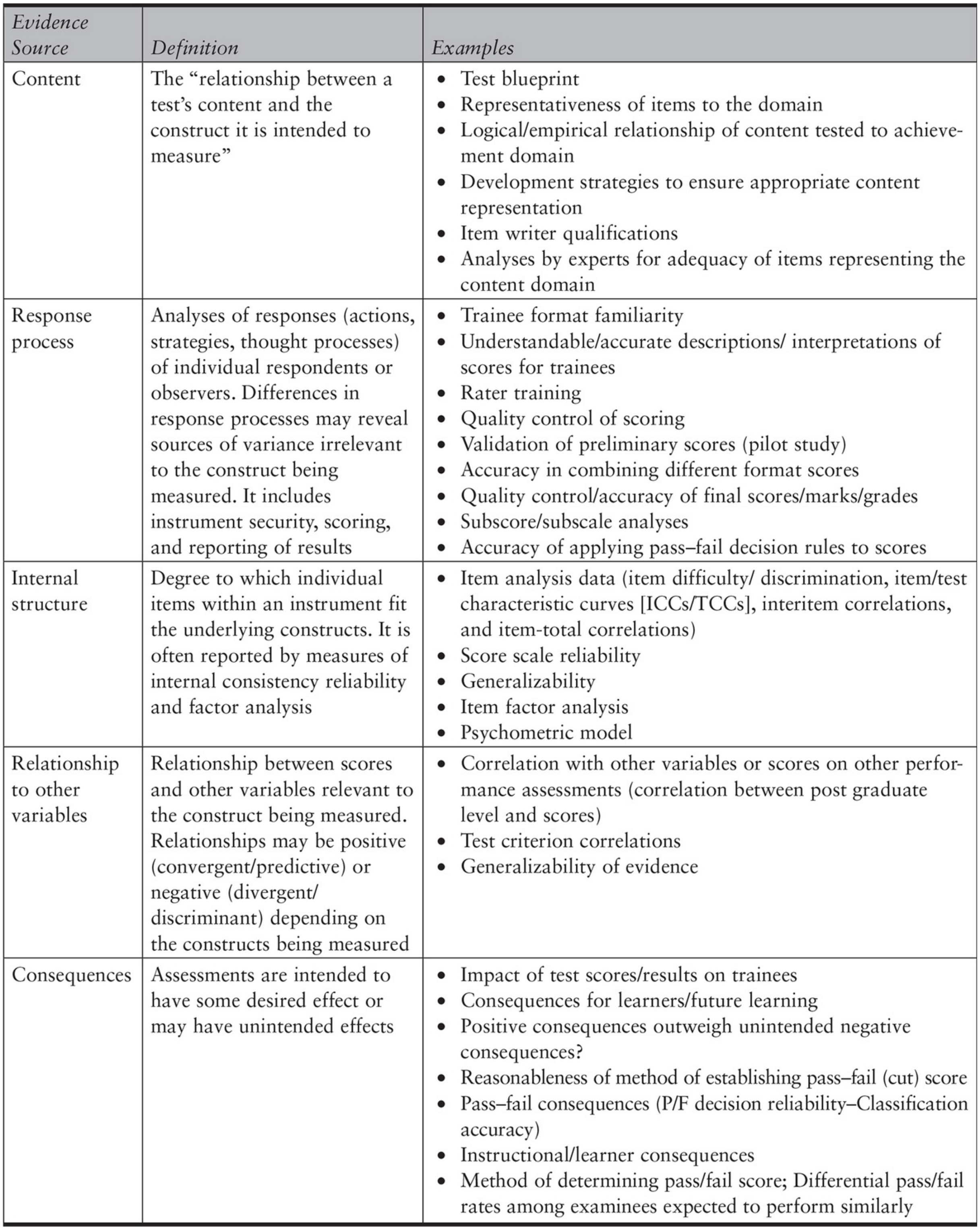
This aspect of validity is the correlation between assessment scores and other variables relevant to the construct being measured. The newer measure is usually validated against external variables. These variables could be criteria that the measure is designed to evaluate such as predicted associations or hypotheses, criteria that predict the measure, or other measures that were designed to measure the same or similar constructs. Therefore, evidence based on relations to other variables refers to traditional forms of criterion-based evidence for validity, such as the correlation between the scores of a test and scores generated by other instruments, other test scores, or an observer’s rating. The old terms such as *convergent*, *divergent* (or *discriminant*), *concurrent*, and *predictive validity* fall under the relationship to other variables in the unitary framework.

Consequences

As noted earlier, consequences of assessments or tests and their social impact are an important aspect of validity in the unitary framework of validity. Consequences can be positive or negative and intended or unintended. The significance of consequences depends on whether the tool is designed for formative assessment (feedback) or summative assessment (pass/fail) and whether it is used for low-versus high-stakes assessments.

Controversies

Although Messick’s interpretation of the validity concept has been widely endorsed, it also has been the subject of criticism. The root cause of this controversy, as Keith Markus explains, lies in synthesis of realism and constructivism with respect to both scientific facts and measurement in this framework. Although union of these two fundamentally different entities has made it appealing, it also gave birth to an entity that harbors conflict by nature. Consequences as a source of validity evidence have been at the center of these debates and criticism. Opponents argue that the consequences of an assessment are beyond the scope of a validity study and should be deferred to policy makers who make decisions about the impact and the appropriateness of its use. The argument is that consequences of a particular use of a test do not necessarily inform us about the meaning of a construct or adequacy of assessment process in measuring the construct of interest. Moreover, the consequences could be political value judgments, which may not provide any information about whether the assessment is a good measure of a construct.



Evidence Source	*Definition*	*Examples*
Content	The "relationship between a test's content and the construct it is intended to measure"	• Test blueprint • Representativeness of items to the domain • Logical/empirical relationship of content tested to achievement domain • Development strategies to ensure appropriate content representation • Item writer qualifications • Analyses by experts for adequacy of items representing the content domain
Response process	Analyses of responses (actions, strategies, thought processes) of individual respondents or observers. Differences in response processes may reveal sources of variance irrelevant to the construct being measured. It includes instrument security, scoring, and reporting of results	• Trainee format familiarity • Understandable/accurate descriptions/ interpretations of scores for trainees • Rater training • Quality control of scoring • Validation of preliminary scores (pilot study) • Accuracy in combining different format scores • Quality control/accuracy of final scores/marks/grades • Subscore/subscale analyses • Accuracy of applying pass–fail decision rules to scores
Internal structure	Degree to which individual items within an instrument fit the underlying constructs. It is often reported by measures of internal consistency reliability and factor analysis	• Item analysis data (item difficulty/ discrimination, item/test characteristic curves [ICCs/TCCs], interitem correlations, and item-total correlations) • Score scale reliability • Generalizability • Item factor analysis • Psychometric model
Relationship to other variables	Relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent/predictive) or negative (divergent/ discriminant) depending on the constructs being measured	• Correlation with other variables or scores on other performance assessments (correlation between post graduate level and scores) • Test criterion correlations • Generalizability of evidence
Consequences	Assessments are intended to have some desired effect or may have unintended effects	• Impact of test scores/results on trainees • Consequences for learners/future learning • Positive consequences outweigh unintended negative consequences? • Reasonableness of method of establishing pass–fail (cut) score • Pass–fail consequences (P/F decision reliability–Classification accuracy) • Instructional/learner consequences • Method of determining pass/fail score; Differential pass/fail rates among examinees expected to perform similarly

Source: Ghaderi *et al.* (2015).

In contrast, advocates of including consequences in instrument validation argue that consequences reflect the soundness of test-based decisions. They believe that the consequences of an assessment can reflect flaws in the conceptualization of the assessment tool and interpretation of the scores. For example, standard setting is a common challenge whereby the examiner has to determine a cut score for the test. Obviously, a cut score must be empirically justified and cannot be arbitrarily defined. In this context, any score and performance data (evidential data) should be used as a means to support expert judgment about competency. The cut score derived from the juncture between contrasting groups (competent students who would pass and incompetent ones who would fail the test) should be used as a starting point for judges to determine an appropriate standard, after adjusting for issues such as the false-positive and false-negative classifications or measurement errors.

Whether one believes consequences of assessment or scores should be part of validity evidence or not, the consequences of each assessment should be carefully examined, and the evidence for its appropriateness of use must be demonstrated.

Concluding Remarks

The dynamic field of validity has benefited from both advancement in science of measurement as well as challenges researchers encounter in the real world, which force them to find practical yet evidence-based approaches using rigorous scientific methods. This provided a productive environment for scholars and experts to put forward their ideas and conceptualize such challenging concepts. This endeavor won’t halt in the foreseeable future.

Iman Ghaderi

See also Conceptual Framework; Construct-Related Validity Evidence; Generalizability Theory; Reliability; Validity

Further Readings

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. American Psychological Association.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. The American Journal of Medicine, 119,166–167.

Cronbach, L. J., Gleser, G. C., Nanda, H., Rajaratnam, N. (1972). The dependability of behavioral measurements. New York, NY: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281.

Downing, S. M., & Yudkowsky, R. (2009). Assessment in health professions education. New York, NY: Routledge.

Ghaderi, I., Manji, F., Park, Y. S., Juul, D., Ott, M., Harris, I., & Farrell, T. M. (2015). Technical skills assessment toolbox: A review using the unitary framework of validity. Annals of Surgery, 261(2), 251–262. doi:10.1097/SLA.0000000000000520

Joint American Educational Research Association. (2009). The standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38(4), 319–342.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635–694.

Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick’s synthesis possible? Social Indicators Research, 45(1–3), 7–34.

Mehrens, W. A. (2005). The consequences of consequential validity. Journal of

Educational Measurement, 16, 16–18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. ETS Research Report Series, 1993(2), 1–18.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741–749.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. Journal of Educational Measurement, 28, 3–9.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. Handbook of Statistics, 26, 81–124.

Lyman L. Dukes Lyman L. Dukes Dukes, Lyman L.

Karla Kmetz Morris Karla Kmetz Morris Morris, Karla Kmetz

Zachary Walker Zachary Walker Walker, Zachary

Universal Design in Education Universal design in education

1756

1760

Universal Design in Education

Universal design in education (UDE) is a framework in which the design and delivery of instructional programming is made accessible for the widest range of students regardless of personal characteristics. UDE (the term is being used generically to refer to all UDE applications) emphasizes that both the design of the course *and* the pedagogy, that is, the teaching methods employed, are as barrier free as possible. UDE stresses that diversity is considered a typical aspect of the human condition; therefore, it is a necessary aspect of instructional planning and implementation. This does not mean potential barriers will cease to exist, simply that the use of accommodation and modification can be significantly reduced if UDE is employed. This entry describes the origins of universal design (UD) and the rationale for and application of UDE. The entry concludes by noting various UDE approaches and provides an in-depth description of universal design for learning (UDL), which is perhaps the most well-known UDE framework. UDE is offered as a tool for addressing diversity in educational contexts whose goal is including the widest range of learners.

Origins of UD

The term and idea of UD, formulated by Ron Mace, an architect and professor at North Carolina State University, promote the notion that both products and the built environment apply esthetic and usability objectives from the outset of the design and development process. The intent is to create products and, in particular, physical spaces that allow for use by the greatest number of individuals regardless of ability. Mace himself was a wheelchair user and he had a particular passion for design that did not require subsequent adaptation for people with disabilities. A number of societal changes set the stage for the movement to adopt barrier-free concepts, including UD. These include innovations such as adaptive and assistive technologies, legislative mandates such as the Americans with Disabilities Act, and societal demographic changes that have resulted in longer life expectancy and, subsequently, more people with impairments later in life.

The Center for Universal Design at North Carolina State University explicates the following seven UD principles: (1) equitable use, (2) flexibility in use, (3) simple and intuitive, (4) perceptible information, (5) tolerance for error, (6) low physical effort, and (7) size and space for approach and use. Many examples of UD are now a common part of daily life, including curb cuts that allow wheelchair users or persons with strollers greater accessibility, public transit equipped with ramps that provide access for people using mobility aids, and closed captioning for persons with hearing impairment or someone viewing television in a loud or distracting environment.

Why UDE

The school population is in a state of flux. Students now represent a variety of ethnic backgrounds, language learning abilities, and documented or undocumented impairments while others are classified as gifted. Indeed, all learners in today’s schools have an array of learning characteristics and preferences. It is commonly reported that students with disabilities spend more than 80% of the school day in general education settings. Moreover, there is an increased emphasis on the value of inclusive instructional environments. Currently, many general education teachers are not trained to educate students with unique needs.

These realities have created both opportunities and challenges for teachers in today’s classroom. For example, classrooms that include a diverse student body promote collaboration among school professionals and also provide students the opportunity to interact with a diverse group of peers. However, challenges exist, with the most significant likely the lack of suitable professional preparation for meeting the diverse needs of students in today’s classrooms. Many experts believe that educational outcomes for all students may be improved through the application of UDE.

The Concept of UDE

The goal of UDE is to make teaching and learning accessible to students with a diverse array of abilities. The principles of UDE spell out that course designers assess the needs of their students and prepare lessons that account, in advance, for any potential barriers to learning. Instructors are encouraged to utilize resources and apply UDE principles to build in supports where needed, with the goal of providing all students a barrier-free path to achieve their respective learning objectives. Ideally, such a teaching and learning approach eliminates the need for the majority of retroactive accommodations, modifications, and ancillary services often used by students with diverse learning needs.

UD Frameworks in Education

Three of the most notable UDE frameworks are universal design of instruction (UDI), universal instructional design (UID), and UDL. Each of these frameworks draws its inspiration from the original architectural movement, yet each has expanded beyond the product and physical access considerations to include principles that support access to the general curriculum for the widest range of students. Both UDI and UID are applied in postsecondary settings, whereas UDL is most often applied in secondary settings. Nonetheless, each respective design has more in common than not. All models focus on the needs of diverse learners and attempt to address their educational goals using flexible resources and pedagogy when designing and delivering instruction. Each reflect, in some general fashion, the original seven UD principles but apply them to educational environments.

UDI is the proactive use of design and instructional strategies that promote the inclusion of a wide range of learners including learners with disabilities. It has been primarily associated with postsecondary education. It consists of nine principles, the first seven of which were adapted from the UD principles formulated by Mace for products and the built environment. Its principles are as follows: (1) equitable use, (2) flexibility in use, (3) simple and intuitive, (4) perceptible information, (5) tolerance for error, (6) low physical effort, (7) size and space for approach and use, (8) a community of learners, and (9) instructional climate.

UID focuses exclusively on the application of its principles in postsecondary educational settings. Although its seven principles read differently, like other UDE approaches, they are intended to maximize learning opportunities for all students and minimize the need for accommodations. Its seven principles are as follows: (1) be accessible and fair; (2) provide flexibility in use, participation, and presentation; (3) be straightforward and consistent; (4) be explicitly presented and readily perceived to be supportive; (5) provide a supportive learning environment; (6) minimize unnecessary physical effort or requirements; and (7) ensure learning spaces that accommodate both students and instructional methods.

UDL, which is highlighted in the following sections, was developed with an emphasis on the K–12 educational system, with a focus on the traditional classroom setting and the three primary learning networks of the brain.

The UDL Framework

The Center for Applied Special Technology, better known as CAST, formed in 1984 to specifically examine how technology could impact learning and education for those with disabilities. Today, UDL, which was developed by CAST, is the framework for all of the organization’s research and development. Following CAST’s lead, scholars and practitioners have broadly characterized the UDL principles in terms of providing options for students with regard to how they take in information (representation), how they practice new content (engagement), and how they express what they know (expression). The key to UDL is considering these options before teaching and learning commences. UDL suggests that curriculum design should be proactively constructed to meet the needs of diverse learners instead of applying reactive curriculum adaptations when lesson objectives are not achieved.

The three principles of UDL are as follows: 1. *Providing multiple means of representation*. Individuals learn in a variety of ways. This includes how they perceive and understand the content that is presented to them. Information presentation (inputs) can make the difference between comprehension and confusion. For example, learners with sensory disabilities (e.g., blindness or deafness), learning disabilities (e.g., dyslexia), or language or cultural differences may receive content in different methods. There are learners who understand information efficiently through visual or auditory means, whereas others may learn best with traditional printed text. All people benefit when content is represented in multiple ways as the brain has the opportunity to receive and process multiple stimuli. When information is presented in various ways, students have the opportunity to make connections within, as well as between, concepts. In short, because there are multiple learning styles, it is worthwhile to provide information in several formats so educators do not limit learning to one type of learner.

One way to provide multiple means of representation is to allow multiple sources of content. For example, when presenting a biology lesson on a plant type, the teacher can provide inputs from the standard text (e.g., a book, an article), tools or resources that include the same content with a text-to-speech option, videos about the content, audio files, and plant samples that can be physically manipulated. Providing these options allows learners to gain knowledge through a variety of experiences, including the method that may best meet their learning preferences.

2. *Provide multiple means of action and expression*. Like learning inputs, people communicate and express what has been learned in different ways. For example, individuals with significant physical impairments (e.g., cerebral palsy), those who struggle with planning and choice making (executive function disorders), and those who have language barriers due to disability or are second-language learners may think about, manipulate, and articulate content very differently. Students with autism spectrum disorder may express themselves well in writing but not feel comfortable giving a speech. A student with dyslexia may not do well when writing a paper but can deliver a rousing presentation. Action and expression require a great deal of strategy, practice, and organization, and it is important to give all learners exposure to many different forms of expression while allowing them to choose and practice what they feel is most important. There is not one means of action and expression that is optimal for all learners, so it is essential to provide multiple options for conveying learning.

One way to provide multiple means of action and expression is to give multiple forms of assessment. For example, by studying world conflicts, some students may want to write a paper, some may want to create a video, and others may want to give a speech. All of these options allow learners to express their learning in ways that they perceive as comfortable while also allowing the teachers to assess understanding of the lesson objective.

3. *Provide multiple means of engagement*. The background that a student brings to the learning environment is a critical element of acquiring the information presented. Based on past experience and styles of learning, people’s opinion varies greatly as to what promotes learning. Neurology, culture, personal relationships, foreseen relevance, subjectivity, and previously learned knowledge, all impact a learner’s motivation. While some learners are highly engaged by the spontaneity and novelty of an active classroom, other learners prefer a quiet environment and strict routine. Students with social deficits may prefer to work alone or may not possess the social skills to work in groups, while others may prefer to collaborate with their peers on projects. It is essential to design lessons taking into account the needs of a multitude of learners as well as the desired lesson outcomes.

As with the first two principles, providing options for students is a fundamental component of UDL. Relationship building in the classroom will allow teachers to develop an understanding of student preferences and learning needs. Learning tasks can be arranged so students have options of working alone or in groups. For example, if there are three learning projects throughout the year, teachers can require students to do one in a group, one individually, and one in which they can choose between working in a group or alone. As with all the examples, choice is a powerful motivator in its own right and allowing choices in the learning environment is especially beneficial for learners.

UDL Guidelines

Each UDL principle is also broken down into guidelines for planning. For example, providing multiple means of representation includes the following guidelines: (a) providing options for comprehension; (b) providing options for language, mathematical expressions, and symbols; and (c) providing options for perceptions. Providing multiple means of action and expression includes the guidelines: (a) providing options for executive functions, (b) providing options for expression and communication, and (c) providing options for physical action. Finally, providing options for engagement includes the guidelines: (a) providing options for self-regulation, (b) providing options for sustaining effort and persistence, and (c) providing options for recruiting interest. These guidelines are meant to assist teachers as they lesson plan. All guidelines are intended to provide students options so that learning preferences are addressed and each learners’ abilities and choices are valued.

UDL and Technology

The impact that UDL and technology can have on learning design has long been a topic of significant interest to key stakeholders in education, including policy makers and administrators. However, UDL is more than simply providing assistive technology for learners or allowing the use of devices in the learning environment. Instead, UDL is considered an instructional design approach, and technology is one potential element of a UDL-driven lesson.

There are some exciting applications of technology in the context of UDL. Mobile technology, for example, provides exciting opportunities when considered in light of UDL. One of the most promising aspects of the proliferation of mobile devices is that many contain built-in features such as speech-to-text transcription, hearing aids, voice-over capability, subtitles and captioning, among others. Many of these instructional tools provide educators with additional options for applying instructional interventions for students with diverse learning needs. Developing best practices integrating technology into the UDL framework is challenging, as it requires a multifaceted approach combining theory, research, practice, policy, and innovation. However, it has the potential for great educational benefit.

The Status of Research on UDE

There are significant interests in the application of UDE in learning environments in recent years. Remarkably, evidence supporting its efficacy is not in significant supply. Scholars, however, are moving toward rectifying this circumstance. Certainly, recently reported evidence indicated that, at least on a preliminary basis, the application of UDE has resulted in positive educational outcomes. Researchers are beginning to make a case for the value of UDE but studied a variety of UDE approaches (i.e., UDL, UDI, and UID), applied an array of research methods, and rarely were able to establish causality when drawing conclusions. This has resulted in few, if any, UDE practices that can be defined as scientifically valid.

Nonetheless, it is worthwhile to point out the progress made. For example, studies in which UDE methods are applied show an improvement in student academic performance (e.g., math, science) and content accessibility for students with reading problems. Regardless of disability status, findings, in many cases, were true. That is, studies taking into consideration the disability status, it was were true. That is, studies taking into consideration the disability status, it was determined that all students, with or without disability, benefited from the use of UDE methods. With schools now obliged to adopt programming based upon rigorous research, UDE advocates will continue to move toward empirically validating their methods.

Lyman L. Dukes, Karla Kmetz Morris, and Zachary Walker

See also Classroom Assessment; Instructional Theory; Learning Maps; Learning Styles; Learning Theories; Least Restrictive Environment; Special Education Law; Technology in Classroom Assessment

Further Readings

Burgstahler, S. (2013). Websites, publications, and videos. In S. Burgstahler (Ed.), Universal design in higher education: Promising practices. Seattle: DO-IT, University of Washington. Retrieved from www.uw.edu/doit/UDHE-promising-practices/resources.html

Center for Applied Special Technology. (2012). What is universal design for learning. Retrieved from: http://www.cast.org/udl/index.html

Center for Universal Design: Environments and products for all people. Retrieved from https://www.ncsu.edu/ncsu/design/cud/about_us/usronmace.htm

Edyburn, D. L. (2010). Would you recognize universal design for learning if you saw it? Ten propositions for new directions for the second decade of UDL. Learning Disability Quarterly, 33(1), 33–41.

Hall, T. E., Meyer, A., & Rose, D. H. (Eds.). (2012). Universal design for learning in the classroom: Practical applications. New York, NY: Guilford Press.

Israel, M., Ribuffo, C., & Smith, S. (2014). Universal design for learning: Recommendations for teacher preparation and professional development

(Document No. IC-7). University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center. Retrieved from http://ceedar.education.ufl.edu/tools/innovation-configurations/

McGuire, J. M., Scott, S. S., & Shaw, S. F. (2006). Universal design and its applications in educational environments. Remedial and Special Education, 27, 166–175. doi:https://doi.org/10.1177/07419325060270030501

Palmer, J., & Caputo, A. The universal instructional design implementation guide. Retrieved from http://www.cer.jhu.edu/pdf/uid-implementation-guide-v6.pdf

Rao, K., Ok, M. W., & Bryant, B. R. (2014). A review of research on universal design educational models. Remedial and Special Education, 35, 153–166. doi:https://doi.org/10.1177/0741932513518980

Rao, K., & Tanners, A. (2011). Curb cuts in cyberspace: Universal instructional design for online courses. Journal of Postsecondary Education and Disability, 24(3), 211–229. doi:http://dx.doi.org/10.1080/02680513.2014.991300

Roberts, K. D., Park, H. J., Brown, S., & Cook, B. (2011). Universal design for instruction in postsecondary education: A systematic review of empirically based articles. Journal of Postsecondary Education and Disability, 24(1), 5–15. Retrieved from http://files.eric.ed.gov/fulltext/EJ941728.pdf

Bruce B. Frey Bruce B. Frey Frey, Bruce B.

Universal Design of Assessment Universal design of assessment

1760

1765

Universal Design of Assessment

The concept of universal design began as an architectural and engineering philosophy, but it has spread to education in terms of lesson design and educational measurement. The idea that one should design products and environments to be usable in a meaningful and similar way by all people was popularized by architect Ron Mace, who developed the first U.S. state building accessibility code in North Carolina in the 1970s. The underlying assumption of universal design is that all aspects of our world can be planned from the beginning to allow access and use by everyone. As a coherent philosophy and set of guidelines, universal assessment is only a few decades old. Both classroom teachers and standardized test developers have begun to explore design of assessments that work equally well for every student regardless of their characteristics. Those who use the jargon of measurement would say, more specifically, that assessments should work equally well for every student regardless of their *construct-irrelevant characteristics*. That is, universal design of assessment provides guidelines for developing assessments that are equally valid for all students. This entry reviews the established standards for universal design and examines the application of these standards to classroom assessments, discusses the validity of universal design for educational assessments, and details the steps involved in developing an assessment based on universal design.

Application of Universal Design Standards to Classroom Assessments

There are seven broad, established standards for universal design: (1) equitable use, (2) flexibility in use, (3) simple and intuitive use, (4) perceptible information, (5) tolerance for error, (6) low physical effort, and (7) size and information, (5) tolerance for error, (6) low physical effort, and (7) size and space for approach and use. Each of these general standards has been interpreted in the more specific context of educational measurement.

According to the work of Sandra Thompson and colleagues, the goal of universal design in assessment would be to allow participation of the widest range of students and to produce valid inferences about performance for all students who participate in the assessment. Although no assessment will be completely accessible or valid for all, the objective is to be as inclusive as possible. Some of these interpretations match directly to the broad seven standards and some do not: (1) inclusive assessment population; (2) precisely defined constructs; (3) accessible, nonbiased items; (4) amenable to accommodations; (5) simple, clear, and intuitive instructions and procedures; (6) maximum readability and comprehensibility; and (7) maximum legibility.

So what do these standards look like when they are applied to classroom assessments? A teacher-developed classroom assessment or a standardized commercially produced test built under the philosophy of universal design may on its face look somewhat similar to an assessment from 20 years ago, although there will likely be some noticeable technical differences (e.g., larger fonts, more white space), and the wording of directions and questions may be simplified. More significant differences, however, are likely to be in the choice of tasks, questions, administrative procedures, and in the planning.

Each universal design principle has been translated by educational researchers into concrete implications for assessments: 1. *Inclusive assessment population*. Assessments provide opportunity for participation for all members of the target population regardless of physical characteristics, culture, linguistic background, or cognitive abilities. This information is difficult to “observe,” but most teacher-developed assessments are consistent with this principle.

2. *Precisely defined constructs*. Performance should not be affected by construct-irrelevant variance, processes that are extraneous to the intended construct. Points are awarded for knowledge or performance, not construct-irrelevant tasks (e.g., speed, handwriting, perhaps spelling and grammar). The wording for math problems, particularly, should be simple and clear.

3. *Accessible, nonbiased items*. Items are probably biased if groups of equal ability have different probabilities of answering the questions correctly. Items also should be free of culturally offensive content. Teachers and other

assessment developers should only use words, phrases, and concepts that are commonly used across cultures and languages. Pop culture references (e.g., television, music) should be avoided, and there should be no reference to stereotypes or offensive terms.

4. *Amenable to accommodations*. The way in which a test is presented can easily be changed to remove unintended disadvantages for English language learners or for those with disabilities. Characteristics that tend to make things easier for these populations and others use horizontal text, avoid construct-irrelevant graphs and pictures, and keep the graphics simple and clear. Keys and legends if necessary appear at the top of the page or screen or right of the item. There should be no time limits and the different subsections of tests should be independent of each other.

5. *Simple, clear, and intuitive instructions and procedures*. Directions and procedures should be easy to understand for all students, regardless of their culture or language skills. Format and instructions should be consistent (e.g., circling correct answers). One observable way to tell whether instructions are clear is that students can work independently without asking the teacher questions. Practice or sample items should be provided, and all the questions should be numbered.

6. *Maximum readability and comprehensibility*. Plain language and well-constructed sentences should be used for items and directions. Questions should be clearly framed. Verbal and organizational complexity should be minimized. Universally designed assessments use simple, clear, and common words and avoid unnecessary words. Technical terms should be clearly defined. Sentences should be short and not compound with an obvious link between the nouns and pronouns and verbs. If there are multiple steps in the directions, a clear and, perhaps, numbered sequence should be used.

7. *Maximum legibility*. All parts of the test should be visible without distraction. This applies to tables, figures, and graphics as well as the questions and directions. Legible tests have high contrast, large font size, and much “white space.” White space is empty single-color space around the elements of an item. For paper tests, the paper should be off-white in color. A good rule of thumb is that the page be at least 50% blank or empty. The type should be black. In terms of technical specifics, gray-scale shading should be avoided; the font should be at least 10 point and graphic text at least 12-point font. Purposeful bolding is ok

with standard use of upper and lower case. Text should be unjustified or “jagged.”

Validity

There are only a few research studies on the effect of universal design for educational assessments. Recommendations for the approach are driven primarily by theory and philosophy. That does not mean that universal design will not increase the usefulness of educational assessments and allow for fair access to tests by more populations, it only means that it has not been studied much and the field does not yet know for sure whether it makes a difference regarding assessment and learning. The few studies of universal design principles have to do with their positive effect on standardized test performance. Much of the research effort in universal design of assessment has focused on standardized state tests in the United States because of the U.S. federal mandate to include all populations in statewide testing. Presumably, if application of universal design guidelines positively affects performance on standardized tests, the same applications should also increase performance on classroom assessments designed by teachers.

A second line of research related to universal design is the willingness of teachers to buy in to the philosophy and apply its principles. In 2011, Allison Lombardi and Christopher Murray conducted a large survey of college faculty at a university about their attitudes toward the principles and instructional behaviors and expectations consistent with universal design. The researchers found that teachers who were female, newer on the job, or had been trained to teach students with disabilities felt much more positively than their peers toward minimizing barriers, adjusting assignments and requirements, providing easier access to course materials, and other universal design characteristics.

A fairly modern validity concern with teacher-made, or standardized, tests is the validity of inferences made from such tests for students whose first language is not English. A second somewhat more traditional concern is the validity of these assessments for students with disabilities. Universal design is meant to respond to those validity concerns by producing assessments that not only fairly assess those students but fairly assess all students regardless of their irrelevant characteristics.

Beyond these concerns, though, there is a modern concept of validity that is frequently cited by supporters of the universal design approach. This aspect of validity is known as social consequences validity or *consequential validity*. The usefulness of an assessment is not only whether the test score accurately represents a particular domain of knowledge or skill but includes whether the use of an assessment is *fair* and *just* in a social sense. In 1993, Samuel Messick, a measurement philosopher, first suggested this idea of consequential validity. He pointed out that an assumption underlying the broad concept of validity is that tests should serve the purposes for which they are intended. If a teacher, school system, or state believes that the use of an assessment will ultimately help those involved by improving instruction, for example, or by increasing student learning, that intent becomes part of the validity requirement for the assessment. Messick argued that judging validity in terms of whether a test does the job it is employed to do—that is whether it serves its intended function or purpose—requires evaluation of the intended or unintended social consequences of test interpretation and use.

The educational and psychological measurement field incorporated Messick's arguments into the modern definition of validity as the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Assessment developers with this view of validity are often concerned with the instructional time taken up by assessments, the effects of labeling on students, whether tests are biased, and other issues regarding the consequences on students from assessment. The underlying argument for the need for universal design considerations is that traditional assessment scores may represent something a bit different for each student. If some of the variability in scoring is construct-irrelevant variance, then the validity of those scores is questionable. If assessments are designed from the beginning so that all items are free from cultural bias, all students understand directions, all students can read and comprehend all items, and all students are capable of performing all assessment tasks, then construct-irrelevant variance is minimized.

The application of universal design of assessment primarily improves the quality of measurement through improved validity, but reliability may be improved as well. One benefit of universal design is that to some extent it should have relevance to inter-rater reliability concerns. The level of subjectivity in any scoring system affects inter-rater reliability, and one source of subjectivity is bias. Evaluating unexpected responses or dealing with task performance that does not seem to meet assessment instructions or requirements is difficult.

Responses will be more uniform when directions and tasks are described using text that is easily understood by all students. The range of performances should more closely match the rubric categories and expectations when the assessment is planned from the start following universal design guidelines. So, one might expect less subjectivity in scoring when this modern approach is followed.

The science of universal test design has to do with the physical characteristics of a test that follow the key principles. The art of universal test design comes into play in the actual writing of an assessment. It is word choice in items, directions, and the terms used on an assessment that may lead to construct-irrelevant variance in the scores for some students. Fortunately, researchers have suggested guidelines to follow when composing items and assessment tasks and when formulating directions.

In 1987, Stephen Rakow and Thomas Gee provided general guidelines for knowing whether the content of an assessment follows universal design principles and allows “access” to all students, in their suggestions for improving readability in assessments:

1. All students would likely have the experiences and prior knowledge necessary to understand the question.
2. The vocabulary, sentence complexity, and required reasoning ability are appropriate for all students’ developmental levels.
3. Definitions and examples are clear and understandable.
4. Relationships are clear and precise.
5. Item content is well organized.
6. The questions are clearly framed.
7. The content of items is of interest to all students.

The wording used in assessments can make a difference, and there are a variety of ways to ensure that assessments are written in plain language: First, shorten the length of sentences wherever possible. Reduce needless wordiness and irrelevant text; break complex sentences into several shorter sentences. Second, unless it is important to use the jargon of a field, replace unusual words with more common synonyms. Third, be consistent across assessments and within each assessment. Use the same word for an important concept each time you use it. Finally, number or identify in some way each question.

P. J. Brown found that when students actually know the answers or have the assessed skill, they perform higher on plain language tests, but that performance was not affected for those who did not have the knowledge or skill. This is a good indication that the use of plain language tests affects only the construct-irrelevant variance in performance and increases validity. It increases fairness without disadvantaging any students.

Developing Assessments Using Universal Design

In 2005, Leanne Ketterlin-Geller provided a detailed example of procedures for developing classroom assessments that follows the principles of universal design. Although her example is specifically for designing a computer-enhanced assessment, the principles and applications generalize well to traditional paper-and-pencil tests designed by teachers. The assessment that is described is a third-grade math test. As the author points out, many of the procedures and development strategies used in this test are similar to those for other classroom assessment approaches. The difference is the deliberate consideration of individual needs along the way.

Step 1: Identify and Define the Construct

What skill, ability, attitude, or knowledge domain is meant to be assessed? In Ketterlin-Geller’s example, the construct was mathematical ability. More specifically, the construct was the knowledge and skills identified as standards for the third grade in the state in which the assessment was developed. These were measurement concepts, geometry, probability, statistics, algebra concepts, calculation skill, and estimation skill.

Step 2: Identify and Define the Population

In this example, the population was all third graders. This population included students with a wide variety of disabilities, linguistically diverse students, and students with a wide variety of cultural characteristics and cognitive abilities.

Step 3: Choose the Testing Platform

Will the testing platform be traditional paper-and-pencil, performance assessment, computer-based, or some other assessment environment? At this step, the designers decided that they wanted flexibility in the level of support (e.g., practice items, navigation options, concentration aids, text-to-speech capability) and chose a computer environment.

Step 4: Choose the Item Format

Ketterlin-Geller and colleagues wished to use a traditional multiple-choice format. To increase reliability, they used five answer options instead of four (this reduces the likelihood of randomly guessing the correct answer). A left to right layout was chosen (question on left; answer options on right). Answer options were vertical, one beneath the other, which is consistent with universal design guidelines. Because the answer options would be indicated on a computer screen, they did not need to be labeled with As, Bs, Cs, and so on. Because some students might have physical disabilities, more than one way of indicating the correct answer was available (using a mouse or the keyboard). So that difficulties with attention and concentration were less likely to affect performance, the interface was designed so students could select an answer and review it as long as they wished before submitting it.

Step 5: Compose and Sequence the Test

This computerized, third-grade math test was written so that directions, prompts, and questions were simplified (the text is easy, not the difficulty level). In an example item provided, a two-color graphic is shown of 11 circles. Each circle is either striped or has a crossed-lines pattern (cross hatch). Four of the circles are striped. The question is worded in a straightforward manner without superfluous text: “What is the probability of picking a striped ball?” Because the question is designed to assess understanding of probability concepts, and not geometry terms or anything else, the simpler word *ball* can be used instead of circle. The word *probability* should be used instead of a simpler word, though, because it is terminology central to the targeted skill. Answer options are succinct and only provide information necessary to answer the question (e.g., “4 of 11”).

Step 6: Finalize Accommodation Options

This example had built-in accommodation options, such as text-to-speech options were available by clicking on a “speaker” icon. Students could listen to a question or directions as often as they wished. An alternative form was available with the same math questions in an even more simplified format. Access to the alternative form was automatic based on a brief pretest screening of sorts which assessed reading ability.

The author emphasizes that though this particular case example used computers for administration, the principles applied here can also be applied to traditional paper-and-pencil teacher-made classroom assessment. This is true, of course, as most universal design "rules" apply to wording of items, the layout of test components, and the upfront careful definition of the intended construct for assessment.

Bruce B. Frey

Adapted from Frey, B. B. (2014). Universal test design. In B. B. Frey, *Modern classroom assessment* (pp. 235–262). Thousand Oaks, CA: Sage.

See also Authentic Assessment; Paper-and-Pencil Assessment; Test Bias; Validity

Further Readings

Brown, P. J. (1999). Findings of the 1999 plain language field test. Newark: Delaware Education Research and Development Center, University of Delaware.

Frey, B. B. (2014). Universal test design. In B. B. Frey, Modern classroom assess (pp. 235–262). Thousand Oaks, CA: Sage.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. The Journal of Technology, Learning and Assessment, 4(2), 1–23.

Lombardi, A. R., & Murray, C. (2011). Measuring university faculty attitudes toward disability: Willingness to accommodate and adopt Universal Design principles. Journal of Vocational Rehabilitation, 34(1), 43–56.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.). Washington, DC: American Council on Education.

Rakow, S. J., & Gee, T. C. (1987). Test science, not reading. Science Teacher, 54(2), 28–31.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report). Retrieved from ERIC Document Reproduction Service. (ED467721).

Melissa N. Richards Melissa N. Richards Richards, Melissa N.

U.S. Department of Education

U.S. department of education

1765

1766

U.S. Department of Education

Located in Washington, DC, and in regional offices throughout the United States, the U.S. Department of Education plays a large role in shaping educational policies and programs in the United States through impacting the actions of researchers, parents, community practitioners, and policy makers. The agency supplements and supports state and local school systems to ensure that all U.S. citizens receive access to educational resources. This entry describes the department’s history, duties, and organization.

History

Although the U.S. Department of Education was not formally created until 1980, the United States has collected systematic data about schools and teachers and used this information to shape policy since the 1860s. Prior to the creation of the department, a bureau known as the Office of Education distributed federal funding to support educational programs and also managed research initiatives. Throughout the first half of the 20th century, additional responsibilities were assigned to the office, including managing educational standards for colleges and universities as well as vocational coursework (e.g., home economics and agriculture training).

Wartime events also impacted the historical role of the Office of Education. The post–World War II years saw an increase on federal spending on education. Approximately eight million World War II veterans returned home and used funding they received from the GI Bill to attend college. During the Cold War, after the Soviet Union launched Sputnik, Congress allocated money to support the college-level scholarship of advanced science and engineering. Likewise, at the secondary and elementary school levels, additional monies were allocated to ensure quality instruction in foreign languages, science, engineering, and math.

The role of the federal government began to change again in the 1960s and 1970s, with a new emphasis on ensuring that all U.S. citizens had equal access to education. The passage of laws during this time period made race, sex, and disability discrimination illegal in schools and educational institutions. Furthermore, federal financial aid programs were established for children from disadvantaged backgrounds as well as college students seeking assistance with tuition payment.

The U.S. Department of Education was formally established as a cabinet-level agency in 1980. Today, the U.S. Department of Education oversees millions of students attending elementary, secondary, postsecondary, and vocational schools.

Duties

The U.S. Department of Education engages in four core tasks:

1. Creating policies that determine the allocation of federal funds for education. The Department is responsible for supervising the distribution and use of the funds as well as administering and organizing the programs that the funds support.
2. Conducting research about education in the United States through the collection and analysis of data and distribution of the findings to citizens, educators, and policy makers. The department works with communities to formulate solutions to difficult issues through the use of this information.
3. Detecting weaknesses in the education system and making them a nationwide focus by raising awareness about them.
4. Enforcing protections against discrimination to guarantee that all citizens receive an equal opportunity to utilize educational programs and services.

When completing these duties, the Department of Education strives to improve coordination and management of educational programs that receive federal funding to ensure accountability of these programs to U.S. citizens.

Organization

Organization

The U.S. Department of Education is organized into nine separate program offices that work to achieve the agency’s mission:

Institute of Education Sciences is the research arm of the agency. Institute of Education Sciences conducts rigorous research that produces a strong foundation of information on which to base future policies and education practice.

Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students support the academic achievement of those who are learning English as a second language and the creation of policies that support curricula to foster the education of these students.

Office of Elementary and Secondary Education aims to improve education at the state and local level through providing assistance to stakeholders in those communities. Funding is provided to support students from preschool to secondary school.

Office of Innovation and Improvement organizes and evaluates initiatives that develop innovations in education.

Office of Postsecondary Education creates policies and programs that support postsecondary education. In particular, the office maintains programs that encourage international scholarship and foreign language study.

Office of Safe and Drug-Free Schools manages initiatives that encourage students’ physical and mental health in educational institutions such as drug prevention and character building programs.

Office of Special Education and Rehabilitative Services administers programs that aid children and adults with disabilities to help them reach their full potential in an educational setting.

Office of Federal Student Aid directs federal financial aid to individuals pursing postsecondary education. This office provides information to parents, students, and administrators about the logistics of applying for loans.

Office of Vocational and Adult Education supervises programs that equip adults with the skills needed to complete educational coursework. In particular, this office works to ensure that adult students have the abilities needed to attain a high school diploma.

Melissa N. Richards

See also Federally Sponsored Research and Programs; Institute of Education Sciences; Office of Elementary and Secondary Education

Further Readings

Radin, B. A., & Hawley, W. D. (1988). The politics of federal reorganization: Creating the U.S. Department of Education. Elmsford, NY: Pergamon Press.

U.S. Department of Education. (n.d.). About ED. Retrieved from http://www2.ed.gov/about/landing.jhtml?src=ln

U.S. Department of Education. (2010). Overview of the U.S. Department of Education. Washington, DC: U.S. Department of Education Office of Communications and Outreach. Retrieved from http://www2.ed.gov/about/overview/focus/what.pdf

Amy S. Gaumer Erickson Amy S. Gaumer Erickson Erickson, Amy S. Gaumer

Patricia M. Noonan Patricia M. Noonan Noonan, Patricia M.

Utilization-Focused Evaluation Utilization-Focused evaluation

1766

1770

Utilization-Focused Evaluation

Utilization-focused evaluation is an evaluation framework developed by Michael Quinn Patton that focuses on intended use and intended users throughout development and implementation, with the premise that evaluation should be judged on the actual use of the results. Evaluations are designed and implemented in close collaboration with primary intended users (i.e., the people who will use the results), with the intent that these users feel ownership of the evaluation process and findings, therefore, increasing the likelihood that these users will employ evaluation results in decision making. This is a shift from the evaluator being the primary decision maker in the evaluation to facilitating the decision making of the intended users. This entry first reviews the characteristics of utilization-focused evaluation and then details the steps in conducting such an evaluation.

Characteristics of Utilization-Focused Evaluation

The utilization-focused approach is context specific. It can include any evaluation purposes, theory, model, design, or data. Within this framework, evaluators work to understand the specific situation, intended users, and evaluation purposes. They also build the capacity of the primary intended users to make evaluation decisions. The evaluators then guide these users in determining the evaluation's questions and design, matched to the specific context for the specific purpose of the evaluation. At the same time, the evaluators adhere to professional principles and attend to the evaluation's accuracy and feasibility.

Utilization-focused evaluation is both a comprehensive philosophy and a pragmatic approach. Patton published his first book with this title in 1978 and has since documented numerous evaluations adhering to this approach. His approach to evaluation emerged in a time when numerous researchers were characterizing the impact of evaluation as unsuccessful because the results were not being used. Patton’s research found that evaluations were more likely to affect change when individuals took direct, personal responsibility for making or advocating for decisions based on the results. In evaluations that did not include this ownership component, the results did not impact organizational change.

Conducting Utilization-Focused Evaluation

Patton has developed a 17-step process for conducting utilization-focused evaluation. These steps provide a deeper understanding of the utilization-focused evaluation process. It is important to note that, while these are organized as steps, they do not necessarily represent a linear process. Some steps are undertaken together while other steps must be addressed throughout the entire evaluation process.

Step 1. Assess and build program and organizational readiness for utilization-focused evaluation. Evaluators must strengthen the capacity for primary intended users to understand and value evaluation by learning about their organizational culture, perceptions of evaluation, and what they hope to gain from the evaluation. For example, the evaluator could ask a series of questions to gain an understanding of various perspectives of the organization and stakeholder groups, including their past experience with evaluation. Based on the responses, the evaluator might need to conduct a workshop to build evaluation capacity or support the stakeholder groups to develop a shared vision for the evaluation.

Step 2. Assess and enhance evaluator readiness and competence to undertake a utilization-focused evaluation. Evaluators need to establish credibility by demonstrating both technical competency to conduct the evaluation (e.g., identifying methods and measures, collecting valid data, and reporting accurate results) and situational responsiveness (e.g., cultural competence, group facilitation, conflict management, and adaptability). They must ensure that the evaluation is pragmatic, balancing methodological rigor with authentic response to the unique evaluation needs of the primary intended users. One example of this is to share artifacts from similar evaluation projects that the evaluator conducted, such as logic models, evaluation plans, measures, or summarized data reports. The evaluator should also explain how the tools were codeveloped with stakeholders.

Step 3. Identify, organize, and engage primary intended users. In this step, the evaluators must determine stakeholders who represent diverse constituencies and can transmit and use evaluation findings. To continually build these users' capacity and engage them in the evaluation process, evaluators will also need to teach users about evaluation.

Step 4. Situation analysis conducted jointly with primary intended users. Embedded within, yet building on the first three steps, situational analysis is an ongoing process of deepening understanding of the program and the stakeholders' perspectives by taking into account the prior experiences of the program and the stakeholders. In this step, it is determined how the evaluation will inform decisions by considering both barriers and factors that will facilitate use. This is also the step where necessary resources are identified for collaboratively conducting the evaluation and reporting results that will facilitate use.

Step 5. Identify and prioritize primary intended uses by determining priority purposes. The evaluator facilitates the prioritization of evaluation purposes by the primary intended users. This results in clear goals for the evaluation. For example, the evaluator could facilitate a series of group discussion where users brainstorm ways that the evaluation could inform their program and then prioritize options to determine agreed-upon priorities for the evaluation.

Step 6. Consider and build in process uses if and as appropriate. Utilization-focused evaluation increases the ability of primary intended users to think evaluatively. Instead of being an unintended consequence, this step purposefully considers how involvement in the evaluation can have long-term impacts on the users' capacity to apply evaluation logic.

Step 7. Focus priority evaluation questions. In this step, the evaluators support the primary intended users in identifying evaluation questions that are grounded in the goals for the evaluation, meaningful to the users, can be answered with data, and have the potential to result in actionable change. This step builds upon Step 5 by supporting users in developing evaluation questions based on the priority purposes.

Step 8. Check that fundamental areas for evaluation inquiry are being adequately addressed. Evaluators, in collaboration with primary intended users, must consider the complex system influencing answers to the evaluation questions. For example, if users want to know whether a program results in identified outcomes, they should also consider whether the program is being implemented as intended.

Step 9. Determine what intervention model or theory of change is being evaluated. This step involves articulating a testable intervention model (e.g., logic model) or theory of change, identifying assumptions, and comparing the underlying theory to reality. The intervention model or theory of change creates a framework for the evaluation that guides the evaluation design.

Step 10. Negotiate appropriate methods to generate credible findings that support intended use by intended users. Methods are not determined by the evaluators alone. Instead, the evaluators advise the users of options, including benefits and drawbacks, and negotiate quality methods that produce valid and reliable information for decision making. One example is to communicate evaluation methods and measures alongside clear evaluation questions. This helps the intended user better understand what evaluation questions will be answered and what impact can be assessed.

Step 11. Make sure intended users understand potential methods' controversies and their implications. In conjunction with Step 10, this step creates a shared understanding of the trade-offs in any methodological paradigm, with the goal that users understand the controversies around their identified methods for the evaluation.

Step 12. Simulate use of findings. Prior to data collection, fabricated results are discussed among primary intended users as practice in using the data to enact change. This then determines whether all data necessary to inform decisions are collected and, if not, incorporates these data points into the design.

Step 13. Gather data with ongoing attention to use. Evaluators keep primary intended users informed and involved throughout data collection. This includes regular updates on progress and changes, providing feedback, and sharing interim findings. One example is sharing response rates and initial composite results with users to allow for mid-course corrections, wider dissemination strategies, and modifications to data displays.

Step 14. Organize and present the data for interpretation and use by primary intended users. Evaluators actively engage primary intended users in analysis, interpretation, judgement, and recommendations. To do this, evaluators must answer the primary questions through interpretable presentation of findings.

Step 15. Prepare an evaluation report to facilitate use and disseminate significant findings to expand influence. An evaluation report cannot be all things to all people. Instead, it must meet the information needs of primary intended users so that they can employ the results. Instead of a formal narrative, users might be better able to interpret and use meaningful, easy-to-understand charts and graphs with clustered responses. Statements beneath each graph can summarize the data. For open-ended questions, evaluators can cluster responses around themes. Although a few users may want or need a detailed report, most will likely benefit from a short, visual report.

Step 16. Follow-up with primary intended users to facilitate and enhance use. The evaluation does not end when the report is submitted. Instead, the evaluator continues to support the primary intended users in utilizing the results and follows up with them to determine how findings are actually used. For example, if the users are reporting findings to their funders to articulate modifications to their work or to justify additional funding, the evaluators might tailor a condensed report for this group or develop visual representations of the data that can be disseminated. The evaluators might also codevelop guiding questions for stakeholder groups to facilitate reflection on the evaluation and findings and to determine how the evaluation will impact future work.

Step 17. Meta-evaluation of use: Be accountable, learn, and improve. The final step is evaluating the quality of the evaluation. In addition to the quality of the design, procedures, data collection, analysis, and report, utilization-focused evaluation also considers the outcome of the evaluation (i.e., the degree to which the results were used for the intended purposes). Furthermore, meta-evaluation should consider changes in organizational capacity to undertake evaluation. Finally, evaluators should engage in reflective practice to continually improve their own capacity to conduct utilization-focused evaluation.

In utilization-focused evaluation, the evaluator becomes a facilitator and coach, building decision makers’ capacity to ask evaluation questions, determine methods to answer these questions, analyze results, and use results to enact meaningful change.

Amy S. Gaumer Erickson and Patricia M. Noonan

See also Collaborative Evaluation; Developmental Evaluation; Evaluation Capacity Building; Participatory Evaluation

Further Readings

Better Evaluation. (n.d.). Utilization-focused evaluation. Retrieved from http://betterevaluation.org/plan/approach/utilization_focused_evaluation.

Patton, M. Q. (2008). Utilization-focused evaluation (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2012). Essentials of utilization-focused evaluation. Thousand Oaks, CA: Sage.

Ramirez, R., & Brodhead, D. (2013). Utilization focused evaluation: A primer for evaluators. George Town, Malaysia: Southbound Penang. Retrieved from http://www.managingforimpact.org/sites/default/files/resource/ufeenglishprime

Torres, R. T., & Preskill, H. (2001). Evaluation and organizational learning: Past, present, and future. American Journal of Evaluation, 22(3), 387–395.

Westley, F., Zimmerman, B., & Patton, M. Q. (2007). Getting to maybe: How the world is changed. New York, NY: Random House.

V

Andrew Maul Andrew Maul Maul, Andrew

Validity

Validity

1771

1775

Validity

Validity is a central concept in social science research. At the broadest level, validity refers to the extent to which a claim, result, inference, or argument is well founded. In the social sciences, the term *validity* is often (but not exclusively) used in reference to educational and psychological measurement and assessment, where it is frequently referred to as the most fundamental consideration in developing and evaluating tests. However, despite wide agreement regarding its importance, there is no single conception of validity universally accepted in the scholarly and professional communities, and there remains considerable controversy surrounding the definition of validity and many related concepts and terms.

This entry introduces the topic of validity, concentrating on its applications in measurement and assessment. The entry begins with an overview of basic concepts and terminology, followed by sections describing perspectives on validity and validation roughly following a historical progression. Early perspectives on validity are described first, including the concepts of criterion- and content-related forms of validity. The next section discusses construct validity, as first introduced by Lee J. Cronbach and Paul E. Meehl in the mid-1950s, and the idea of nomological networks. Following this are discussions of the unified perspective on validity due primarily to Samuel S. Messick and the interpretive argument-based approach to validation due primarily to Michael T. Kane. The final section discusses a causal perspective on validity due primarily to Denny Borsboom.

Although this entry does not aim to provide a thorough historical overview of thinking about validity, this historical presentation may nonetheless be helpful in contextualizing the origins of many common ways of thinking about validity, especially insofar as each of these perspectives remain influential, to varying degrees, in different areas of contemporary social scientific scholarship. This entry does *not* aim to provide an introduction to how validation activities do or should take shape in any given application.

Basic Concepts in Validity

One of the earliest proposed conceptions of validity (and one that remains popular among many scholars and practitioners) is that validity refers to the extent to which a test measures what it claims to measure. Validity is often introduced alongside the concept of *reliability*, which refers to the extent to which the results of an assessment are free from random sources of measurement error; in other words, the higher the reliability, the less "noise" there will be in the measurement process. Seen this way, validity refers to the accuracy of an assessment, whereas reliability refers to its consistency. A classic pictorial analogy (Figure 1) helps make this concept intuitive: A test that is both valid (accurate) and reliable (consistent) is analogous to a situation in which the bull's-eye of a target is struck regularly. If a test is reliable but not valid, the results will be consistent but consistently off the mark; if it is valid but has lower reliability, the scatter around the bull's-eye will be greater, but on average, the shots will be in the right location. This image also helps motivate the intuition that at least some degree of reliability is a precondition for validity: If *all* of the observed variation in the results of an assessment were due to random measurement error (i.e., zero reliability), it is difficult to imagine how one could claim that the assessment's results have validity.

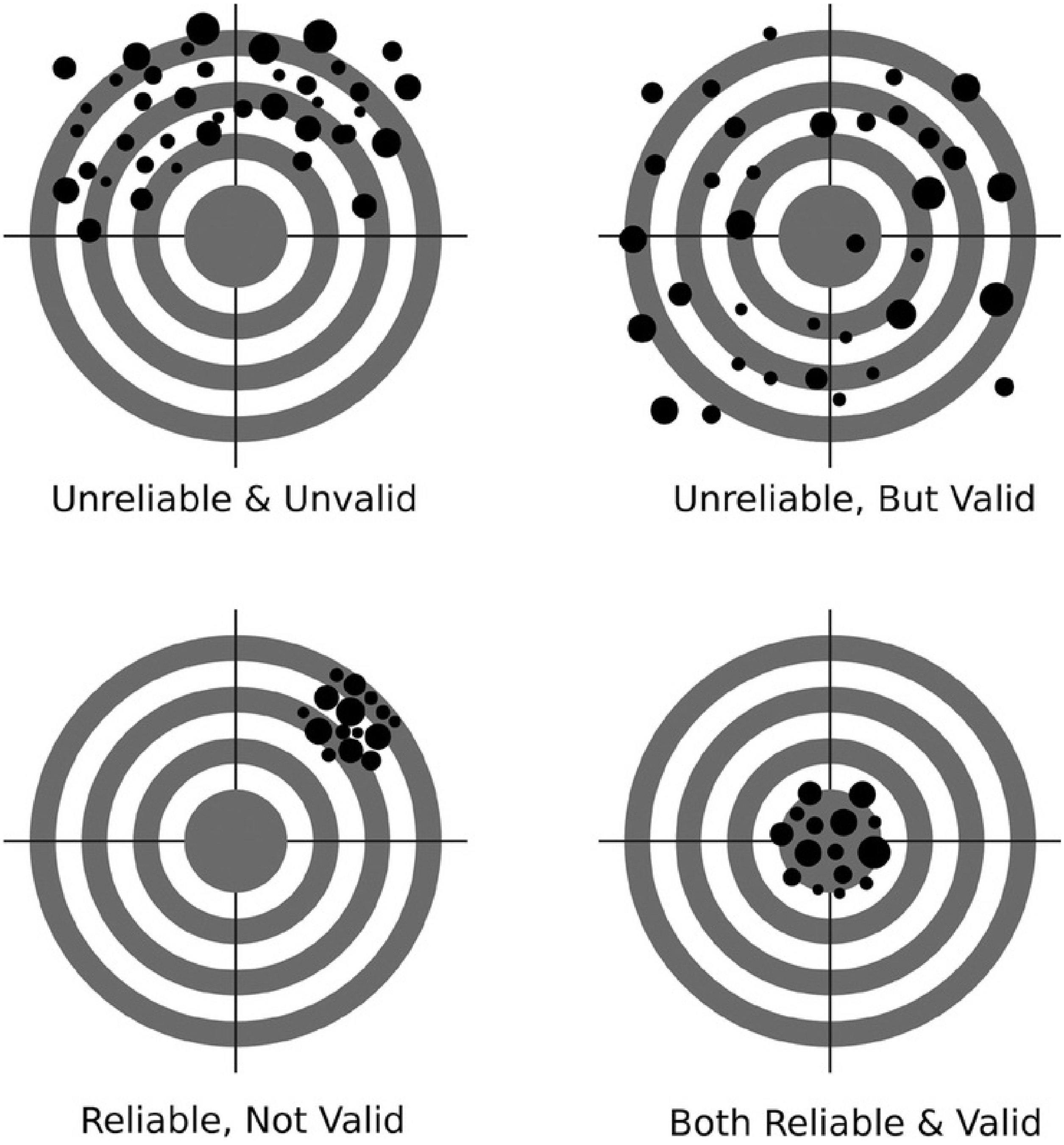


Figure 1 Validity as accuracy, reliability as consistency

Source: © Nevit Dilman, https://commons.wikimedia.org/wiki/File:Reliability_and_validity.svg, licensed under Creative Commons License Attribution-Share Alike 3.0 Unported (CC BY-SA 3.0), https://creativecommons.org/licenses/by-sa/3.0/deed.en

The account given in the previous paragraph focuses on validity in terms of the accuracy of measurement (where *measurement* is broadly understood as the estimation of a person's value of an attribute). It is worth noting that a given test may be interpreted as a measure of more than one attribute or may be reinterpreted after its initial development as a measure of an attribute for which it was not originally intended; thus, a test may be said to be valid for some measurement purposes but not others or to possess greater or lesser degrees of validity for different measurement purposes. Furthermore, in addition to measurement, tests often serve a variety of other purposes as well, and a test may be said to be valid or invalid (or be more valid or less valid) for each of these purposes.

Many tests play a role in decision making for individuals (e.g., job placement tests, college entrance examinations) or groups (e.g., when tests are used as part of educational accountability systems), or play a role in the evaluation of programs and policies and thus help guide policy making. Tests may also serve social purposes such as the signaling of values (to individuals, groups, or society as a whole) and, especially when stakes are attached, may be used to motivate or alter behavior (e.g., as when a student studies calculus because she knows she has an upcoming exam or when teachers are motivated to focus on some topics at the expense of others based on what is covered on year-end assessments). The term *validity* can be applied to describe the adequacy and appropriateness of the test for any or all of these goals. Thus, it is often said that a test cannot simply be claimed to be unconditionally valid or invalid, but rather that understanding a test's validity requires understanding the specific purposes for which the test is intended.

Early Perspectives on Validity

In many of the earliest formal accounts of validity, due to scholars such as Truman L. Kelley and Edward E. Cureton, validity was understood in terms of the correlation between test scores and a criterion measure. For example, the validity of a job placement test might be expressed in terms of its correlation with measures of job performance, or a short version of a test might be evaluated in terms of its correlation with a longer or more thorough battery of tests. Such test-criterion correlations were sometimes referred to as *validity coefficients*.

In other contexts, tests were developed from sets of content specifications (e.g., as on many educational tests, where a primary goal would be to ensure adequate coverage of the domain covered in a course). In such contexts, prediction of a specific external criterion could be regarded as less important than ensuring that the content of the test representatively sampled from the domain of interest; this, in turn, was primarily established via documentation of the test-construction procedures and expert review. This led to a distinction between criterion-related validity and content validity, initially thought of as each applying to different types of tests.

Construct Validity

Although criterion-and content-related forms of evidence seem to be appropriate for many tests, some tests appear to defy this classification. In particular, psychological attributes such as personality characteristics (e.g., aggression, contentiousness) and broadly defined cognitive abilities (e.g., general intelligence) seem difficult to operationalize in terms either of a specific domain of content coverage or in terms of relations with specific external criteria. In 1955, Cronbach and Meehl introduced the concept of *construct validity* to account for such cases, where construct validity was understood primarily in terms of how scores on a given test related to a network of other observables (called a *nomological network*) and the extent to which these relations were consistent with predictions made based on the theory of what the test measured. For example, suppose that a theory of creativity states that creativity should be positively associated with general intelligence, but not with agreeableness. Finding that scores on a creativity test correlate positively with scores on a general intelligence test but not with an agreeableness test would provide corroborative support for the theory that the creativity test is a valid measure of creativity. A wide variety of other observations might also bear on the empirical evaluation of this theory.

It should be noted that this sort of correlational evidence can only be interpreted as evidence of validity to the extent to which it can be compared to a priori, theory-based predictions. To continue with the example in the previous paragraph, suppose that the theory of creativity does not specify whether creativity is expected to be associated with extroversion. Finding that scores on the creativity test correlate with scores on an extroversion test (whether positively or negatively) could be interesting and theory generating, and it may be possible to generate a convincing ad hoc explanation for such an observed association, but the finding could not be interpreted as evidence either for or association, but the finding could not be interpreted as evidence either for or against the validity of the instrument. Furthermore, the finding cannot even be interpreted as evidence of an association between creativity and extroversion, unless it is presupposed that the creativity test is in fact a valid measure of creativity—in other words, presupposing the very claim that the nomological evidence is meant to help test.

One consequence of the focus on construct validity was increased attention to the distinction between a test and the psychological attribute, or construct, putatively measured by the test (as opposed to the earlier, operationalist view that a test simply defined a construct). This opened the door to the possibility that multiple tests could measure the same construct. Donald T. Campbell and Donald W. Fiske popularized the idea of multimethod studies, and in so doing added two new terms into the validity lexicon: *convergent validity*, reflecting the idea that multiple measures of a common construct should exhibit high levels of agreement with one another (i.e., they should “converge” on a common truth), and *discriminant validity*, reflecting the idea that measures of distinct constructs should not be too highly correlated with one another, even if they used the same method of observation (i.e., it should be possible to empirically “discriminate” among theoretically distinct constructs). As a classic example, suppose that extroversion and dominance are both assessed via self-report and the reports of one’s family members. Evidence for convergent validity could take the form of showing that self-reports and family reports of extroversion are highly correlated (and similarly for dominance); evidence for discriminant validity could take the form of showing that self-reports of extroversion and self-reports of dominance are not so highly associated as to render them empirically redundant (and similarly for family reports).

A Unified Theory of Validity

Starting in 1989, Messick offered a new perspective on validity that reflected a significant shift from previous viewpoints in several respects. Messick’s view subsumed disparate lines of validity-related evidence under the generalized concept of construct validity. On this view, validity is a single property of a test—the extent to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores. Thus, the idea of distinct types of validity (e.g., criterion, content, construct) was replaced with the notion of there being distinct types of *evidence* that could be brought to bear on the validity of a given test, depending on the intended purposes of the test. Broadly, these types of evidence help establish that the test assesses as much as possible of what it should assess and as little as possible of what it should not: In Messick's language, this involves minimizing both *construct underrepresentation* and *construct-irrelevant variance*.

This view also had the function of calling greater attention to the intended purposes of tests—including both interpretations of test scores and actions taken on the basis of such scores—and to the idea that quite different types of evidence could be necessary depending on these purposes. One of the more controversial elements of this theory was the proposition that validation explicitly involves a consideration of the consequences of test interpretation and use. For example, if educational tests given to students are used to help inform decisions made about the retention and compensation of teachers, claiming that the tests are valid for this purpose would involve demonstrating not only that they measure the knowledge, skills, and abilities of students they claim to measure but also that using these tests as a basis for high-stakes decisions about teachers has the intended positive consequences and does not have unforeseen negative consequences. This viewpoint could be taken as broadening the concept of validity to include social and moral concerns in addition to more purely epistemic concerns. Although Messick himself only proposed that the consequences of tests could be used as indirect evidence of construct underrepresentation and construct-irrelevant variance, other scholars such as Lorrie A. Shepard made stronger proposals for the explicit consideration of consequences as a primary and independent source of validity evidence.

An Argument-Based Approach to Validation

Messick's unitary view of validity has remained influential since its introduction and is arguably still the dominant conception of validity in the literature on educational assessment and measurement. Using Messick's definition of validity as a starting point, scholars such as Kane have argued that validation should consist of the construction and evaluation of an argument aimed at defending the appropriateness of a test for a particular use and the collection of forms of evidence relevant to that argument. Kane's argument-based approach is an approach to the problem of practical validation and not a new theory about validity itself; this emphasis on validation rather than validity reflects a shift in focus toward pragmatic, context-specific arguments tailored for specific audiences and circumstances.

The argument-based approach emphasizes that any validation effort begins with a clear statement of the proposed uses and interpretations of a test, and that if tests are used for purposes other than those originally intended, this will require a reexamination of the validity argument or the development of an entirely new argument. On this view, consequences of testing would play a central role in a validity argument for a given test insofar as the proposed use of the test implies an intention for certain consequences to happen (or not to happen) as a result.

A Causal Perspective on Validity

The perspectives described in the previous sections could be broadly characterized as representing the mainstream ways of thinking about validity found in the literature on educational and psychological measurement and assessment, as reflected in prominent sources such as the many editions of *Educational Measurement* and the *Standards for Educational and Psychological Testing*. However, these perspectives do not encompass the whole of thinking about validity. To start, it could be noted that the scholarship described herein comes almost exclusively from the United States. Furthermore, it is largely focused on the roles that tests play in education (including both large-scale, standardized achievement tests and smaller scale tests such as those designed primarily for pedagogical purposes) and in other domains in which tests are regularly used to make decisions about individuals and groups, such as in industrial/organizational settings and in counseling and clinical psychology. This focus may help explain the steady shift away from timeless and universal standards of scientific evidence toward a more dynamic, contextualized, and pragmatic focus on the evaluation of the appropriateness of tests for particular uses.

In contrast, other recent scholarship has more strongly emphasized understanding the semantics of validity in terms of factual claims about true states of affairs. In particular, Borsboom and colleagues have developed an account of validity that could be regarded as an extension of the earliest definition of the term (i.e., validity is whether a test measures what it claims to measure): Specifically, a test is a valid measure of an attribute if (a) the attribute exists and (b) variation in the attribute causes variation in the outcomes of the test. This causal perspective on validity emphasizes that whether or not a test is valid as a measure of an attribute is a claim about the state of affairs in the world, and is thus independent of the *evidence* available at any given time, or the extent to which that evidence is found to be persuasive by any given community of observers. On this perspective, validation involves tracing the causal pathway leading from (between and/or within person) variation in the attribute to variation in the outcomes of the testing procedure; thus, this perspective emphasizes the importance of strong cognitive theory explaining the relationship between test scores and the attribute being measured.

Joel Michell offers another measurement-focused perspective on validity, even more strongly dissenting from the "mainstream" perspectives described previously. Michell argues that the concept of measurement is classically understood as *the estimation of ratios of magnitude* and thus requires that an attribute both exist and be quantitatively structured in order to be measurable; furthermore, he argues that this is not something regularly or adequately tested in applied educational and psychological research.

It could be noted that both Borsboom and Michell focus on the validity of tests as *measures* of an attribute. As previously noted, interpreting test scores as measures in this sense is only one among many possible interpretations, and tests are routinely put to uses that, strictly speaking, do not appear to require that any measurement take place at all; thus, the focus on measurement could be described as a special case of a focus on test score interpretations and uses. But especially insofar as many test interpretations and uses do at least appear to depend on measurement claims, both Borsboom's and Michell's perspectives call attention to the need to articulate and justify claims about measurement separately and in addition to other claims about test interpretation and use more broadly.

Andrew Maul

See also Concurrent Validity; Consequential Validity Evidence; Construct Irrelevance; Construct-Related Validity Evidence; Construct Underrepresentation; Content-Related Validity Evidence; Criterion-Based Validity Evidence; Predictive Validity; Unitary View of Validity; Validity, History of

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014).

Validity. In Standards for educational and psychological testing (pp. 11–31). Washington, DC: American Psychological Association.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061–1071.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Markus, K. A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. Routledge.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741.

Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions and applications (pp. 135–170). Information Age.

Newton, P., & Shaw, S. (2014). Validity in educational and psychological assessment. Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16(2), 5–8. doi:10.1111/j.1745-3992.1997.tb00585.x

John D. Hathcoat John D. Hathcoat Hathcoat, John D.

Nicholas A. Curtis Nicholas A. Curtis Curtis, Nicholas A.

Courtney B. Sanders Courtney B. Sanders Sanders, Courtney B.

Shengtao Liu Shengtao Liu Liu, Shengtao

Validity, History of Validity, history of

1776

1780

Validity, History of

Validity is arguably one of the most important, yet widely misunderstood, concepts in educational measurement, research, and evaluation. Misconceptions about validity may partly be due to the fact that the term *validity* has distinct meanings both between and within academic disciplines. Even within educational research, validity is described differently within the context of experimental design than it is within measurement. To make matters more complicated, the concept of validity has continued to evolve within the field of educational measurement, and it is this latter domain that is the focus of the present entry.

The concept of validity has changed from a simple typology believed to be a property of tests to a unified concept reflecting the extent to which evidence and theory support score-based interpretations and proposed uses of a test. The *Standards for Educational and Psychological Testing*—hereafter referred to as the Standards—provides a vantage point to frame this discussion about the history of validity. As of 2017, the Standards, including the technical recommendations serving as a forefather of this publication, have been revised 6 times since 1952, with the most recent version published in 2014. For the sake of simplicity, the concept of validity is described across three historical periods. This entry first discusses the concept of validity prior to the 1950s, which emphasized correlational evidence as well as test content. This is followed by the introduction of construct validity in the early 1950s as one of three types (i.e., criterion, content, and construct). Next, this entry discusses the unification of validity beginning in the 1970s throughout the 1990s and concludes with a brief outline of contemporary perspectives toward this concept.

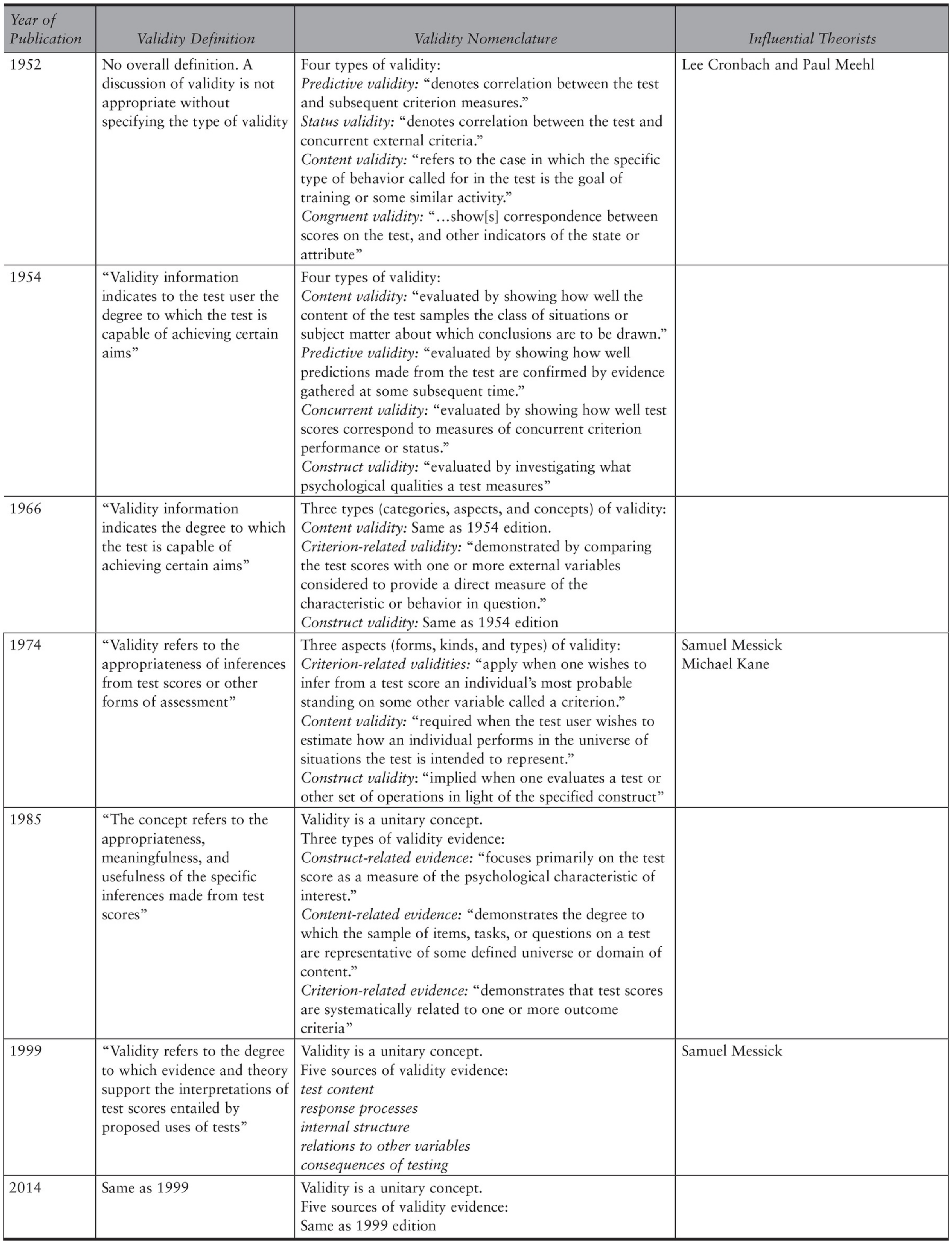
The Concept of Validity Prior to the 1950s

Validity theory has primarily been a concern among theorists in education and other social sciences. The advent of validity theory within these disciplines coincides with early efforts to legitimize the practice of educational and psychological measurement as a scientific activity. As educational and psychological testing became more prominent in the early 1900s, theorists sought to provide an account of validity to justify the consequential actions derived from test scores. Two developments influenced conceptions of validity before the publication of the first technical recommendations in 1952: The formalization of the correlation coefficient and concerns raised by educational researchers about the connection between test content and validity. Both developments are consistent with the view that tests themselves are valid or invalid.

Correlations, which reflect the direction and magnitude of a relationship between two variables, are crucial for making empirical predictions. The formalization of the correlation coefficient led to the concept of criterion-related validity. Criterion-related validity was a concern when researchers were interested in using a test as an indication of a criterion of ultimate interest. The criterion of interest may reflect the future or current status on a variable. Tests were therefore viewed as valid for anything with which it correlated. However, educational researchers indicated that the criterion-related view failed to address the role of test content. These researchers argued that the content of a test, such as an achievement test, should be evaluated for its relevance to, and representation of, a specified domain. In sum, prior to the 1950s, validity was largely viewed as two distinct, though in some sense connected, types which included criterion-related validity and content validity.

The Birth of Construct Validity in the Early 1950s

The concept of construct validity was discussed in the 1952 and 1954 technical recommendations developed by American Psychological Association committees (Table 1). Lee J. Cronbach and Paul E. Meehl, who served on each committee, later published an article in 1955 aiming to further describe construct validity as a third type (i.e., content, criterion, and construct). Although the nomenclature for this typology, often referred to as the Three Cs, slightly changes across publications, one can recognize the Three Cs in both the 1966 and 1974 Standards. Under this view, the three types of validity correspond with three inferences (i.e., generalization of content, present or future status of a criterion, and a theoretical construct). A construct is defined as a theoretical term employed by researchers to explain or summarize consistencies in empirical observations. For example, the terms *intelligence*, *self-esteem*, and *leadership* are constructs that cannot, strictly speaking, be directly observed. At best, constructs are indirectly investigated. In the 1955 article, by Cronbach and Meehl, construct validity concerns obtaining evidence that a hypothesized construct accounts for variation in observed scores.



Year of Publication	*Validity Definition*	*Validity Nomenclature*	*Influential Theorists*
1952	No overall definition. A discussion of validity is not appropriate without specifying the type of validity	Four types of validity: *Predictive validity:* "denotes correlation between the test and subsequent criterion measures." *Status validity:* "denotes correlation between the test and concurrent external criteria." *Content validity:* "refers to the case in which the specific type of behavior called for in the test is the goal of training or some similar activity." *Congruent validity:* "…show[s] correspondence between scores on the test, and other indicators of the state or attribute"	Lee Cronbach and Paul Meehl
1954	"Validity information indicates to the test user the degree to which the test is capable of achieving certain aims"	Four types of validity: *Content validity:* "evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn." *Predictive validity:* "evaluated by showing how well predictions made from the test are confirmed by evidence gathered at some subsequent time." *Concurrent validity:* "evaluated by showing how well test scores correspond to measures of concurrent criterion performance or status." *Construct validity:* "evaluated by investigating what psychological qualities a test measures"	
1966	"Validity information indicates the degree to which the test is capable of achieving certain aims"	Three types (categories, aspects, and concepts) of validity: *Content validity:* Same as 1954 edition. *Criterion-related validity:* "demonstrated by comparing the test scores with one or more external variables considered to provide a direct measure of the characteristic or behavior in question." *Construct validity:* Same as 1954 edition	
1974	"Validity refers to the appropriateness of inferences from test scores or other forms of assessment"	Three aspects (forms, kinds, and types) of validity: *Criterion-related validities:* "apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion." *Content validity:* "required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent." *Construct validity*: "implied when one evaluates a test or other set of operations in light of the specified construct"	Samuel Messick Michael Kane
1985	"The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores"	Validity is a unitary concept. Three types of validity evidence: *Construct-related evidence:* "focuses primarily on the test score as a measure of the psychological characteristic of interest." *Content-related evidence:* "demonstrates the degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content." *Criterion-related evidence*: "demonstrates that test scores are systematically related to one or more outcome criteria"	
1999	"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests"	Validity is a unitary concept. Five sources of validity evidence: *test content* *response processes* *internal structure* *relations to other variables* *consequences of testing*	Samuel Messick
2014	Same as 1999	Validity is a unitary concept. Five sources of validity evidence: Same as 1999 edition	

The introduction of construct validity in the 1950s had a few implications on subsequent validity frameworks, many of which are apparent in contemporary validity theory. Contrary to earlier views, the 1952 and 1954 recommendations validity theory. Contrary to earlier views, the 1952 and 1954 recommendations abandoned the idea that tests were either valid or invalid. In fact, both recommendations warn that it is inappropriate to use the unqualified statement "this test is valid." Instead, validity was conceived as a property of score-based inferences. Investigating construct validity required strong theory and an examination of multiple lines of evidence aiming to demarcate the meaning of theoretical constructs used to describe a set of scores. Construct validation, as described by Cronbach and Meehl, consisted of an ongoing investigation. Finally, validity theorists recognized interconnections between the Three Cs. For example, theoretical constructs may inform test content that can, in turn, be used to assist in the identification of criterion variables.

Efforts to Unify Validity: 1970s–1990s

The Three Cs was eventually abandoned in favor of a unified view of validity (Table 1). An effort to unify validity under a single framework was initiated during the 1970s, even though the 1974 Standards still employed a typological language. The "seeds" of unification can be found in various sources; however, a unified view of validity was perhaps most vehemently argued by Samuel S. Messick, who conceived of construct validity as an evaluative judgment about the extent to which evidence and theory support the adequacy and appropriateness of score-based inferences and actions resultant from test use. Under Messick's view, all validity is construct validity, which includes the consequential aspects of testing.

Messick was ultimately concerned about the meaning attributed to a set of scores. Any evidence bearing on the meaning of scores informs construct validity. Test content, as well as the relationship of scores to other variables, informs score meaning. Thus content and criterion-related validity can be subsumed as an aspect of construct validity. The inclusion of consequences as an aspect of construct validity largely derived from Messick's position that the meaning attributed to scores cannot be separated from social values. In other words, social consequences inform, and are informed by, the meaning attributed to a set of scores; thus, the consequential aspects of testing can also be subsumed as a part of construct validity.

The 1985 Standards is the first version of this publication to explicitly provide a unified view of validity, which has remained consistent in each subsequent publication. Prior versions of the Standards had distinct sections for each of the three types (e.g., content validity is addressed in a different section than construct validity). The language used in the 1985 Standards reflects a radical departure from this simple typology by discussing construct-related, criterion-related, and content-related “evidence” as opposed to validity. Although the ideas of Messick are apparent in the 1985 Standards, this text does not claim that all validity is construct validity. Validity is instead unified when defined as the appropriateness, meaningfulness, and usefulness of score-based inferences. The 1985 Standards also depart from Messick’s framework by excluding the consequential aspects of testing as a central validity issue.

Contemporary Views Toward Validity

Most contemporary theorists view validity as a unitary concept denoting a property of score-based inferences and uses of a test. Scores can be interpreted in various ways and a test could be adopted to accomplish multiple purposes. Both the 1999 and 2014 Standards thus emphasize the process of investigating score-based inferences as opposed to describing the concept of validity. Validation, defined as the process of accumulating relevant evidence to provide a scientific basis for score-based inferences, has become increasingly important. This may be attributed to the work of Michael T. Kane who has advanced an argument-based approach to validation. Under his view, validation involves constructing an argument consisting of an overall evaluation of each claim, plausible alternative explanations, and the tenability of underlying assumptions.

Both the 1999 and 2014 Standards align with an argument-based approach in their emphasis of five sources of validity evidence. These sources of evidence include test content, response processes, internal structure, relations to other variables, and test consequences. The inclusion of test consequences has remained a controversial topic among validity theorists. However, according to the current standards, intended consequences should be investigated as a validity issue. Negative consequences are a validity issue when they result from construct underrepresentation (i.e., test is missing something important) or construct-irrelevant variance (i.e., scores are influenced by something unintended). Numerous controversies continue to exist among contemporary validity theorists; hence, it is likely that the concept of validity will continue to evolve as the field of educational and psychological measurement confronts existing and new challenges.

John D. Hathcoat, Nicholas A. Curtis, Courtney B. Sanders, and Shengtao Liu

See also Reliability; Validity; Validity Generalization

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51, (Suppl. 2), 1-38. doi:http://dx.doi.org/10.1037/h0053479

American Psychological Association. (1966). Standards for educational and psychological tests and manuals. Washington DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). Standards for educational and psychological tests. Washington DC: American Psychological Association.

American Psychological Association, & Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. American Psychologist, 7, 461–475. doi:http://dx.doi.org/10.1037/h0056631

Cronbach L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302. doi:http://dx.doi.org/10.1037/h0040957

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50, 1–73.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement (3rd ed.). Washington, DC: American Council on Education.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions and applications (pp. 19–37). Charlotte, NC: Information Age Publishing.

Sheila K. List Sheila K. List List, Sheila K.

Michael A. McDaniel Michael A. McDaniel McDaniel, Michael A.

Validity Coefficients Validity coefficients

1780

1782

Validity Coefficients

Validity, the accuracy of a conclusion or inference, is central to research, measurement, and evaluation. If inferences made from the results of a test or measure are not considered valid, it suggests that the test may not be measuring what is supposed to be measured. Several sources of evidence can be used to establish validity. One source of validity evidence is criterion-related evidence. Criterion-related validity represents the relationship between scores on a test or measure and a criterion, or outcome, variable. Validity coefficients are correlations that quantify this relationship; thus, they are essential for establishing criterion-related validity.

Interpreting Validity Coefficients

Validity coefficients can be described with respect to direction and magnitude and can vary between −1 and +1. A validity coefficient of 0 indicates that there is no linear relationship between scores on the measure and scores on the criterion variable. Such a correlation has no magnitude and no direction. A coefficient of .90 has a large magnitude (a value of 1 would be largest possible magnitude) and has a positive direction such that as one variable increases (e.g., a reading readiness test score), the other variable increases (e.g., reading performance in the classroom). A coefficient of −.15 has a small magnitude and a negative direction such that as one variable increases (e.g., school days absent), the other variable decreases (e.g., family socioeconomic status). A validity coefficient describes a relationship, but it does not necessarily imply causation.

The closer the validity coefficient is to 1, the more confident one can be in the accuracy of the inferences drawn from scores on the measure. For example, if a test designed to predict success in college has a validity coefficient of 1, it means that the test has a perfect linear relationship with the criterion of interest (e.g., freshman year grade point average). In other words, scores on the test perfectly predict freshman grade point average, thus the higher an individual’s score on the test, the higher the individual’s future grade point average. In this situation, test users (i.e., college admissions staff) can feel very confident that they can make accurate decisions based on test performance. If the validity coefficient of this test was 0, it would indicate that test performance does not accurately predict college success and should, therefore, not be used.

Much of social science research relies on Jacob Cohen’s guidelines for interpreting the magnitude of a correlation. He suggested that correlations of .1 represent a small effect, correlations of .3 represent a moderate effect, and correlations of .5 or greater represent a large effect. Using these guidelines, a validity coefficient of .35, though far from a perfect correlation of 1, may be considered useful. In fact, any nonzero correlation provides some predictive value. Although these guidelines can be useful, they do not provide insight into the practical impact of using the measure. Therefore, other methods, such as Taylor–Russell tables, utility analyses, and tests of sensitivity and specificity, are also commonly used.

In the field of personnel selection, Taylor–Russell tables use the validity coefficient, selection ratio (number selected or hired), and the proportion of candidates that would perform well on the criterion variable to determine the probability that a candidate who performs well on a selection test will perform well on the job. The results of the analysis provide test users with an estimation of how much using the test improves their hiring decisions as compared to not using the test. Utility analysis builds on the Taylor–Russell tables by considering the monetary impact of using the test to make hiring decisions. Utility analyses can be helpful for organizations deciding whether or not to use a test for selection purposes because they provide the organization with a cost versus benefit estimation.

In other fields, sensitivity and specificity tests are used more frequently. These are particularly useful when the outcome variable, or criterion, is dichotomous. Specifically, these tests are used to determine the probability that an individual’s score on a measure will lead to a correct categorization. For instance, if a measure is designed to determine which students are “gifted” and should attend a classroom for gifted students, sensitivity and specificity tests would provide classroom for gifted students, sensitivity and specificity tests would provide insight into the accuracy of the categorization of students based on test scores.

Factors Affecting Validity Coefficients

Several factors affect the size of the validity coefficient, including the type of validation study, range restriction, as well as unreliability and measurement error. Concurrent validation studies use test scores and criterion scores that are gathered at the same time and tend to result in larger validity coefficients. Predictive validation studies use test scores collected at one point in time, and criterion scores collected at a later point in time and tend to result in smaller validity coefficients. Range restriction also affects validity coefficients. For example, if an educational test is particularly easy and most students answer every item correctly, the lack of variance in test scores will result in a lower correlation between that test and future academic performance. Lastly, measurement error will affect the magnitude of the validity coefficient, specifically the more measurement error, the larger the underestimation of population validity coefficient. Given that these factors have such a significant influence on the validity coefficient, it is essential to consider the impact that each of these may have had when interpreting a validity coefficient.

Sheila K. List and Michael A. McDaniel

See also Correlation; Criterion-Based Validity Evidence; Psychometrics; Reliability; Validity

Further Readings

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Furr, R. M., & Bacharach, V. R. (2008). Psychometrics: An introduction. Thousand Oaks, CA: Sage.

Markus, K. A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. New York, NY: Routledge.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 23, 565–578. Retrieved from http://dx.doi.org/10.1037/h0057079

Sheila K. List Sheila K. List List, Sheila K.

Michael A. McDaniel Michael A. McDaniel McDaniel, Michael A.

Validity Generalization Validity generalization

Validity Generalization

To ensure that researchers and practitioners understand the degree of accuracy of their decisions, collecting validation evidence for the measures they use is crucial. Validity coefficients, correlations that represent the linear relationship between scores on a predictor and scores on a criterion, are quantitative indices of the value of an assessment in predicting future outcomes. Validity generalization is an application of meta-analysis that is used to estimate the mean and variance of a collection of validity coefficients. The validity generalization estimate of the mean validity coefficient is likely to be more accurate than local validation studies that often have limited sample sizes, outcome variables with measurement error, and possibly data with a restricted range of scores. This entry details the value of summarizing individual validation studies, presents a conceptual description of validity generalization techniques, and discusses special considerations for its application and use.

Validation Studies and Situational Specificity

The higher the validity coefficient, the more accurate the decisions made as a result of scores on the measure. Therefore, it is in the best interest of practitioners and researchers to use measures that have sufficient validity evidence. In addition, legal and professional guidelines may suggest the necessity for conducting validation studies. It was previously believed that local validation studies were required to determine the validity of using a measure. This meant that validation studies had to be conducted at each separate location or situation in which a measure was used. This was necessary because the view at the time was that there was something unique about each location or situation in which a test was used that caused a measure to be valid in one location but not in another location. This is known as the situational specificity hypothesis. However, conducting validation studies is time-consuming and costly. Furthermore, depending on the location, the test user may only have a small sample size from which to collect data. This small sample size affects the variance of validity coefficients across situations due to random sampling error. In fact, researchers have documented that variations in validity coefficients across locations are primarily due to studies having small sample sizes and have nothing or little to do with the characteristics of, or differences across, locations.

Meta-Analysis as a Tool

To address these sampling, measurement error, and range restriction issues, researchers can employ meta-analysis. Meta-analysis is a set of research synthesis methods, which combines the results of several primary studies leading to a cumulated sample size that substantially mitigates the effects of random sampling error. Furthermore, psychometric meta-analysis allows for the correction of errors due to range restriction in predictors and measurement error. By increasing the sample size and correcting for error, meta-analysis provides an estimate of the mean population validity and the variance in population validity. If the results of the meta-analysis indicate that 90% or more of the validity coefficient estimates are above 0, the measure is said to demonstrate validity generalization. This eliminates the need for costly, time-consuming, and underpowered local validation studies.

Special Considerations

This section offers several considerations in validity generalization. First, if one is correcting for measurement error when determining validity, one will likely only correct the validity coefficient for measurement error in the criterion. This is because the test, with its inherent measurement error, is being used in the screening. Second, in order to use validity generalization evidence, one should consider whether the test and criterion included in the meta-analysis are comparable to fit one’s needs. For instance, consider this scenario: A school principal wants to use a measure of conscientiousness to predict the likelihood that a teacher will turnover. In an effort to determine whether the measure is a valid predictor of turnover, the principal locates a meta-analysis concerning conscientiousness. The meta-analysis indicates that conscientiousness has a validity coefficient of .35; however, the criterion variable in the meta-analysis was job performance, not turnover. Therefore, this meta-analysis does not provide evidence for the principal's objective. If the meta-analysis had used turnover as the criterion but had used a measure of personality (e.g., locus of control) that measured something meaningfully different than conscientiousness, then the evidence is also not useful. Lastly, one needs to determine whether the population in the meta-analysis suits one's needs. For instance, if a meta-analysis includes only studies with individuals in primary school, but the test user wishes to use the measure with teens, it is not clear whether the meta-analytic evidence is applicable to the test user's needs.

In addition to these considerations, test users must also carefully consider the transparency of the meta-analytic study. For instance, given that psychometric meta-analysis can be used to correct for range restriction and unreliability, it is important that the meta-analyst describes from where the range restriction and reliability estimates were obtained and how these corrections were applied with respect to the mean and variance of the set of validity estimates. In addition, the meta-analyst should provide a table listing all of the data used in the meta-analysis, the sources of these data, and a description of how these data were identified and selected for inclusion. Lastly, meta-analyses should include assessments of publication bias. Publication bias exists when the studies included in the meta-analysis is not representative of all studies that have been conducted on the topic of interest. This bias can occur for a number of reasons and can influence the magnitude of validity coefficients. Because publication bias affects the meta-analytic estimate of the validity coefficient, determining whether the set of validity coefficient data is affected by publication bias is an important, but often neglected, consideration.

Sheila K. List and Michael A. McDaniel

See also Criterion-Based Validity Evidence; Meta-Analysis; Psychometrics; Reliability; Validity; Validity Coefficients

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures*. Federal Register, 43(166), 38290–39315.

McDaniel, M. A. (2007). Validity generalization as a test validation approach. In S. M. McPhail (Ed.), Alternative validation strategies: Developing new and leveraging existing validity evidence (pp. 159–180). San Francisco, CA: Wiley.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529–540. doi:10.1037/0021–9010.62.5.529

Cassandra Guarino Cassandra Guarino Guarino, Cassandra

Value-Added Models

Value-added models

1783

1787

Value-Added Models

Value-added models are statistical models that try to identify the impact of programs, people, or environments on a specific outcome. In education research, the term typically refers to models that look at the impact of various inputs on student growth in achievement, where achievement is generally measured by standardized test scores. In this sense, these models fit into the broader investigation of the *education production function,* which describes the process of producing learning as a formula that relates various inputs to the output of learning. So, within this context, for example, educators and researchers can use a value-added model to ask whether a particular program or intervention helps raise achievement—to see “what works” in improving student learning. Such an investigation would fall under the heading of program evaluation. Alternatively, educators and researchers could also use value-added models to estimate how effective individual teachers or schools tend to be at raising their students’ achievement. These types of analyses are generally used for accountability purposes—to see who is performing well.

In all cases—whether value-added models are oriented toward program evaluation or accountability systems—the models estimate how the features, institutions, or individuals of interest improve student achievement during a specific time period. To help restrict their focus to a particular time period, the models always adjust the estimates of impact in some way for a student’s prior achievement to determine the “value added” of the input of interest after that point. It is this feature of the model—the adjustment for prior achievement—that gives it its name.

This entry begins by explaining how to compute a value-added model. It then

This entry begins by explaining how to compute a value-added model. It then discusses concerns with these models and how those concerns are commonly addressed. Next, the entry presents debate and controversy and concludes by reviewing research findings and summarizing the current and future outlook.

Gain Score Model

The simplest way to compute a value-added model involves averaging a simple growth measure by group over the time period of interest. So, for example, if one is interested in estimating a teacher’s value added for her students in a particular school year, one could very simply compute the average difference in each of her student’s test scores between the end of the year and the end of the prior year. As long as test scores are scaled along a continuum over time, this would produce a simple rough measure of the average growth in student learning for the teacher’s students.

Clearly, this value-added approach is better than simply judging teachers on the basis of the average test scores of their students at the end of the year. The simple average of end-of-year test scores does not take prior achievement into account and will thus provide an unfair basis of comparison for teachers. It will be unfair because some teachers might have been assigned to a classroom full of students who were high performing from the beginning while others might have been assigned to a classroom of low-performing students. If teachers were evaluated on the basis of test scores of their students at only one point in time—the end of the year—teachers with high-performing students could appear to be more effective than others, even if they actually provided very little helpful instruction. The simple value-added model described here—that is, the average “gain score” model—is a better measure by which to compare the impact of teachers than an average that does not take prior test scores into account.

Concerns and Approaches for Addressing Concerns

The simple gain score model, however, immediately highlights a reason why many people worry that even value-added estimates can be unfair—or “biased” in statistical terminology. Some teachers may have students who learn a lot outside of school—in families that supplement their education through parental involvement, tutoring, or enrichment programs. Other teachers may have students who lack support for their studies at home or even proper nourishment. Thus, teachers might not be fully responsible for the achievement growth or lack thereof of their students over the course of a year. It should be noted that if students were randomly assigned to teachers, these types of concerns about bias would be unfounded. However, students are often tracked into classrooms of similar ability and, on top of that, purposefully sorted to particular teachers. Therefore, the concern is a valid one.

The same type of worry exists when evaluating programs rather than teachers. For example, a certain type of learning intervention—say, a particular mathematics curriculum—may be given to certain kinds of students and not others, so it will be difficult to determine the true effect of the curriculum if the model is not able to distinguish that effect from all other factors that might simultaneously affect the amount students learn.

Generally speaking, value-added models can be structured to reduce concerns about bias. If the statistical model is structured so that it can “adjust” for other types of factors that contribute to learning, then the value-added estimates of the effectiveness of teachers, schools, or programs are more believable. However, many states and districts use value-added models that are not well structured to reduce concerns about bias—even for accountability purposes that have stakes attached in terms of incentives or sanctions for schools and teachers. Many studies of value-added methods have shown that different modeling approaches can produce different estimates of effectiveness, so it is important to use stronger models where possible, particularly if there are policy stakes attached to the value-added scores.

The primary way that value-added models adjust for other factors contributing to learning is by controlling for demographic characteristics of students that might in some way proxy for the amount of out-of-school support or learning issues students might have—characteristics such as family income, non-English-speaking home, or special needs status. However, it is not enough just to control for these characteristics. It is important that the value-added model include indicator or “dosage” variables representing the teacher of record (or the program) in the model. This is so that statistical regression can “partial out” the effect of the student characteristics from the teacher performance estimate. In nontechnical terms, this means that a teacher’s effect estimate will take into account the characteristics of the teacher’s students.

There are two other challenges that value-added models must surmount, however, in order to produce good measures of the performance of individual teachers, schools, or programs. One is the issue of sample size. This comes into teachers, schools, or programs. One is the issue of sample size. This comes into play primarily in teacher evaluation, where teachers may have small classes—say, just 15–25 students. It is known from statistics that small sample sizes do not produce very precise estimates. This concern can be alleviated somewhat by computing teacher value-added over more than 1 year at a time—thus building up the number of students contributing to a particular teacher's estimate to a higher number. Of course, this approach assumes that teachers do not fluctuate much in their effectiveness from year to year.

The other concern is about measurement error in the outcome variable. Measurement error in test scores poses a problem because students may post inaccurate learning gains if the tests being used do not adequately measure their progress. Therefore, their teachers will be judged on the basis of student achievement measures that are subject to error. Test scores are known to contain a certain amount of imprecision, particularly at the tails of the achievement distribution—that is, the students with very high and very low achievement scores. Measurement error can thus become particularly troublesome if students with different characteristics have different amounts of error in their test scores. Thus, it is incumbent upon any accountability system that relies on test scores to evaluate teachers, schools, or programs to use standardized assessments that are as precise as possible.

Debate and Controversy

Much debate has taken place regarding the usefulness and fairness of value-added models. Their use in teacher evaluation with stakes attached has been particularly controversial. Teachers' unions, in general, do not support the use of value-added measures in performance evaluation. Some public reporting of individual teacher value-added scores by the media, such as the *Los Angeles Times* and the *Wall Street Journal*, have exacerbated the issue, generated a negative effect on teacher morale, and led to discontentment with the use of these measures. A sometimes excessive emphasis on test score–based measures of teacher and school performance in some states and districts have led to high-profile cheating scandals in Atlanta and other locations, in turn, leading to further problems for morale in the teaching profession. As studies by David Figlio and others have shown, accountability systems can place strains on educational systems that lead to improper behavior and perverse incentives.

Federal endorsement of the use of value-added and other test score–based models of teacher performance as a component of teacher evaluation systems through the Race to the Top competition under former U.S. Department of Education Secretary Arne Duncan, promoted the widespread use of the measures throughout the nation. However, in response to the backlash by teachers' unions and other constituents, the newer reauthorization of the Elementary and Secondary Education Act, entitled the Every Student Succeeds Act, passed in 2015, no longer emphasizes teacher evaluation.

Research Studies

Over time, many research studies have weighed in on the issues of bias and imprecision in value-added models. Some studies have provided evidence to suggest that the potential for bias in teacher performance measures based on value-added models may be fairly low. In their study comparing experimental value-added estimates of teacher performance to earlier nonexperimental estimates in the Los Angeles Unified School District, Thomas Kane and Douglas Staiger found that the two sets of measures were similar. In another study that allayed concerns regarding bias, Raj Chetty, John Friedman, and Jonah Rockoff found that student achievement responded in expected ways to the entry and exit of teachers with differing value-added to a school. Studies by Brian Jacob and Lars Lefgren and by Douglas Harris and Tim Sass have found that value-added measures are positively correlated with the subjective judgments of school principals.

On the other hand, many studies, such as those by Daniel Aaronson, Lisa Barrow, William Sander and Dan McCaffrey, Tim Sass, J. R. Lockwood, and Kata Mihaly, have shown that value-added measures for individual teachers show a fair amount of instability from year to year, leading one to question how accurately they represent a teacher's true competence. Jesse Rothstein devised falsification tests that challenge the validity of value-added models of teacher performance in North Carolina; however, Dan Goldhaber and Duncan Chaplin, Josh Kinsler, Cassandra Guarino, Mark Reckase, Brian Stacy, and Jeffrey Wooldridge have shown that such tests can be misleading. Several studies by Steven Dieterle, Cassandra Guarino, Michelle Maxfield, Paul Thompson, Brian Stacy, Jeffrey Wooldridge, and others have explored different methods of computing value-added models and found that certain models do better than others in dealing with bias.

Current and Future Outlook

Current and Future Outlook

Although concerns remain about whether value-added models may be capable of evaluating all teachers fairly, their use has grown in district and state accountability systems throughout the nation, and research continues to compile more helpful information about their strengths and limitations. Most accountability systems to this day use value-added models as just one component of teacher evaluation, and the weight placed on it varies considerably from system to system.

All in all, value-added models, if carefully structured, can provide useful information to judge teachers, schools, and programs. However, a certain degree of caution is always needed when policy stakes in the form of rewards and sanctions are attached to the measures produced by these models.

Cassandra Guarino

See also Growth Curve Modeling; Teacher Evaluation

Further Readings

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. Journal of Labor Economics, 25(1), 95–135. doi:10.1086/508733

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. American Economic Review, 104(9), 2593–2632. doi:10.1257/aer.104.9.2593

Dieterle, S., Guarino, C., Reckase, M., & Wooldridge, J. (2015). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added. Journal of Policy Analysis and Management, 34(1), 32–58. doi:10.1002/pam.21781

Figlio, D. N. (2006). Testing, crime and punishment. Journal of Public of Economics, 90(4–5), 837–851. doi:10.3386/w11194

Figlio, D. N., & Winicki, J. (2005). Food for thought: The effects of school accountability plans on school nutrition. Journal of Public Economics, 89(2–3). doi:10.3386/w9319

Goldhaber, D., & Chaplin, D. (2015). Assessing the Rothstein falsification test: Does it really show teacher value-added models are biased? Journal of Research on Educational Effectiveness, 8(1), 8–34. Retrieved from http://dx.doi.org/10.1080/19345747.2014.978059

Guarino, C., Maxfield, M., Reckase, M., Thompson, P., & Wooldridge, J. (2015). An evaluation of empirical Bayes' estimation of value-added teacher performance measures. Journal of Educational and Behavioral Statistics, 40, 190–222. Retrieved from https://doi.org/10.3102/1076998615574771

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2015a). A comparison of student growth percentile and value-added models of teacher performance. Statistics and Public Policy, 2(1), e1034820. doi:10.1080/2330443X.2015.1034820

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2015b). Evaluating specification tests in the context of value-added estimation. Journal of Research on Educational Effectiveness, 8(1), 35–59. Retrieved from http://dx.doi.org/10.1080/19345747.2014.981905

Guarino, C., Reckase, M., & Wooldridge, J. (2015a). Can value-added measures of teacher performance be trusted? Education Finance and Policy, 10(1), 117–156. doi:10.1162/EDFP_a_00153

Guarino, C., Reckase, M., & Wooldridge, J. (2015b). Policy and research challenges of moving toward best practices in using student test scores to evaluate teacher performance. Journal of Research on Educational Effectiveness, 8(1), 1–7. Retrieved from http://dx.doi.org/10.1080/19345747.2015.986943

Harris, D., & Sass, T. (2011). Teacher training, teacher quality, and student achievement. Journal of Public Economics, 95(7–8), 788–812.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. Journal of Labor Economics, 26(1), 101–136. doi:10.1086/522974

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation (Working Paper 14607). National Bureau of Economic Research.

Kane, T., & Staiger, D. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains (MET project Research Paper). Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T., & Staiger, D. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project’s 3-year study. Seattle, WA: Bill and Melinda Gates Foundation.

Kinsler, J. (2012). Assessing Rothstein’s critique of teacher value-added models. Quantitative Economics, 3, 333–362. doi:10.3982/QE132View

Koedel, C., & Betts, J. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. Education Finance and Policy, 6(1), 18–42. doi:10.1162/EDFP_a_00027

McCaffrey, D. F., Lockwood, J. R., Louis, T., & Hamilton, L. (2004). Models for value-added models of teacher effects. Journal of Educational and

Behavioral Statistics, 29(1), 67–101. doi:10.3102/10769986029001067

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. Education Finance and Policy, 4(4), 572–606. doi:10.1162/edfp.2009.4.4.572

McGuinn, P. (2012). The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems. Washington, DC: Center for American Progress.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. Quarterly Journal of Economics, 175–214. Retrieved from https://doi.org/10.1162/qjec.2010.125.1.175

Sanders, W., & Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. Journal of Personnel in Education, 8, 299–311.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? Journal of Educational and Behavioral Statistics, 38(2), 142–171. Retrieved from https://doi.org/10.3102/1076998611432174

Thomas I. Vaughan-Johnston Thomas I. Vaughan-Johnston Vaughan-Johnston, Thomas I.

Leandre R. Fabrigar Leandre R. Fabrigar Fabrigar, Leandre R.

Values

Values

1787

1789

Values

Values are abstract principles about what is worth doing in life and how life should be lived. For example, *achievement* is a value that describes what is worth doing (e.g., working hard, learning) and how life ought to be lived (e.g., through striving for success such as through hard work and goal setting). Values are typically measured by having respondents rank the personal importance of a set of listed values. By collecting large samples of such responses, intervalue systems can be extrapolated (e.g., that valuing *achievement* is associated with valuing *power*). Understanding the origins and consequences of values has significance for educational research in that values help determine and prioritize the academic and life goals of people such as students and teachers and can predict their motivation to work, study, and achieve.

Values are a distinct member of a larger family of evaluative constructs. For instance, while values characterize broad views of what is important in life, *attitudes* are the evaluation of specific objects as good or bad. These are naturally related; for instance, individuals who value *equality* may have more favorable attitudes toward specific egalitarian political policies. Similarly, theorists have suggested that many animals beyond humans have *needs* (core survival drives, such as thirst, or group belonging). Many values represent intellectual abstractions of these needs, but these abstractions are considered unique to humans. For instance, a need to belong may promote valuing *benevolence* or *self-transcendence*. This entry first describes the structure, measurement, and stability of values and then discusses research investigating antecedents and some important consequences of value for human behavior.

Structure, Measurement, and Stability

Most psychologists posit a finite number of discrete human values but dispute the precise number and nature of these. For example, the Rokeach Value Survey lists two subtypes of values: 18 instrumental values (ways of living life, e.g., *honesty, courage, responsibility*) and 18 terminal values (end goals of life, e.g., *happiness, salvation, freedom*). Respondents rank the importance of values within each list.

By contrast, Schwartz's popular intervalue system posits 10 fundamental values, eliminating the instrumental-terminal distinction. These are arranged into four broad clusters: *openness to change*, *conservation*, *self-transcendence,* and *self-enhancement,* with the value clusters forming quadrants of a circle. A self-report measure, the Schwartz Value Survey, measures individuals' position in the circle by asking respondents to rate each value's importance on a semantic differential scale. For each value, some values are harmonious (adjacent on the circle), some irrelevant (90° away), and some antagonistic (180° away). Some supportive evidence for this structure has been demonstrated; for example, making particular values salient increases the accessibility of related values (those in agreement and disagreement with the primed value) and decreases the accessibility of irrelevant values.

Scholars are concerned with the universality of human values (that the same core set of values exist globally) and of value systems (that values are interrelated in the same ways, across cultures). Evidence supports both kinds of universality, particularly research using the Portrait Value Questionnaire, which has substantially reduced cross-cultural measurement variance issues. It should be stressed that this invariance of what cultures construe as values, and how values interrelate, does not mean that all cultures rank the values equivalently. Instead, cultures sharply vary in their prioritizing of values, with important political and economic consequences.

Although most measures of values are explicit self-reports, implicit measures of values have also been developed, such as the Value-SC-IAT. Such implicit measurements may be particularly useful when individuals are reluctant to express their true values due to situational pressures.

Values are usually considered to be stable across time. When change does occur, it happens among respondents who have expressed that they are dissatisfied with a value's rank. Psychologists usually study value change through manipulations. One example is the value-self-confrontation paradigm, in which participants are told that their peers differed from them on their value ratings. This produces convergence toward the peer ratings. Second, individuals will sometimes show value changes when they are simply asked to explain why they hold their values. This may occur because values can be unquestioned "truisms," which individuals are ill-prepared to defend.

Antecedents and Consequences

In earlier values research, there was an emphasis on how values forecast specific behaviors. To a large extent, the emphasis on correlational research renders ambiguous whether values are the cause and/or consequence of these variables. For example, when data suggest that prisoners rank honesty lower than a control group, it is unclear whether the low importance of honesty produces criminal activity and/or whether a criminal lifestyle encourages individuals to deprioritize honesty. In other cases, causality is clearer. For example, certain age trends have been noted, often following a curvilinear trend (e.g., values including *wisdom* and *imagination* seem to start at a low rank, peak in young adulthood, and decrease among seniors). Similarly, it is usually assumed that culture and personality are the causes, not consequences, of associated values.

Nonetheless, values have important consequences for human behavior. They predispose individuals to certain political, religious, and economic perspectives and guide the selection of career paths and academic focuses. They promote pro- or anti-social orientations toward social groups and provide standards against which the self and others can be compared. These standards are emotionally impactful: When individuals work against a personal value (e.g., writing a counter-value essay), they experience negative emotions proportional to the subjective importance of that value. A lower income is more distressing to individuals who value *materialism*.

Values are an integral part of several psychological theories, including terror management theory (where they afford long-term sources of self-esteem) and contemporary evolutionary theory (guiding individual survival and group cooperation needs). They are pivotal in research regarding attitudes (in that values may help to organize individuals' many specific object evaluations), culture (in that value differences may explain many cross-cultural differences and misunderstandings), and human development (in characterizing how values fluctuate across the life span). Thus, values play central roles in psychological theory and are an indispensable element of the field.

Thomas I. Vaughan-Johnston and Leandre R. Fabrigar

See also Attitude Scaling; Goals and Objectives

Further Readings

Maio, G. R. (2010). Mental representations of social values. Advances in Experimental Social Psychology, 42, 1–43.

Rokeach, M. (1973). The nature of human values. New York, NY: The Free Press.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in Experimental Psychology, 25, 1–65.

Dorothy J. Musselwhite Dorothy J. Musselwhite Musselwhite, Dorothy J.

Brian C. Wesolowski Brian C. Wesolowski Wesolowski, Brian C.

Variance

Variance

1789

1793

Variance

Variability is a numerical description that refers to the spread values within a given distribution. Variability can be described using four common measures: range, the mean deviation score, standard deviation, and variance. Variance is most useful in determining the overall spread of a data set. Range simply takes two data points, the lowest and highest values, and measures the space between them. The interquartile range also interprets the amount of spread within a set of data by using only the middle 50% of scores. Unlike range, variance takes into account every data point within the set and measures each distance from the mean. Variance is calculated using a number of values from data and the associated distribution. Specifically, variance requires raw data scores and the sample size of the data.

The sum of squares (or sum of squared errors) is also used as a method to determine total spread or dispersion. The problem with sum of squares is this value cannot be compared across samples that differ in size. Variance deals with average spread, which is comparable across groups that may change sizes.

Variances are most impactful when comparing multiple distributions. Furthermore, this statistic becomes the basis for various statistical comparisons, including an analysis of variance (ANOVA). In experiments, people or groups undergo treatments to potentially elicit various responses. If all scores, or responses, are different, variability will be large. If all scores are exactly the same, variability will be zero.

Variance is used in many different experimental designs. After discussing a brief history, variance is described not only by its formulae but also by its conceptual properties. Variance is then applied to statistical testing, including ANOVA and multiple regression.

History

Although credit for the concept of variance is given to Ronald Fisher, it is apparent that the concept of variance existed long before Fisher, in the work of Carl Friedrich Gauss and his endeavor to estimate the locations of stars. Within his search, Gauss encountered a probability distribution with deviations that may have given way to the current concept of variance.

Fisher introduced the concept of variance in his 1918 paper “The Correlation Between Relatives on the Supposition of Mendelian Inheritance.” Fisher was the first to introduce the test now known as ANOVA. Most of his later work involved significance and hypothesis testing.

Formulae

When research designs are considered, it is the goal of the researcher to collect as many real-world observations as possible, knowing that the collected observations will not account for all possible observations that exist. As mean and variance are calculated from these select observations, it can be assumed that the values will not perfectly match the values that would have been obtained using every possible observation. Therefore, the researcher must estimate the mean and variance using an equation to account for observational bias. Two formulae exist to calculate variance, one describing the population variance and the other describing the sample variance.

Population variance is calculated using the mean of the squared deviations from the distribution mean. Deviation is the difference between a specific value of data and the distribution’s mean. The formula for population variance can be defined as

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N},$$

where σ^2 represents the population variance, $X - \mu$ represents the deviation between each score and the population mean, and N is the population size. This formula is conceptual and represents a hypothetical population. A population refers to the entire group being defined. For example, if the experiment involves 20-year-old females, the population would be every 20-year-old female on the planet. Because the population in question is not always readily available, the researcher must obtain a sample or a proportion of the population. This sample is meant to provide a snapshot of what the population actually looks like.

If a population is extremely large, specifically when it is not possible to count every observation available, a sample of the population must be used to compute variance. This formula for sample variance is used most frequently and can be defined as:

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1},$$

where S^2 is the variance, represents the deviation between each score and the sample mean, and n is the sample size. The numerator is often referred to as the sum of squared errors, or *SS*. Using $n - 1$ as the denominator for sample variance is called Bessel’s correction. The result is an unbiased estimator of the population variance called the corrected sample variance or unbiased sample variance. By taking the sum of the deviations and dividing by Bessel’s correction, what results is an average of all deviations. The sum of the deviations of any data set will always be zero; therefore, the deviations must be squared before being summed to account for negative values. If the distribution mean is not easily attainable, another formula may be used while employing raw data:

$$S^2 = \frac{\sum X^2}{N} - \frac{\left(\sum X\right)^2}{N^2}.$$

Both formulas yield identical answers but are dependent on the data and statistics available.

Once variance is calculated, variability can be further explored through another statistic, standard deviation. Standard deviation is another method of determining spread, specifically focusing on the data and its location from the mean. Variance is more about the average distance away from the mean and is reported in square units. Standard deviation can be calculated by taking the square root of the variance:

$$S = \sqrt{S^2}.$$

By taking the square root of the variance, the units are also restored to their original form. Standard deviation is more appropriate when units do not make sense to be squared (e.g., people).

Properties of Variance

Variance will always yield a nonnegative value because squares are positives or zero. If a constant is applied to all values of a variable through addition, variance will remain unchanged. In other words, variance is invariant as it relates to the location parameter. However, if a constant is applied through multiplication, or scaling, variance will adjust by the square of that constant.

Variance is often the preferred method of determining the spread of a given data set due to the fact that the variance of the sum of the uncorrelated (independent) random variables can be calculated using the sum of the variances. The Bienaymé formula is a derivation of the aforementioned concept. This derived formula was discovered in 1853 and is defined as:

$$\mathrm{Var}\,(\bar{X}) = \frac{\sigma^2}{n},$$

where refers to the sample mean. The Bienaymé formula can be interpreted to indicate that the mean in every experiment with a sample size of *n* will fall within ±σ*n* of the true population 68% of the time. The 68% refers to the area under the normal curve within ±1 standard deviation from the mean. This formula also shows that the variance of the mean will decrease when the sample size, *n*, increases when the variables have the same variance.

Correlated variables are treated differently in terms of the variance of their sum. Instead of adding the variances as in uncorrelated variables, the variance of the sum of correlated variables is equal to the sum of their covariances. Because the variables are now related, the covariance must be used to determine how the variables change together. Variance of the mean is calculated differently when the variables are correlated and have equal variance:

$$\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2,$$

where ρ refers to the average correlation of distinct variables.

The variance sum law states that the variance of a difference between two independent variables can be calculated using the sum of their respective variances. For example, a test is given to fifth graders to determine level of mathematics ability based on gender. Therefore, boys and girls are split into two subgroups for testing. The two populations are fifth grade boys and fifth grade girls. For each group, the variance is calculated separately. When the two variables are independent, the variance can be added together to show the variance of the sum. The standard error of that variance can be calculated by taking the square root of the aforementioned sum.

Variance should always be reported in units squared. For example, if data points are measured in centimeters (cm), variance will be reported as cm^2. This reporting can be problematic, depending on the units. It may be reasonable to report cm^2 or kg^2, but, as noted earlier, when the units are people, it does not make sense to report people. When this is the case, standard deviation is often reported instead, by taking the square root of the variance and defining by the same units as the original variable.

Between Groups and Within Groups Variance

Within one set of data exists two types of variance: between-groups variance and within-groups variance. These types of variance are most relevant when conducting an ANOVA, specifically when multiple groups are being compared. Between-groups variance is frequently referred to as explained or systematic variance. Systematic variance is derived from the independent variable or among the groups in the study. Explained variance is accompanied by a direct manipulation from the researcher. Within-groups variance refers specifically to the error variance and comes from within each group. It does not come from the independent variable. This unsystematic variance is due to variations in performance, by factors outside of the manipulation directly from the researcher. This type of variance is not consistent and may fluctuate from one testing administration to the next. Comparing examinees in the same group would result in within-groups variance. Comparing examinees in two different groups would result in between-groups variance. Both of these variances will impact the validity and reliability of tests.

Homogeneity of Variance

The assumption of homogeneity of variance, or homoscedasticity, is used when several groups of examinees are tested for differences in group means. In such a design, it is assumed that all samples come from populations of equal variance. This assumption may also be applicable in a correlational design, where the variance of the dependent variable should be equal at all levels of the independent variable. Any fluctuation of the variance would be a violation of the assumption of homogeneity of variance.

If the spread around the scores are roughly the same from each sample taken, the assumption has been satisfied. If the spread fluctuates greatly from sample to sample, this assumption has been violated and is called heterogeneity of variance or heteroscedasticity. The violation of this assumption is most important when group sizes are unequal.

Homogeneity of variance is most important when using the method of least squares to estimate parameters. Unequal variances will create a bias and an inconsistent estimate of the standard error. Inconsistent standard error will directly affect confidence intervals and significance tests. Levene’s test will test the null hypothesis, stating different groups have equal variances. This test is executed through a one-way ANOVA, where a significant p value indicates the variances among groups are significantly different. This significant p value from Levene’s test, therefore, indicates a violation of the assumption of homogeneity of variance. However, a large sample size may result in an incorrect reporting of significance. Small differences among group variances when the sample size is large will generate a significant Levene’s test. It is important to note that Levene’s test works best with equal group sizes.

Another way to compare groups is to use the variance ratio or Hartley’s F_{max}. Hartley’s F_{max} is equal to the ratio of the largest group variance to the smallest group variance. This value is then compared to critical values that are dependent on the group size and the number of variances being evaluated. The variance ratio is expected to be smaller than the critical values specified by Hartley’s table to be nonsignificant.

Homogeneity of variance should be evaluated by both Levene’s test and the variance ratio, as false significance may by reported by one test.

Variance in Multiple Regression

In a multiple regression design, fit of the regression line is measured by correlation (r) and r^2. In relation to variance, however, R^2 is often reported specifically in reference to explained variance. R^2, often called the coefficient of determination, may be reported as a decimal ($R^2 = .71$) or a percent. When reported as a percentage, it is often stated as the percentage of variance in the independent variable as explained by the model. This value does not make the implication that causation occurs. Using the output from statistical analysis software, or from hand calculations, R^2 can be calculated using the following formula:

$$R^2 = \frac{SS_M}{SS_T},$$

where SS_M refers to the model sum of squares and SS_T refers to the total sum of squares. The square root of this value yields Pearson’s correlation coefficient, r. Adjusted R^2 will often be reported in addition to R^2, which removes the bias associated with the latter by diminishing its value.

Dorothy J. Musselwhite and Brian C. Wesolowski

See also Analysis of Variance; Correlation; Multiple Linear Regression; Normal Distribution; Standard Deviation; Standard Error of Measurement

Further Readings

Boyle, J. D., & Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York, NY: Schirmer Books.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. London, UK:

Sage.

Gravetter, F. J., & Wallnau, L. B. (2011). Essentials of statistics for the behavioral sciences. Belmont, CA: Wadsworth.

Kubiszyn, T., & Borich, G. (2003). Educational testing and measurement: Classroom application and practice. New York, NY: Wiley.

Payne, D. A. (2003). Applied educational assessment. Toronto, Canada: Wadsworth.

Reid, H. M. (2014). Introduction to statistics: Fundamental concepts and procedures of data analysis. Thousand Oaks, CA: Sage.

Xi Wang Xi Wang Wang, Xi

Craig Stephen Wells Craig Stephen Wells Wells, Craig Stephen

Vertical Scaling Vertical scaling

1793

1796

Vertical Scaling

In educational measurement, evaluation of students’ growth in knowledge, skills, or aptitude over time has important implications to educators and policy makers. One common method of measuring growth is through the use of vertical scales. Vertical scaling is a special form of linking, which aims at adjusting score differences on tests that differ in content and/or difficulty. Successful linking enables educators and researchers to make statements such as “a score of 155 on Test *X* corresponds to a score of 365 on Test *Y*,” where the content specifications in Tests *X* and *Y* need not be exactly the same. Vertical scaling is intended to establish the concordance relationship between scores on tests measuring educational achievement or aptitude at different academic grades. Proper vertical scaling enables educators and researchers to answer questions such as “given a raw score of 20, what are the corresponding scale scores in Grade 3 and 4, respectively?” Furthermore, the vertically scaled scores in different grades are directly comparable with each other. However, different from “equating,” which conducts linking on test forms with very similar content and statistical specifications, there often exist systematic differences in both content and difficulty between test forms in vertical scaling, so vertically scaled scores on different forms cannot be regarded as interchangeable. This entry reviews the applications, advantages, implementations, and limitations of vertical scaling in educational measurement.

Applications and Advantages in Educational Measurement

Some well-known practices of vertical scaling in educational achievement

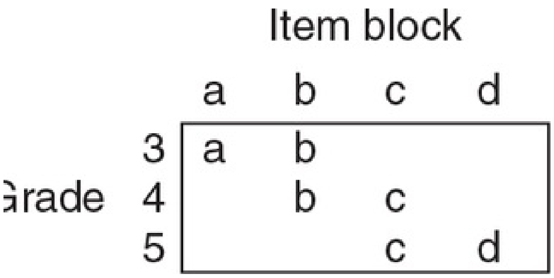
Some well-known practices of vertical scaling in educational achievement batteries include its applications on the Iowa Tests of Basic Skills, the California Test of Basic Skills, and the California Achievement Tests. As vertical scaling creates a common scale for students' scores over grades, one can compare students' achievement from one grade to another so as to evaluate students' growth pattern over time. In addition to evaluating students' performance, vertically scaled scores are used as the input in some value added models to evaluate a teacher's contribution to students' achievement. Another advantage of vertical scaling is that it also places item statistics from different grades onto the same scale, which may lead to more efficient use of the field-test items. For instance, some field-test items initially constructed for Grade 4 may be found more appropriate for Grade 3, so instead of being discarded from the item pool, those items can be used to construct the test forms for Grade 3.

Implementations

Data Collection Design

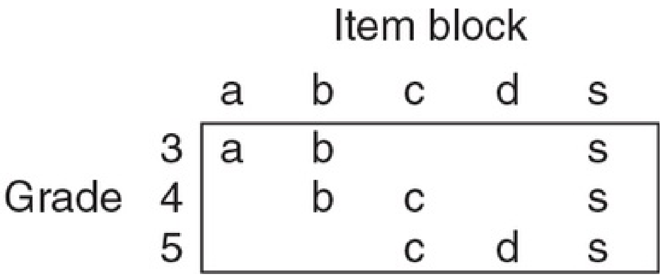
To create vertical scales, three data collection designs are often used to adjust test form differences: a common-items design, scaling test design, and equivalent-groups design. Figures 1 and 2 show the basic setups for each of the three designs. Each design has some variations. As shown in Figure 1, a common-items design administers a common block of items between every two adjacent grade levels. Under this design, students' responses to the common items serve as an anchor to estimate their ability differences and further to put their scores on the entire test onto a same scale. Linking is conducted between every two adjacent grades. For example, fourth graders' performance is linked to the third graders' through block *b*, the fifth graders' is linked to the fourth graders' through block *c*, and so on. The choice of common items plays a critical role in the linking result as well as the growth measurement. With this design, growth is measured grade by grade over the content areas represented by each common item block. For example, the growth from Grades 3 to 4 is measured on the content areas covered by block *b* but not on the unique content areas covered by block *c*.

Figure 1 Examples for the common-items design (left) and scaling test design (right). Letters *a*–*d* represent item blocks administered to each grade and *s* represents the scaling test



Item block

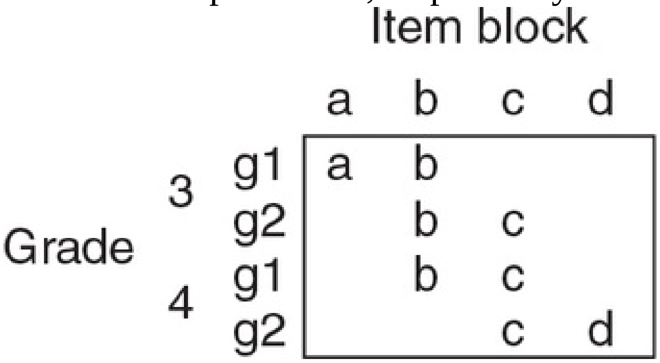
		a	b	c	d
Grade	3	a	b		
	4		b	c	
	5			c	d



Item block

		a	b	c	d	s
Grade	3	a	b			s
	4		b	c		s
	5			c	d	s

Figure 2 Examples for equivalent-groups design. “g1” and “g2” represent random Groups 1 and 2, respectively



Item block

			a	b	c	d
Grade	3	g1	a	b		
		g2		b	c	
	4	g1		b	c	
		g2			c	d

The scaling test design, in comparison, can be used to measure growth over the range of content areas taught in different grades. A scaling test is constructed to cover the content areas across all grade levels. For instance, the scaling test in Figure 1 could consist of the content areas in blocks *a*, *b, c*, and *d*, simultaneously. The scaling test is administered to all grade levels, and responses to the scaling test are used to define the score scale. In addition to taking the scaling test, each student takes the test form appropriate for a specific grade level, and scores on each grade-level test are linked to the scale constructed by the scaling test.

In equivalent-groups design, students in each grade are randomly assigned to take either a test for the given grade level or a test for the adjacent grade levels. For instance, as shown in Figure 2, about half of students in Grade 4 are assigned the Grade 4 test, and the other half are assigned the Grade 3 test. With the random assignment, the two groups taking each test can be regarded as equivalent, and scores on the two test levels can be linked through equivalent groups.

Statistical Methods for Vertical Scaling

Three methods are commonly used to establish a vertical scale: Hieronymus scaling, Thurstone scaling, and the item response theory (IRT) scaling. The Hieronymus and Thurstone scaling are based on raw scores, and the IRT scaling assumes a person’s responses follow a probabilistic model. With the scaling test design and the common-items design, Hieronymus scaling conducts linking based on the test score distributions in each grade on the common items. The medium score at each grade level is used to define the scale. For instance, the medium score for the third graders on the scaling test is defined as a scale score of 3 and the medium score for the fourth graders is defined as a scale score of 4. The remaining raw scores on the common items can be transformed to the scale scores according to test developers’ theory about the year-to-year growth. For instance, a transformation can be made such that the within-grade score variability decreases or increases as the grade increases. After the score scale is constructed based on the common items, the scores on each grade-level test are linked to the score scale.

Thurstone scaling, also referred to as Thurstone’s absolute scaling method, assumes that scores are normally distributed within each group. Thurstone scaling begins by transforming raw scores on common items to normalized scores for each group. The score scale can be predefined by fixing the mean and standard deviation of the scale scores for one group. For example, the score scale can be defined by setting the mean and standard deviation of the third graders to $\mu 3(sc)$ and $\sigma 3(sc)$, respectively, and the normalized scores obtained in the first step for the third graders are then linearly transformed to the scale scores. The scale scores for the other groups are found through the relationship of their normalized scores with the baseline group that is used to define the score scale. For instance, to determine the scale scores for the fourth graders, for each given raw score y on the common items, the normalized score in Grade 4 (i.e., $z4 \times (y)$) is paired with that in Grade 3 (i.e., $z3 \times (y)$). Across different values of y, $z4 \times (y)$ and $z3 \times (y)$ are expected to have a linear relationship, that is, $z4 \times y = Agz3 \times y + B$. The coefficients A and B can be determined using the mean and standard deviation of $z4 \times y$ and $z3 \times y$, and the mean ($\mu 4(sc)$) and standard deviation ($\sigma 3(sc)$) of scale scores for Grade 4 are given by $Ag\mu 3(sc)$+B and $Ag\sigma 3(sc)$, respectively. After finding the mean and standard deviation of the scale score in each grade, each raw score on the common items is transformed to the scale score by multiplying its corresponding z score with the scale score standard deviation and then adding the scale score mean, that is, $sc3y = \sigma 3(sc)z3 \times y + \mu 3(sc)$ and $sc4y = \sigma 4(sc)z4 \times y + \mu 4(sc)$, where sc3y and sc4y represent the scale score in Grades 3 and 4 corresponding to a given raw score (y) on common items. After the score scale is constructed based on the common items, the scores on each grade-level test are linked to the score scale through the common items.

IRT scaling is based on the parameter invariance property in IRT models. IRT models describe the probability of a correct response on an item as a function of the item parameters and a person's latent ability (θ). The invariance property implies that the item parameter values in the IRT model remain the same across different ability groups. IRT scaling can be conducted through either concurrent or separate estimation, both of which make use of responses to the common items between grade levels. In concurrent estimation, responses from different grade levels are combined into the same file, with responses to items not administered in a certain grade being coded as "not reached." Parameter estimation for items in all test forms is conducted simultaneously and the ability estimates of students from different grade levels are placed onto the same scale automatically.

In separate estimation, parameters of items from different test forms are estimated separately. Due to the parameter invariance property, the common items have the same parameter values regardless of the group from which they are estimated. However, because θ does not have an inherent scale, the item parameters may appear different when the scale of θ is fixed differently for different groups, but the discrepancy is subject to a linear relationship. For example, for the same group of students, consider two approaches to fix the scale for θ. The first approach fixes θ on a scale with a mean of $\mu1$ and a standard deviation of $\sigma1$, while the second approach uses a linear transformation to fix the mean and standard deviation to be $\mu2$ and $\sigma2$, where $\sigma2 = A\sigma1$ and $\mu2 = A\mu1 + B$. If the item discrimination and difficulty parameter are a and b corresponding to the first scaling method, the second scaling method will result in the two parameters being a/A and $Ab + B$, respectively, so as to produce invariant item response functions. Similarly, if the mean and standard deviation of θ are fixed to be the same in two different ability groups, which is a standard practice in a lot of IRT estimation software, item parameter estimates on the same items from separate estimation runs will follow the linear relationship described earlier. After estimating the linking constants A and B through some statistical methods, θ in one group can be linearly transformed to be on the same scale as that in the other group.

Challenges and Limitations

Michael J. Kolen and Robert L. Brennan pointed out that the differences in content and statistical specifications between different grade-level tests could limit the score interpretations in vertical scaling. The difference in test difficulty may result in incomparable measurement precision for the same scaled score on different tests. For instance, the Grade 3 test may have larger measurement error at a higher scaled score than the Grade 4 test, so the linking between the two tests is less accurate at higher scaled scores. The difference in test content may result in different meanings for the same scaled score. For instance, if the Grade 3 test does not cover “geometry,” but the Grade 4 test does, the same scale score does not reflect students’ ability on geometry in the Grade 3 test, but that will be reflected in the Grade 4 test.

In addition to the limitations in score interpretability, existing research shows there are several challenges in vertical scaling that may affect its practical uses. First of all, studies have shown that vertical scaling results are likely to be affected by many factors, such as the data collection designs, statistical scaling methods, scoring methods, and examinee samples. No combination of those factors have been found to work best in a generalized context. As a function of those factors, inconsistent grade-to-grade growth patterns are also observed among different studies. The inconsistent growth pattern creates some difficulty in choosing and/or evaluating a vertical scale. Second, Derek C. Briggs questioned the assumption that the vertical scale has interval properties, and he argued that this assumption needs to be justified to support the equal interval interpretation in growth evaluation. If a vertical scale does not have interval-level properties, one unit of change in the scale score will represent different levels of growth as a function of a student’s starting point, which is analogous to measuring length using an unstandardized ruler with the same unit representing different lengths. Third, when IRT methods are used for linking, commonly used IRT models assume different tests measuring the same content have a unidimensional structure, but this assumption is unlikely to hold over multiple grades.

Xi Wang and Craig Stephen Wells

See also Score Linking; Standard Setting; Standardized Tests

Further Readings

Briggs, D. C. (2013). Measuring growth with vertical scales. Journal of

Educational Measurement, 50(2), 204–226. doi:10.1111/jedm.12011

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). Linking and aligning scores and scales. New York, NY: Springer.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices (3rd ed.). New York, NY: Springer.

Anh Andrew Nguyen Anh Andrew Nguyen Nguyen, Anh Andrew

Leandre R. Fabrigar Leandre R. Fabrigar Fabrigar, Leandre R.

Visual Analog Scales

Visual analog scales

1796

1800

Visual Analog Scales

Visual Analog Scale (VAS) is a type of psychometric scale that uses a continuous measurement indicator rather than the multiple discrete indicators more common in many other types of measurement scales. Respondents to VAS make a subjective judgment on where their answer lies on a continuum and then mark their response on a VAS line. The VAS tends to be used to measure the same kinds of psychological phenomena that traditional rating scales are used to measure, such as mood, satisfaction, well-being, or psychological pain. For instance, respondents may be shown a horizontal line and asked to mark on the line their level of pain, where the left-most point represents *no pain* and the right-most point represents *extreme pain*. Researchers can then use the continuous distance between the mark and an end point as a subjective indicator of pain. In contrast, discrete scales, such as rating scales, will ask respondents a similar question, but with response options broken up into discrete points, which then serve as the subjective indicator of pain.

The continuous nature of measurement of the VAS allows respondents to potentially provide more precise answers and for a finer distinction between subjective states than traditional rating scales allow. Because traditional rating scales constrain respondents to predetermined ratings, such as 1–7, and because most psychological phenomena, such as pain or mood, have no objectively discrete intervals, the VAS should have an advantage over traditional rating scales in that it allows for the maximum possible distinction, limited only by the physical width of the response mark itself. This entry briefly reviews the administration and scoring of the VAS and then discusses variations in the VAS.

administration and scoring of the VAS and then discusses variations in the VAS.

Administration and Scoring

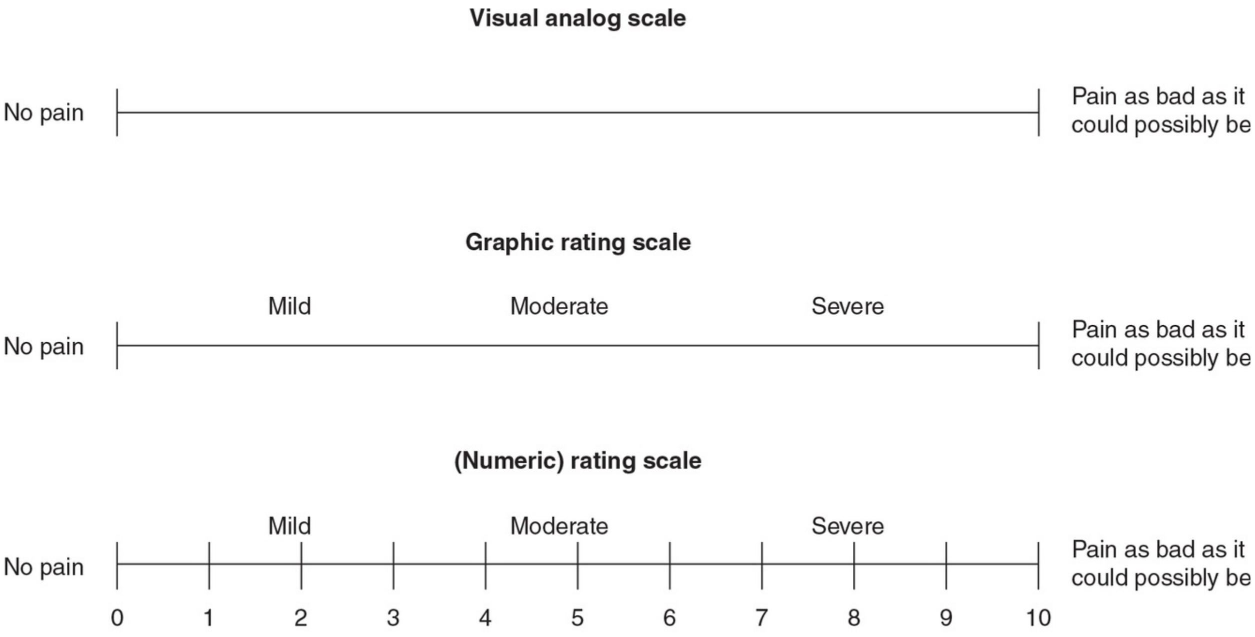
Respondents are asked a question, such as “How much pain are you in?” and are instructed to mark the location on a VAS line that best reflects their answer (in this case, the pain they are feeling). Administrators then measure the distance from the left end point to the mark, and this distance is the respondent’s VAS score. Specific characteristics of the line and anchor descriptors vary depending on the version of the VAS used.

Variations

Types of VAS

There are generally three separate types of VAS: the traditional VAS, the Graphic Rating Scale (GRS), and the Numeric Rating Scale (NRS; see Figure 1 for samples of each type). The GRS is a form of the VAS where descriptor labels, such as “mild,” “moderate,” and “severe,” are placed at appropriate intervals on the evaluation line. The GRS may help participants more easily recognize where to mark their position on the evaluation line. However, the position of descriptor labels on the evaluation line can influence the distribution of ratings, so care should be taken when using the GRS.

Figure 1 The Visual Analog Scale, Graphic Rating Scale, and (Numeric) Rating Scale.



The NRS is often categorized as a type of VAS, but much more closely resembles a rating scale in that it does not use a continuous response line, but rather a line with discrete numbers representing equal intervals. Similar to rating scales, respondents select, or otherwise indicate, the number that best matches their response to the question. Indeed, the NRS shares more common elements with traditional rating scales than the VAS, and comparisons between the NRS and the VAS should be seen as comparisons between rating scales and the VAS.

Thus, the scales can be understood as ranging from least to most continuous, with the NRS the least continuous, the GRS moderately continuous, and the VAS the most continuous. There is some evidence that as the type of VAS scale becomes less continuous and more response indicators are added, respondents are better able to understand the scale, and consequently, response compliance is increased.

Line Length and Properties

Although the most common presentation form of VAS uses a horizontal line 100 millimeters in length, there are a number of VAS variations in use with altered line lengths or properties. One such variation, the vertical VAS, uses a vertical line instead of a horizontal line. Although there is some research suggesting that the respondents are more likely to overestimate their scores on the vertical VAS and that the vertical VAS is more sensitive to changes than the horizontal VAS, this research has been limited to small sample sizes, and its effects have not been well replicated. There is little evidence that the length of the line significantly affects respondent scores; whereas due to mechanical reasons, very short lines will present a greater challenge for respondents to accurately mark down their intended response. Because the VAS can be difficult to understand for some respondents, and because they sometimes provide responses that fall beyond the boundary of the VAS line, it is generally recommended that clear end points be created, with descriptors placed adjacent to the ends of the lines rather than positioned below the ends of the lines.

Unipolar and Bipolar VAS

The VAS can be administered as a unipolar or bipolar scale. The unipolar VAS measures characteristics or properties that range from complete absence to overwhelming presence. Anger, for example, is a unipolar property that ranges from *absence of anger* to *extreme anger*. Although the unipolar VAS is more widely in use, there are concepts that may be more easily assessed with a bipolar VAS. The bipolar VAS measures characteristics or properties that range from overwhelming presence of one property to overwhelming presence of an incompatible, opposite property. A VAS that tries to assess mood, ranging from *very happy* to *very sad*, is an example of a bipolar VAS.

Reliability

Although the reliability of the VAS has been investigated, the majority of this research has been conducted within the context of pain research. Multiple studies have shown the VAS to have test–retest reliabilities of .80 or higher, depending on the construct and the context under which it is measured. For the most part, the VAS and rating scales produce comparable reliability ratings across a variety of constructs and contexts.

The reliability of different aspects of the VAS has been assessed. One aspect, the length of the evaluation line, shows no significant difference in mean error for lines of 5-, 10-, 15-, and 20-cm lengths. Comparisons between the vertical and horizontal VAS generally result in high correlations between the two versions, although participants have been noted to express a preference for the horizontal over the vertical VAS. There is also some evidence that reproducibility of marks along the VAS line varies along the length of the line, with the midpoint least reproducible and increasingly reproducible toward the extremes. Overall, most analyses of the reliability of the VAS, especially within the domain of pain research, as reviewed by Marianne Jensen Hjermstad and colleagues in 2011, have found it to produce reliability scores comparable to that of rating scales used within the same studies.

Validity

A number of investigations into the validity of the VAS have been conducted, with the majority of these studies having focused on criterion-related validity within the context of judgments of pain. The VAS has been found to be moderately to highly correlated with established, validated criterion measures, such as gold standard measures of pain or quality of life. Similarly, researchers have found that the VAS was able to detect varying levels of pain as applied through thermal stimuli at varying temperatures. In addition, when used to assess the efficacy of analgesics on pain following the removal of a molar tooth, the VAS was shown to be more highly sensitive to changes in pain than comparable numerical rating scales. Other assessments of its validity have found the VAS to be capable of differentiating between separate categories of pain intensity. However, at least for pain assessments, rating scales have generally provided comparable results to the VAS. Overall, the VAS can be considered to demonstrate validity and reliability comparable to that of similar rating scales for a number of phenomena.

A 2011 review by Hjermstad and colleagues that investigated the usage of pain scales, including the VAS and NRS, showed a moderate trend toward a preference for the NRS over the VAS among both administrators and the respondents. Although these measures generally correlate highly with each other, in studies where the NRS was preferred over the VAS, reasons listed included lower compliance rates on the VAS and generally lower error rates and higher validity for the NRS. Respondents, particularly children or the elderly, similarly tend to subjectively prefer the NRS due to the difficulty of conceptually understanding the VAS. Other comparisons involving the use of VAS and rating scales to measure quality of life indices show comparable results between the types of scales. On occasion, when the VAS and rating scales were compared and produced similar results, those who completed the VAS were more likely to misunderstand and produce errors in response. Careful training on the part of scale administrators, as well as careful instruction for respondents on how to properly use the VAS, is recommended to help lower error rates on the scale.

Strengths and Limitations

The VAS has emerged as a viable option for researchers who wish to quickly gather data. Perhaps, the major strength unique to the VAS is its presumed ability to capture responses on a continuum as opposed to discrete intervals or categories. This characteristic can prove to be an advantage when measuring phenomena without clear jumps between intervals, assuming the more precise responses are highly reproducible and do not reflect underlying error. Thus, even though the VAS has found common usage in clinical contexts, it can also be theoretically used in other contexts, such as in education, by administrators or educators to assess students’ reading/writing ability, or by students to report ratings of their interest in particular subjects. However, researchers considering use of the VAS in educational settings should note that some research suggests that very young children (e.g., children under the age of 7) may, at least in some contexts, have difficulty understanding the procedure well enough to accurately complete the measure.

Commentators have noted other strengths such as adaptability across languages and cultures, flexibility of measurement across multiple different types of constructs, and simplicity and ease of use. However, these strengths are also true of traditional rating scales.

Some methodologists have commented on the ease with which the VAS can be scored. Owing to the requirement that scorers must measure each response, the traditional pen-and-paper VAS is naturally slower to score than a categorical or discrete response scale. Other objections include the specificity of photocopied materials, as well as the mobility required to mark down responses, which may not be met for hospitalized patients in severe pain. However, the increasing trend of administering the VAS online has alleviated some of these reproduction-and mechanical-level issues. Ease of scoring is increased through the automation of many of the VAS procedures, including the measurement step, and reproduction errors are virtually eliminated with lines displayed on standardized screen resolutions. Overall, the online VAS shows promise as the most practical and easy-to-use version of the VAS.

VAS Versus Other Rating Scales

The VAS, as with every other measure, has certain strengths and drawbacks. Usage of the VAS is most conceptually appropriate for cases, either with populations and/or constructs, where there is reason to expect respondents to be capable of producing finer distinctions between subjective states than the intervals provided by traditional rating scales. In cases where there is little reason to believe that respondents have the capability or motivation to make these fine distinctions, there may be little advantage to a VAS over similar rating scales. Indeed, given the greater difficulty in understanding the VAS, usage of the VAS over rating scales in these cases may even prove counterproductive.

Anh Andrew Nguyen and Leandre R. Fabrigar

See also Rating Scales; Scales

Further Readings

Ahearn, E. P. (1997). The use of visual analog scales in mood disorders: A critical review. Journal of Psychiatric Research, 31(5), 569–579.

Hjermstad, M. J., Fayers, P. M., Haugen, D. F., Caraceni, A., Hanks, G. W., Loge, J. H., & European Palliative Care Research Collaborative. (2011). Studies comparing Numerical Rating Scales, Verbal Rating Scales, and Visual Analogue Scales for assessment of pain intensity in adults: A systematic literature review. Journal of Pain and Symptom Management, 41(6), 1073–1093. doi:10.1016/j.jpainsymman.2010.08.016

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of Visual Analogue Scales as ratio scale measures for chronic and experimental pain. Pain, 17(1), 45–56.

Reips, U. D., & Funke, F. (2008). Interval-level measurement with Visual Analogue Scales in Internet-based research: VAS generator. Behavior Research Methods, 40(3), 699–704.

Wewers, M. E., & Lowe, N. K. (1990). A critical review of Visual Analogue Scales in the measurement of clinical phenomena. Research in Nursing &

Health, 13(4), 227–236.

Williamson, A., & Hoggart, B. (2005). Pain: a review of three commonly used pain rating scales. Journal of Clinical Nursing, 14(7), 798–804.

W

Nicholas F. Benson Nicholas F. Benson Benson, Nicholas F.

Ashley Donohue Ashley Donohue Donohue, Ashley

Emily Ward Emily Ward Ward, Emily

W Difference Scores

W Difference scores

1801

1802

W Difference Scores

The W scale was developed by Richard Woodcock and Marshall Dahl in consultation with Benjamin Wright. The W scale is simply a transformation of the ability/item score from a Rasch analysis that uses a logarithm with base 9 (log9) instead of the more common base e (ln). Base 9 was used because Woodcock believed it aided in interpreting the difference between personal ability and item difficulty values.

In the simplest Rasch model, the probability that person *n* correctly answers item *i*, P_{ni}, is

$$P_{ni} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)},$$

where B_n is person *n*'s ability (on a logit scale) and D_i is the item's difficulty (on the same logit scale as B_n). Equation 1 can be rearranged to isolate the relation between *B* and *D*:

$$\mathrm{In}\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i.$$

Converting *B* and *D* to the *W* scale simply involves the following linear transformation:

$$W = 9.1024\ (A) + C,$$

where *A* is either the person’s ability (*B*) or the item difficulty (*D*), and *C* is some arbitrary constant used to reduce the likelihood of having a negative value. Originally, *C* was 100, but a value of 500 is used in most current applications. A value of 500 is customarily set to be the ability on the measured trait associated with a student beginning fifth grade (grade norms) or a child of age 10 years, 0 months (age norms). This adjustment is made to set a reference point for proficiency, as W ability scores are intended to be used for measuring change in proficiency over time.

The 9.1024 multiplier changes the scale of the ln logit in a way that is equivalent to using a value of 20 for the base 9 logit. Using the logarithm change-of-base formula, this value can be derived as:

$$\frac{20\log_9 M}{\text{In } M} = \frac{20}{\text{In}9} \approx 9.104,$$

for any value of $M > 0$.

The version of Equation 1 that uses W scores is

$$P_{ni} = \frac{\exp\left(\dfrac{W_{B_n} - W_{D_i}}{9.104}\right)}{1+\exp\left(\dfrac{W_{B_n} - W_{D_i}}{9.104}\right)},$$

where W_* is the result of applying Equation 3 to either *Bn* or *Di*.

If, say, (i.e., the person’s ability is the same as the item difficulty), the person has a 50% chance of answering the item correctly. When a , the probability of a correct response is greater than .50. Likewise, when , the probability of a correct response is less than .50. Differences of +10, +25, and +50 correspond to probabilities of .75, .94, and .995, respectively, while differences of −10, −25, and −50 correspond to probabilities of .25, .06, and .004, respectively.

For a given group (e.g., age, grade), one can calculate the median W score, which is typically called the Reference W. The difference between a person n's W score and the Reference W score is called the W Difference score:

$$W\text{ Difference} = W_{B_n} - \text{Reference }W.$$

W difference scores are used in a number of commercially available tests, perhaps most notably the *Woodcock-Johnson IV Tests of Achievement* (Schrank, Mather, & McGrew, 2014), *Woodcock-Johnson IV Tests of Cognitive Abilities* (Schrank, McGrew, & Mather, 2014), and *Stanford–Binet Intelligence Scales-Fifth Edition* (Elliott, 2007). The *W* difference score is the value from which other scores (e.g., standard scores, percentile ranks) are derived. Adjustments to the formula for calculating W difference scores can be used to obtain other useful measures of growth. For example, the relative proficiency index is a modification of the W difference score, where the reference W is set at a value of 20 W units below the median. This adjustment facilitates prediction of success with items that same age or same grade peers answer correctly 90% of the time. The relative proficiency index is expressed as a fraction where the denominator is set to 90 to represent the probability of success for same-age or same-grade peers, depending on the type of norms used, and the numerator represents the probability of success for a given examinee. Thus, a ratio of 40:90 would indicate that a person has a 40% chance of responding correctly when same-age or same-grade peers have a 90% chance of responding correctly.

In addition to the measures of growth, W difference scores can be converted to traditional standard scores. This conversion involves using a W difference score and the standard deviation for an individual test or a composite to create a Z score. Then, the Z score can be converted to the standard score scale by multiplying Z by the standard deviation of the scale (15) and adding the mean of the scale (100).

Nicholas F. Benson, Ashley Donohue, and Emily Ward

Further Readings

Elliott, C. D. (2007). Differential Ability Scales 2nd edition administration and scoring manual. San Antonio, TX: Harcourt Assessment.

Jaffe, L. E. (2009). Development, interpretation, and application of the W score and the relative proficiency index (Woodcock-Johnson III Assessment Service Bulletin No. 11). Rolling Meadows, IL: Riverside.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). Woodcock-Johnson IV Tests of Achievement. Rolling Meadows, IL: Riverside.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV Tests of Cognitive Abilities. Rolling Meadows, IL: Riverside.

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance? In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement: What every psychologist and educator should know (pp. 105–127). Mahwah, NJ: Erlbaum.

Woodcock, R. W., & Dahl, M. N. (1971). A common scale for the measurement of person ability and item difficulty (AGS Paper No. 10). Circle Pines, MN: American Guidance Service.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23–48. doi:10.1177/001316446902900102

Angela Broaddus Angela Broaddus Broaddus, Angela

Stephen Keith Sagarin Stephen Keith Sagarin Sagarin, Stephen Keith

Waldorf Schools

Waldorf schools

1802

1806

Waldorf Schools

Waldorf schools, also called Steiner schools, are characterized by approaches to children and their education deriving from the writings and teachings of Rudolf Steiner (1861–1925). Waldorf education aims to develop a person with practical skills for work, wisdom of heart for sensitivity to others, and clarity of thought for discerning purpose in the world. This entry first discusses Steiner’s ideas and the history of Waldorf schools. It then discusses the human developmental stages conceptualized by Steiner, the education methods used in Waldorf schools, and research on Waldorf schools.

A prolific scholar, Steiner published more than 5,000 lectures and books that together provide a basis for understanding his claims regarding meditatively acquired knowledge of higher worlds and a method for attaining this knowledge. Together, this body of knowledge and method are known as *anthroposophy* and may be seen as his response to inadequacies he found in an earlier engagement with theosophy. Roughly 10% of Steiner’s work directly addressed education by describing an approach to teaching and curriculum that responds to human development and the spiritual essence of individual children.

Steiner described a 3-fold human being, consisting of body, soul, and spirit, and prescribed education as a means to assist the healthful unification of the physical (body) with the spiritual (consciousness) by engaging the soul (capacities of thinking, feeling, and will). The 3-fold human being engages with the world by thinking, feeling, and willing, metaphorically by head, heart, and hand. That is, human beings know with their heads, feel with their hearts, and experience with their hands.

This 3-fold perception of human nature requires individuals to acknowledge their individuality and simultaneously respond to and attend to their purpose in the world. Individual inner freedom may only be realized through free and independent thought, which may be developed by education that appeals to children’s individuality and prioritizes the healthy development of head, heart, and hand. Meaning and purpose, inner freedom, and ethical individuality, for Steiner, are spiritual capacities of consciousness that can flourish later in life if the soul is nurtured appropriately in childhood and youth.

History of Waldorf Schools

In lectures following World War I, Steiner sought to promote a social consciousness that could work toward peace and justice. Emil Molt (1876–1936), manager of the Waldorf-Astoria cigarette factory in Stuttgart, Germany, and an anthroposophist, asked Steiner what he could do to further this work. Steiner’s answer was to open a school for the children of the factory workers based on educational principles that Steiner had been espousing for more than a decade.

The Independent Waldorf School (*Freie Waldorfschule*) opened in September 1919 following a summer of preparation and a 2-week training course for the first teachers. The lectures and workshops that constituted this course have been published in three volumes (*Foundations of Human Experience*; *Practical Advice for Teachers*; and *Discussions with Teachers*) and form the core of the principles that guide Waldorf or Steiner education. These were supplemented during the rest of Steiner’s life by almost 300 additional lectures on education, delivered in Stuttgart, in other cities in Germany, and in Switzerland, Norway, and England.

The first school was coeducational and independent, both of which were unusual in Germany at the time. The school, which still exists, opened with more than 200 students in Grades 1–8, and, within a few years, had grown to close to 1,000 students and added Grades 9–13 (German high schools include a 13th year that focuses on preparation for university entrance). By the turn of the 21st century, there were more than 1,000 Waldorf or Steiner schools worldwide.

The Rudolf Steiner School in New York City, founded in 1928 by a group of anthroposophists, was the first Waldorf school in the United States. Since then, approximately 200 Waldorf schools have been founded in the United States. Roughly 150 of these are independent, 50 are charter schools, and a handful are district-run public schools that use Waldorf methods. The number of independent schools is relatively static, and the number of charter schools has increased steadily since the early 1990s. Collectively, these schools enroll roughly 50,000 students.

Human Developmental Stages

The driving force behind Steiner's educational methods is his theory of human development, in which physical developmental milestones are connected to cognitive stages. The first three stages are initiated by birth, change of teeth (6–7 years), and puberty (13–14 years). During each stage, teachers in Waldorf schools use methods and curricula that address children's developmental needs and capacities.

From birth to change of teeth, roughly, children are imitative beings, guided primarily by their will and responsive to the moods and wills of their caregivers. Parents and teachers of these children should model desirable behaviors and attitudes because these examples are the most powerful instruction during this stage. These children do not easily make sense of verbal instructions nor are they ready for conceptual explanations of the world. In contrast, they prefer imagination and wonder, which should be kept alive.

From change of teeth to puberty, children are devoted to loving authority, guided by their observations of the world, their feelings, and their imagination. They require aesthetic experiences to understand and learn best through pictures and symbols that appeal to their feelings. Thus, these children are not easily swayed or led by logical explanations of why things are or work the way they do. Early in this stage, at around 9 years, children generally acknowledge their inner self or "I" in a more conscious way, which permits them to see themselves as increasingly separate from the sensory world. This change prompts more objectivity and allows for the gradual introduction of comparative, abstract, and linear thinking. During the latter part of this stage, around 11–12 years, children begin to judge and to better understand cause and effect relationships and simple physical phenomena.

From puberty into adulthood, children develop their ability to think for themselves. They can think abstractly, creatively, and synthetically, and begin to apply independent judgment. During this stage, they are ready to acquire formal conceptual knowledge and benefit from a diverse and integrated curriculum. The cognitive activity occurring during this stage builds on the will-based and feelings-based activities of previous stages, thereby offering these children means for continually evaluating and making sense of their inner being and its relation to the world in which they live.

These stages are cumulative and integrative; capacities developed in one stage transform into new capacities in those that follow. For instance, young children imitate unconsciously; older children mimic deliberately; and adolescents may develop empathy as a form of inner, imaginative imitation.

Educational Methods

Steiner described an educational approach in which the nature and needs of a developing child are the bases for both teaching methods and curriculum. The goal of education, then, is to develop a child’s personal and social consciousness for reflecting on what lives in each human being and also what looks toward the world in which we live to fulfill our nature and purpose in relation to that world. Steiner sees the creativity and insight necessary for such fulfillment as fundamental human capacities that may be strengthened through education.

Teaching is regarded as an art and prioritized as a “doing” practice, in contrast to the practice of facilitating learning. Teachers narrate, tell, illustrate, show, inspire, and in a very real sense they provide the curriculum. Few or no textbooks are used until late middle or high school grades. Instead, teachers present carefully crafted lessons during which they provide material and engage students through stories. In practice, students document the curriculum and their learning by recalling, reflecting on, and recording their lessons in notebooks filled with reflective essays and illustrations. Assessment is typically formative in nature, whereby teachers aim to characterize students’ work rather than criticize. Students learn to consider the value of their own work, and the notion of comparisons among students is rather irrelevant because of the focus on the desirable individuality of each human being.

Recalling that the nature of children between change of teeth and puberty is feeling centered, Waldorf teachers optimize experience-based learning opportunities that appeal to all senses. In addition, constructive activities are often selected in advance of receptive activities, particularly in earlier grades because constructive experiences permit young children to naturally engage using their will and feelings instead of their cognition, which is required of receptive activities. In other words, children should express their learning physically with their bodies; through art, which calls on their feelings; and through writing before they are challenged to learn passively or through reading. Accordingly, children taught with Waldorf methods typically learn to write before they learn to read, which is in contrast to other learning sequences.

Writing Then Reading

Steiner recommended that elementary children should first experience writing by way of artistic expression, which inherently relates to the whole child by appealing to his or her feelings and worldly sensitivities. Thus, children in Waldorf schools learn to read through a gradual and multisensory approach that culminates with their early reading experiences occurring as they reflect on their own writings. This approach is seen to be more consistent with children’s nature and prevents the bewilderment that ensues when children try to learn standard characters and numerals, which are abstract and disconnected from life, without a feelings-focused, multisensory basis.

Mathematics

Students should develop understanding of mathematical truths in terms of their meaning in the world. This goal is in contrast to experiences whereby mathematics is perceived as a set of abstractions with little relevance to reality. For example, when students learn what a line is, it is important for them to understand it as an idea that cannot be directly observed due to its very nature. A line, in fact, has no width and thus cannot be seen, yet we represent a line with lead, chalk, or ink. Furthermore, a line is infinite; any representation of a line we create only inaccurately depicts its limitless nature. The invisible reality of a line is but one example of the many mathematical truths that should be learned as ideas that are relevant to and applied in the world.

Science

Interest in and wonder about the world should guide learning experiences. Children should respond to the world in ways that are consistent with their

Children should respond to the world in ways that are consistent with their developmental level, that is, with their feelings during elementary and middle grades. After children have been introduced to a phenomenon to be investigated scientifically, a teacher aims to pose problems or questions that prompt the children to explore and discover the rules or truths that undergird the phenomenon under consideration, fostering questioning, investigation, and rediscovery that follow actual scientific practices. Students, then, do not learn science as a body of knowledge but as a creative, investigative practice. It is particularly important to refrain from supplying young children with abstract answers to questions about the world because they generally develop the capacity to reconcile abstract concepts with their worldly experiences beginning in puberty.

Handwork and Crafts

Handwork is an important element of Waldorf education because it enhances the development of a child’s will, emotions, and intellect. Children learn a variety of crafts not only to practice artistry but also to develop appreciation for how their hands allow them to engage with the world and transform its materials through their work. Thus, handwork and crafts maintain a connection with the world while also stimulating imagination and thought. As children mature as artists, they apply a balance of thinking, feeling, and will when they practice handwork and crafts.

Research on Waldorf Schools

In 2007, an investigation focused on whether Waldorf methods are effective for reforming the education of traditionally underserved populations by analyzing data collected from public schools practicing Waldorf methods and a survey conducted among American Waldorf school graduates from 1943 through 2005. This study examined how Waldorf methods involve individual attention and challenging, meaningful instruction that is relevant to students’ lives, or what researchers on school reform have referred to as “rigor, relevance, and relationships.” The findings indicated that Waldorf methods contributed positively to public school reform efforts by exposing richer qualities of rigor, relevance, and relationships than other reform efforts.

Angela Broaddus and Stephen Keith Sagarin

See also Constructivist Approach; Critical Thinking; Curriculum; Formative Assessment; Zone of Proximal Development

Further Readings

Franceschelli, A. (1998). Mathematics in the classroom: Mine shaft and skylight. Chestnut Ridge, NY: Mercury Press.

Mitchell, D., & Livingston, P. (1999). Will-developed intelligence. Waldorf Publications.

Oberman, I. (2007, April). Learning from Rudolf Steiner: The relevance of Waldorf education for urban public school reform. Online Submission. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. Retrieved from https://eric.ed.gov/?id=ED498362

Steiner, R. (1967). Discussions with teachers. Helen Fox (Trans). London, UK: Rudolf Steiner Press.

Steiner, R. (1921/1996). Education for adolescents. Hudson, NY: Anthroposophic Press.

Steiner, R. (1996). Foundations of human experience. R. Lathe & N. P. Whittaker (Trans). CW 293. Hudson, NY: Anthroposophic Press.

Steiner, R. (2000). Practical advice to teachers. 14 Lectures, Stuttgart, 1919. CW 294. Great Barrington, MA: SteinerBooks.

Steinmann, L. (2005). *For life and for now: On the question of commitment and learning*. Rundbrief, 22. Dornach, Switzerland. Retrieved from http://www.waldorflibrary.org/articles/728-for-life-and-for-now-on-the-question-of-commitment-and-learning

Von Hydebrand, C. (1932). On the real nature of will in the child. Anthroposophical Quarterly, 7.

Christopher R. Niileksela Christopher R. Niileksela Niileksela, Christopher R.

Wechsler Intelligence Scales Wechsler intelligence scales

1806

1809

Wechsler Intelligence Scales

The Wechsler Intelligence Scales are individually administered, standardized assessments of cognitive functioning. There are three Wechsler intelligence batteries, each designed for different age ranges: the *Wechsler Preschool and Primary Scales of Intelligence* (WPPSI), the *Wechsler Intelligence Scales for Children* (WISC), and the *Wechsler Adult Intelligence Scales* (WAIS). In 1939, the first test published by David Wechsler was intended for adults and was called the Wechsler–Bellevue. In 1955, the Wechsler–Bellevue became the first version of the WAIS. The first version of the WISC was developed in 1949, and the first version of the WPPSI was developed in 1967. With each revision of the Wechsler Scales, the subtests have been updated, new tests have been introduced, other tests have been removed, and changes have been made to the organization of the composite scores based on theory and research. Many of the subtests from the original Wechsler–Bellevue, though now different in item content, continue to be used on the current versions of the tests.

The Wechsler Scales provide scores for general intelligence and several other cognitive abilities, including verbal comprehension, fluid reasoning, visual spatial, working memory, and processing speed. There are also a number of ancillary composites included on the fourth edition of the WPPSI (WPPSI-IV), published in 2012, and the fifth edition of the WISC (WISC-V), published in 2014, that may enhance clinical interpretation. These ancillary composites include quantitative reasoning, vocabulary acquisition, and a nonverbal index that include tests that reduce the influence of expressive language. The Wechsler Scales provide age-based standard scores for composites (mean of 100, standard deviation of 15), age-based scaled scores for subtests (mean of 10, standard deviation of 3), and percentile ranks for all subtests and composites.

Each of the Wechsler batteries are used for a different age-group, although there

Each of the Wechsler batteries are used for a different age group, although there is some slight overlap among the different tests (e.g., a 16-year-old may be administered the WISC-V or the WAIS-IV). The WPPSI-IV may be used for individuals between the ages of 2 years 6 months and 7 years 7 months, the WISC-V can be used with individuals between the ages of 6 years 0 months and 16 years 11 months, and the WAIS-IV may be used with individuals between the ages of 16 and 90 years. The normative sample for each test was large and representative of the U.S. population, including 1,700 children for the WPPSI-IV, 2,200 children and adolescents for the WISC-V, and 2,200 adolescents and adults for the WAIS-IV. The entry reviews the various composites and subtests as well as the scales' psychometric properties.

Composites and Subtests

Across the most recent versions (as of 2017) of the Wechsler Scales, there are 27 different subtests, although not all subtests are on all batteries. The WPPSI-IV includes 15 subtests, the WISC-V includes 21 subtests, and the WAIS-IV includes 15 subtests. There is a substantial amount of overlap among the subtests included on each battery, but some subtests are specific to certain batteries. For example, the working memory tests on the WPPSI-IV are more visual than auditory, which may be more developmentally appropriate for young children who may not have developed language skills necessary for working memory tests that include words or numbers. In addition, the WISC-V includes a number of new tests that may appear on future versions of the WPPSI and WAIS. In this section, all the subtests included on these three batteries are briefly described, and the cognitive ability and test batteries that the subtest is on are specified.

- *Vocabulary* (verbal comprehension; WPPSI-IV, WISC-V, WAIS-IV)—provide the definitions of words presented orally.
- *Similarities* (verbal comprehension; WPPSI-IV, WISC-V, WAIS-IV)—describe how two words are conceptually similar (e.g., How are a *tiger* and *lion* alike?).
- *Information* (verbal comprehension; WPPSI-IV, WISC-V, WAIS-IV)—answer general knowledge questions related to a number of subjects, such as history, science, or the humanities.
- *Comprehension* (verbal comprehension; WPPSI-IV, WISC-V, WAIS-IV)—explain social rules or conventions (e.g., Why do we say *excuse me*?).
- *Picture vocabulary* (verbal comprehension; WPPSI-IV)—provide the name of a picture that is presented.

- *Receptive vocabulary* (verbal comprehension; WPPSI-IV)—choose a picture from a set of four that best represents a word presented orally.
- *Matrix reasoning* (fluid/perceptual reasoning; WPPSI-IV, WISC-V, WAIS-IV)—determine which choice in a set of distractors logically completes a matrix or sequence of pictures.
- *Figure weights* (fluid/perceptual reasoning; WISC-V, WAIS-IV)—determine which set of shapes would correctly balance a scale.
- *Picture concepts* (fluid/perceptual reasoning; WPPSI-IV, WISC-V)—choose pictures from two or three arrays that conceptually go together.
- *Block design* (visual spatial/perceptual reasoning; WPPSI-IV, WISC-V, WAIS-IV)—recreate a picture using a set of colored blocks.
- *Visual puzzles* (visual spatial/perceptual reasoning; WISC-V, WAIS-IV)—mentally identify which three shapes in a set of distractors would make a target puzzle.
- *Picture completion* (perceptual reasoning; WAIS-IV)—identify what part is missing in a picture.
- *Object assembly* (visual spatial; WPPSI-IV)—put together puzzles of common objects.
- *Digit span* (working memory; WISC-V, WAIS-IV)—repeat increasingly longer strings of numbers in the order they were presented (digit span forward), in reverse order (digit span backward), or in numerical order (digit span sequencing).
- *Letter–number sequencing* (working memory; WISC-V, WAIS-IV)—repeat a sequence of numbers and letters that were presented in numerical and alphabetical orders.
- *Arithmetic* (working memory also measures verbal comprehension and fluid reasoning; WISC-V, WAIS-IV)—mentally complete simple arithmetic problems presented orally.
- *Picture span* (working memory; WISC-V)—remember the order a set of pictures were presented.
- *Picture memory* (working memory; WPPSI-IV)—identify the pictures that were presented on one page in an array of distractors on another page.
- *Zoo location* (working memory; WPPSI-IV)—put pictures of animals in the place where they were seen on a previous page.
- *Coding* (processing speed; WISC-V, WAIS-IV, WPPSI-IV [called animal coding on the WPPSI-IV])—copy symbols paired with numbers, shapes, or animals as quickly as possible.
- *Symbol search* (processing speed; WISC-V, WAIS-IV, WPPSI-IV [called bug search on the WPPSI-IV])—identify which of a set of shapes (or

pictures of bugs on the WPPSI-IV) is the same as a target shape.

- *Cancellation* (processing speed; WISC-V, WAIS-IV, WPPSI-IV)—cross out a certain type of stimulus (e.g., only specific shapes or animals) among a set of distractors.
- *Naming speed literacy* (storage and retrieval; WISC-V)—identify the name, size, and color of pictures as quickly as possible and name letters and numbers as quickly as possible.
- *Naming speed quantity* (storage and retrieval: WISC-V)—identify the number of blocks in small arrays (i.e., between one and five blocks) as quickly as possible.
- *Immediate symbol translation* (storage and retrieval; WISC-V)—learn the words associated with a set of novel symbols and translate increasingly longer sentences that use the symbols.
- *Delayed symbol translation* (storage and retrieval; WISC-V)—recall the names of the symbols presented during immediate symbol translation after a 20-to 30-minute delay.
- *Recognition symbol translation* (storage and retrieval; WISC-V)—recognize the name of the symbols from immediate symbol translation in a multiple choice format after a 20-to 30-minute delay.

The subtests from the Wechsler Scales may be combined into a number of different composite scores. The tests that are included in these composite scores differ slightly from one battery to the next. Details for which tests are included on each composite can be found in the manual for each battery.

- *Full-Scale IQ* (WPPSI-IV, WISC-V, WAIS-IV) is an estimate of general intelligence and includes one or more subtests from each of the different cognitive abilities measured on the Wechsler Scales, including verbal comprehension, fluid reasoning, visual spatial, working memory, and processing speed.
- *General ability index* (WPPSI-IV, WISC-V, WAIS-IV) is meant to estimate general intelligence without the influence of working memory or processing speed. This may be a clinically useful composite if the scores for one or more working memory or processing speed subtests are substantially lower than the other tests. The general ability index includes tests that have been highly associated with general intelligence and higher order thinking and include subtests from the verbal comprehension, fluid reasoning, and visual–spatial composites.
- *Verbal comprehension* (WPPSI-IV, WISC-V, WAIS-IV) represents a

person’s breadth and

- depth of vocabulary knowledge, ability to reason using language, and general knowledge.
- *Fluid reasoning* (WPPSI-IV, WISC-V) represents a person’s ability to use inductive and deductive reasoning to solve novel problems.
- *Visual–spatial* (WPPSI-IV, WISC-V) represents a person’s ability to solve visually based puzzles and mentally visualize stimuli.
- *Perceptual reasoning* (WAIS-IV) is a mixture of tests from fluid reasoning and visual–spatial composites. Previous versions of the WPPSI and WISC included a perceptual reasoning composite, but the WPPSI-IV and WISC-V have separated the tests included on the previous perceptual reasoning composite into two composites: fluid reasoning and visual–spatial.
- *Working memory* (WPPSI-IV, WISC-V, WAIS-IV) represents a person’s attentional and mental control, where the person must hold information in short-term memory stores and then use it to answer a question or problem. Tests of working memory are either auditory (e.g., remembering letters or numbers) or visual (e.g., remembering a sequence of pictures).
- *Processing speed* (WPPSI-IV, WISC-V, WAIS-IV) represents a person’s ability to complete simple tasks quickly and efficiently.

There are a number of ancillary indexes that are included with the WPPSI-IV and WISC-V that are intended to enhance clinical interpretation.

- *Auditory working memory* (WISC-V). This composite is a measure of working memory subtests that only use auditory stimuli.
- *Quantitative reasoning* (WISC-V). This composite measures reasoning skills that require numbers and the logic of numbers to solve problems.
- *Nonverbal index* (WPPSI-IV, WISC-V). This composite includes tests that do not rely on expressive language for responses.
- *Cognitive proficiency* (WPPSI-IV, WISC-V). This composite includes tests of working memory and processing speed, which is meant to represent cognitive processing efficiency (e.g., ease of remembering and quickly processing information).
- *Naming speed* (WISC-V). This composite represents a person’s automaticity with
- identifying letters and numbers, small quantities, and pictures.
- *Symbol translation* (WISC-V). A measure of visual–auditory associative memory, this composite includes all of the symbol translation tests.
- *Storage and retrieval* (WISC-V). This composite is meant to be a broad

measure of long-term retrieval (e.g., efficiency of retrieval, associative memory) and includes all tests included in the naming speed and symbol translation composite.

- *Vocabulary acquisition* (WPPSI-IV). This composite measures general vocabulary development for young children.

Psychometric Properties

Overall, the psychometric properties of the Wechsler Scales are very strong. For reliability, the average internal consistency estimates are excellent for subtests on the WPPSI-IV (.71−.95 for subtests and .85−.96 for composites), the WISC-V (.81−.94 for subtests and .88−.96 for composites), and the WAIS-IV (.78−.94 for subtests and .90−.98 for composites). Validity evidence presented for the Wechsler Scales consists of correlations among all the subtests and composites, confirmatory factor analyses, correlations with previous versions of the Wechsler Scales, and other major cognitive, academic, executive functioning, and behavioral measures. In addition, validity studies were conducted with clinical populations to examine how these groups score on the various subtests and composites. For example, some of the clinical populations included individuals who were intellectually gifted and individuals who had intellectual disabilities, learning disabilities, attention-deficit/hyperactivity disorder, dementia, traumatic brain injury, and autism spectrum disorders. Overall, the Wechsler Scales are excellent tests that have a long history as some of the premier cognitive tests in psychology and can be useful for measuring a number of cognitive abilities in both research and practice settings.

Christopher R. Niileksela

See also Intelligence Quotient; Standardized Scores; Standardized Tests; Stanford-Binet Intelligence Scales

Further Readings

Wechsler, D. (1939). The measurement of adult intelligence. Baltimore, MD: Williams & Wilkins.

Wechsler D. (1949). Wechsler Intelligence Scale for Children. New York, NY: Psychological Corporation.

Wechsler, D. (1955). Wechsler Adult Intelligence Scale. New York, NY: Psychological Corporation.

Wechsler D. (1967). Wechsler Preschool and Primary Scale of Intelligence. New York, NY: Psychological Corporation.

Wechsler, D. (2008). Wechsler Adult Intelligence Scale–Fourth Edition: Technical and interpretive manual. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2012). Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition: Technical manual and interpretive manual. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2014). Wechsler Intelligence Scale for Children–Fifth Edition. Technical and interpretive manual. Bloomington, MN: Pearson.

Ji An Ji An An, Ji

Laura M. Stapleton Laura M. Stapleton Stapleton, Laura M.

Weighting

Weighting

1809

1812

Weighting

The term *weighting* refers to the process of incorporating sampling weights into analyses when using educational data collected through a complex sampling procedure, one in which the sample cannot be assumed to be from a simple random sampling process. The use of sampling weights is often a challenge for data analysts who are new to sampling theory. This entry introduces different types of sampling weights and the importance of incorporating sampling weights into analyses when using data collected through a multistage sampling procedure.

Introduction

For data selected following a complex sampling design, the sample itself may not reflect the population characteristics in known ways. Data analysts therefore must incorporate sampling weights to more appropriately mimic the population and thus to obtain unbiased estimates of population parameters. For example, suppose a population of 20 individuals consisted of 10 females and 10 males. In the sample, two males and four females were randomly selected using stratified sampling, reflecting probabilities of selection of π_{male} = 2/10= .2 and π_{female} = 4/10 = .4. If the population average height is calculated based on this sample, which contains proportionally more females than males, the parameter estimate will be likely biased toward the population mean for males.

The sampling weight is generally defined as the inverse of the selection probability and can be considered the number of units each observation in the sample represents. In our example, the sampling weight for males w_{male} is 1/.2 = 5, representing 5 males in the population. Similarly, the sampling weight w_{female} is 1/.4 = 2.5 for females and represents 2.5 females in the population.

In addition to this simple sampling weight based on stratification, other types of weights are available with large-scale data in education. The definitions and use of a variety of weights, as well as weight adjustment approaches, are presented in the following sections.

Types of Weights

A number of different weight variables may be available in any large-scale data set. A researcher needs to fully understand the differences across these weights to be able to select the appropriate ones for analyses of interest.

Base Weights

In large-scale educational survey, observations are often selected following a multistage framework, within which weights are provided for each stage. A typical three-stage survey, for example, may include the selection of geographic areas, schools in those selected areas, and then students within the selected schools. Two-stage surveys are also common in education, including the selection of schools, followed by the selection of students. The units being selected at each stage (e.g., the county, the school, or the student) are based on some predefined probability of selection.

Consider a two-stage sample as an example. It is rare that schools are selected using a simple random design; instead, the school has a given selection of probability and thus has a sampling weight that may differ from other schools. Researchers frequently use a sampling procedure called probability proportional to size sampling. In particular, the selection probabilities for schools vary depending on, and are proportional to, the schools’ size. In other words, large schools have relatively larger selection probabilities, π_j, as compared to smaller schools, and vice versa. Therefore, larger schools tend to have smaller sampling weights (taking the inverse $1/\pi_j = w_j$) than smaller schools. The school sampling weight for a given selected school is the number of schools it represents in the sampling frame. Once schools are selected, in the second stage, students are selected perhaps using categories of a given characteristic (e.g., age and race). These categories are referred to as *strata*. Because the student selection probability, $\pi_{i|j}$, is the selection probability of the student within the selected school, it reflects the conditional selection probability. Thus, the second-stage within-school conditional sampling weight is $1/\pi_{i|j} = w_{i|j}$.

The overall, or unconditional, weight of the two-stage sample is then the product of the two base weights (i.e., the inverse of the selection probability of the school and the inverse of the conditional selection probability of the classroom or student), that is, $w_{ij} = w_j \times w_{i|j}$.

Weight Adjustments

The responses from a complete selected sample are not always available. In practice, it is very common that some schools or students did not agree to participate in the study or did not respond. The initial sampling weights then need to be adjusted to accommodate the nonparticipation or nonresponse to remain representative of the population. Suppose that in the aforementioned simple example, two of the four selected females (out of 10 in the population) did not respond and the researcher should make a nonresponse adjustment to the initial sampling weight. In this example, the nonresponse adjustment is 4/2 = 2 (the selected number of females divided by the number of responses); this way, the weights of the existing two responses are inflated to represent the two missing responses. These nonresponse adjustment factors may exist for any selection level (e.g., county, school, and student). Note that nonresponse adjustment factors might be much more complicated and not based on just the explicit gender information as shown in the example; they might have been derived from the sampling frame information or auxiliary information about the selected units. The simple adjustment approach introduced here is referred to as a weighting class adjustment method by Roderick J. A. Little and Donald B. Rubin. Other methods of nonresponse adjustments, such as the use of propensity score methods and response propensity models, are available.

Replicate Weights

In addition to obtain unbiased parameter estimates, estimating accurate sampling variance is also a challenge in analyzing data obtained from complex sampling designs. Replicate weights is a set of weights that are obtained from replication methods (e.g., jackknife repeated replication; balanced repeated replication) and are used to compute more precise sampling variances of survey estimates. The basic idea of replicate weights is that for each observation or primary sampling unit (PSU; e.g., schools), the weight has either been set to 0 or inflated to account for the weights of their neighbor observation (from the same PSU) or PSU (from the same stratum) that were set to 0. In the case of jackknife repeated replication, this process is repeated so each observation or PSU has the chance to be left out and others reweighted; therefore, multiple sets of replicates are available. With that said, the replicate weights are not to be used within a single analysis; they should be used only in a sampling variance estimation procedure, that is, obtaining a set of model parameter estimates by running the analysis for each set of replicate weights and then computing the empirical estimate of the sampling variance of the parameter estimates. In data sets where the replicate weights are not provided, they can be created by software as needed if the stratum and PSU indicators are available.

Other Weights

The most frequently encountered weights on large-scale educational data sets, including base, conditional, adjusted, and replicate weights, have been discussed. Other weights, such as panel weights and linkage weights, may also play important roles in complex survey data.

In longitudinal surveys, panel weights are usually provided to address the fact that some participants may not complete all waves of the study (e.g., dropout). With panel weights, the user could appropriately weight the sample to accommodate the fact that only some of the original participants in the study are included in the data.

In cases in which more than one informant is associated with an observation, linkage weights are needed. Consider the “student–teacher” linkage weight, for example, when using the teacher report data to estimate the characteristics of a student who received reports from more than one teacher, the students’ overall weight would need to be partitioned appropriately so that they do not count more than they really represent.

Unweighted Approaches to Analyses

Unweighted Approaches to Analyses

Given the introduction to the types of weights that are often seen in large-scale survey data, one may wonder if sampling weights, when available, should always be included in data analysis. The answer is no because the inclusion of weights might be disadvantageous if the corresponding selection mechanism is noninformative (e.g., with regard to the first example provided in this entry, if the average height of men and women does not differ in the population). While not improving the accuracy of the parameter estimates (the parameter estimates from an unweighted analysis are already accurate in this case), including weights in the analysis may unnecessarily inflate the standard errors and thus will lead to a loss of precision and lower power for statistical inference.

In addition, suppose that the sampling mechanism is informative: it is possible to take a model-based approach in place of the sampling weight, that is, to use an analytic model that contains the information that defined the sampling mechanism. For example, with the simple scenario used in this entry, the mean height could be calculated separately for men and women instead of performing the calculation for the population across gender or in a regression model whereby an indicator variable for gender is used. Note that a fully model-based approach would quickly become complicated, as the sampling mechanism gets more complex. In the example, if other information such as age is part of the selection mechanism in addition to gender, or even a second-level sampling is involved (e.g., individuals, males and females, were selected from European and Asian countries), obtaining the average height would require a more complex model. Therefore, to better understand the implications of the inclusion of specific weights in the analysis, researchers are advised to run the analysis with and without weights. Tom A. B. Snijders and Roel J. Bosker have introduced a variety of empirical methods to determine whether weights need to be incorporated in the analysis. It is important to know that the decision should depend on sampling error, as the difference in inference given weighted and unweighted analysis may differ by sample.

Final Thoughts

Large-scale educational survey data include a vast amount of useful information for researchers to develop models and to explain the development of students' abilities and skills. However, the complex sampling nature of the data could be a source of confusion, and it requires appropriate methods to analyze the data to obtain unbiased parameters estimates and their sampling variances.

Although undertaking weighted analyses may lead to inflated sampling variances when the corresponding sampling mechanism is not informative, for most cases where the mechanism is informative, incorporating sampling weights is a straightforward and desirable approach. In addition to sampling weights, model-based approaches can be another option to obtain unbiased estimates, however, and may easily become complex along with the sampling design—all design information must be available and appropriately modeled. When such conditions for the model-based methods are unlikely met, use of weighted analyses is the simplest approach.

Ji An and Laura M. Stapleton

See also Cluster Sampling; Estimation Bias; Stratified Random Sampling; Survey Methods

Further Readings

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). Applied survey data analysis. Boca Raton, FL: Chapman Hall/CRC Press.

Kalton, G. (1983). Models in the practice of survey sampling. International Statistical Review, 51, 175–188.

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, NJ: Wiley.

Lohr, S. (2010). Sampling: Design and analysis (2nd ed.). Pacific Grove, CA: Duxbury Press.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 117–153). London, UK: Chapman Hall/CRC Press.

Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). London, UK: Sage.

Stapleton, L. M. (2013). Incorporating sampling weights into single-and multilevel analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 363–388). London, UK: Chapman Hall/CRC Press.

Thu Suong Nguyen Thu Suong Nguyen Nguyen, Thu Suong

Brendan Maxcy Brendan Maxcy Maxcy, Brendan

Wicked Problems

Wicked problems

1812

1814

Wicked Problems

Commonly attributed to Horst Rittel and Melvin Webber, the concept of wicked problems was first used to describe those problems facing social planners (e.g., city planners who must anticipate and plan for a variety of outcomes and contingencies) that are particularly complex in contrast to easier to define and better behaved problems with which scientists dealt. Terming the latter *tame problems*, Rittel and Webber argued such problems are clearly defined, solvable, and have few or no consequences for social systems. In contrast, urban planners contend with *wicked problems*, which are inherently ill-defined, largely intractable and for which implementation of provisional solutions has significant consequences for social systems. Although the concept originated in the planning literature in the 1960s, there is a resurgent interest in use and development of the idea in a wide range of fields and in the social sciences, particularly in education. This entry focuses on the original formulation of wicked problems in the context of planning, but the reader likely can easily translate the characteristics of wicked problems into educational research and program evaluation examples.

Rittel and Webber delineate 10 distinguishing characteristics of wicked problems: (1) no definitive formulation; (2) no criteria for determining when the problem has been solved, in other words, a planner can always attempt to do better; (3) solutions are not true or false, rather, they are better or worse; (4) no immediate or ultimate tests for solutions to a wicked problem; (5) each implemented solution is consequential; (6) an exhaustive set of solutions cannot be enumerated because wicked problems are ill-defined; (7) each wicked problem is unique; (8) each wicked problem is generated from and generates new problems; (9) numerous explanations can be given for wicked problems and each explanation determines how the problem is resolved; and (10) the planner has no right to be wrong.

These characteristics reflect Rittel and Webber’s efforts to provide an alternative to the dominant linear stepwise approaches to social problem solving. The concept of wicked problems arose out of concerns about the ability of professionals to address complex social problems in the planning and policy arenas. Informed by a context of social upheaval and racial tension of the 1960s, Rittel and Webber were not simply concerned with questions of how planners (and researchers) define, and thus, solve problems. Centering on questions of equity, they emphasized the need for planners to inquire into divergent values and interests brought to bear on social problems, as well as the potential consequences of various solutions for different vested communities. This approach diverged from prior work in the science of design and design thinking, which tended to view problems as definable and solvable through decontextualized study and rational planning.

The predominant mode of planning in the 20th century suggested that complex social problems needed to be reduced to manageable components where rational planning might be better applied. However, according to Rittel, most problems of design are wicked problems in that there is a high degree of complexity, uncertainty, and divergence of values involved in the social policy realms. The idea of wicked problems has since been taken up by those attempting to understand problems in various disciplines including architecture, public management, education, and health policy, to name a few. In part, the indeterminacy of wicked problems has to do with the infinite range of subjects to which it may be applied. The distinguishing characteristics presented earlier were meant to help planners in the identification and resolution of wicked problems. However, this may have limited use of the concept to solely a descriptive tool. Some suggest that wicked problems may be more productively utilized when used as a tool toward inquiry. In the field of education, those interested in management and administration have taken up the wicked problem as one way to advance greater understanding and further inquiry where problems appear or are intractable.

Almost 3 decades after Rittel and Webber’s formulation, Keith Grint elaborated on the model with particular focus on leadership and management studies. His on the model with particular focus on leadership and management studies. His contributions offer important insights into connections between problem definitions and leadership, management, and command. First, Grint expands the typology to include critical problems or crises, which differ qualitatively from both tame and wicked problems. Second, Grint characterizes the responses to problem types: tame problems are amenable to planning; crises demand command; wicked problems are addressed through leadership qua questioning to reveal, reconsider, and resolve. Third, Grint draws on theories of social construction of problems to argue that problems are not inherently tame, wicked, or critical. Rather, problem definition is a social construction dependent on the choices of those individuals and groups determining the definition. In this way, Grint highlights the role of agency in defining and responding to problems.

Extending the idea that problems are socially constructed, Linda Sue Warner along with Grint later highlight the *culturally* constructed nature of problems as well. In doing so, they are better able to address and center Rittel and Webber's earlier concern with equity. These elaborations are taken up in more recent research focused on the education of indigenous and refugee communities.

In his reconsideration of wicked problems, Richard Coyne draws from poststructural theories and argues that Rittel and Webber stop short of a more radical and perhaps more productive view of wicked problems. Although Rittel and Webber argue that in contrast to tame scientific problems most social policy problems are wicked, Coyne advances the notion that all problems may be framed as wicked and that it is the nature of our encounters with problems that constrains or highlights the "wickedness" of each. This seems consistent with Warner and Grint's notions of socially and culturally constructed problems. That is, while agency exists in problem construction and thus response, both are socially and culturally conditioned.

Thu Suong Nguyen and Brendan Maxcy

See also Experimental Designs; Postpositivism; Qualitative Research Methods

Further Readings

Buchanan, R. (1992). Wicked problems in design thinking. Design Issues, 8(2), 5–21.

Coyne, R. (2005). Wicked problems revisited. Design Studies, 26(1), 5–17.

Grint, K. (2005). Problems, problems, problems: The social construction of leadership. Human Relations, 58(11), 1467–1494. doi:10.1177/0018726705061314

Head, B.W. (2008). Wicked problems in public policy. Public Policy, 3(2), 101–118. doi:https://doi.org/10.1177/0095399713481601

Kameniar, B., Imtoual, A., & Bradley, D. (2010). "Mullin' the yarndi" and other wicked problems at a multiracial early childhood education site in regional Australia. Educational Policy, 24(1), 9–27. doi:https://doi.org/10.1177/0895904809354321

Nguyễn, T. S. T., Scribner, S. M. P., & Crow, G. (2012). Tangled narratives and wicked problems: A complex case of positioning and politics in a diverse school community. Journal of Cases in Educational Leadership, 15(4), 49–64. doi:https://doi.org/10.1177/1555458912470657

Rittel, H., & Webber, M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2), 155–169. doi:10.1007/BF01405730

Warner, L. S., & Grint, K. (2006). American Indian ways of leading and knowing. Leadership, 2(2), 225–244. doi:https://doi.org/10.1177/1742715006062936

Jill S. M. Coleman Jill S. M. Coleman Coleman, Jill S. M.

Wilcoxon Signed Ranks Test Wilcoxon signed ranks test

1814

1817

Wilcoxon Signed Ranks Test

Several statistical tests exist for comparing data groups depending on the group number, sample size, data collection method, and population distribution. The Wilcoxon signed ranks test is the nonparametric equivalent of the matched-pairs difference *t* test for two dependent samples. The test examines whether or not the differences between the ranks of paired data come from a population with a median equal to zero. This entry reviews the assumptions and requirements for this text, explains the test statistic, and illustrates its application in education with an example.

Data Assumptions and Requirements

The Wilcoxon signed ranks test has four major assumptions: (1) dependent observations, (2) random sampling, (3) continuous dependent variable, and (4) ordinal-level measurement. In the case of dependent observations, the data are paired or related according to a repeated measurement on an equivalent scale and are drawn from a single sample. Each sample member (e.g., location, individual) has two values or pairs that are obtained for two different time periods or two variables. The paired data are also presumed to be randomly and independently drawn from the population and are continuous. Multiple observations from the same data pair are not permissible. Lastly, a major assumption is that the data are at the ordinal-level measurement scale or at the interval or ratio scale downgraded to the ordinal scale; nominal-level data are not appropriate. Ordinal-level data enable comparison of the differences in the absolute value of the ranks between each data pair.

As a nonparametric test statistic, the Wilcoxon signed ranks test does not assume a Gaussian (normal) distribution. Nonparametric statistics are appropriate for data sets that may have outliers, notable skewness, or multimode distributions, characteristics common in small samples. Nonparametric tests are also useful in situations where a traditional mean and standard deviation cannot be calculated, as in the case of ordinal-level data, or is not a suitable measure of central tendency. The Wilcoxon signed ranks test does not require data to conform to a particular distribution. The differences in the paired data values though are assumed to be distributed symmetrically about their median (or the middle of the distribution at the second quartile).

Test Statistic

The Wilcoxon signed rank test (Z_W) has a similar structure to the parametric matched-pairs test but utilizes ranked values (ordinal-level data) instead of interval-ratio level variables such as the sample mean and standard deviation. For sample sizes greater than 10, the test statistic may be approximated as a normal distribution Z score using the following formula:

$$Z_{\mathrm{W}} = \frac{T - \overline{X}_T}{s_T},$$

where *T* is the rank differences for the matched pairs of one sign (positive or negative), X_T and s_T are the mean and the standard deviation based on the number of matched pairs (*n*), respectively. Smaller samples should calculate *p* values using a *t* distribution based on sample size.

The Wilcoxon signed ranks test examines the differences between the first and second observations of each matched pair. The smallest paired difference is given a rank of one while the largest paired difference is given a rank equal to the total number of paired observations. In the cases when the matched-pair differences are tied for a rank position, the average rank value is assigned to the associated pairs. Matched pairs with no differences (0 values) are not ranked and the sample size is reduced as a result.

After the data pairs are ranked based on their differences, the ranks are summed into one of two variables based on their sign: (1) T_p for positive differences whereby the first observation is greater than the second observation or (2) T_n for negative differences whereby the second observation is greater than the first observation. In cases where the sample has very little difference between the paired data values, the positive (T_p) and negative (T_n) rank sums will be similar. However, large differences between the paired observations will yield large differences in the ranks and the disparity between T_p and T_n will also be large. Only one of these rank summations will be used for the Wilcoxon *T* component of the test statistic.

The selection of either the positive or negative rank sum for the Wilcoxon *T* is contingent on whether or not the alternative hypothesis being tested is directional dependent. If no direction between the paired observations is hypothesized, then the *smaller* of T_p and T_n is selected and a two-tailed test is applied. A nondirectional hypothesis indicates the researcher does not suspect a specific negative or positive difference to occur but only a difference between the observations. If the alternative hypothesis is directional and constructed as a one-tailed test, then either the positive or negative differences between the matched pairs is anticipated to dominate. For directional cases, the *smaller* of the paired differences suspected will be used as the Wilcoxon *T*; hence, T_p is used if the negative differences are expected to be the largest and T_n is used if the positive differences are expected to be the largest.

Computation of the mean X_T and standard deviation s_T equivalents in the Wilcoxon signed ranks test is more straightforward than determining *T*. The mean and standard deviation are computed as

$$\overline{X}_T = \frac{n(n+1)}{4},$$

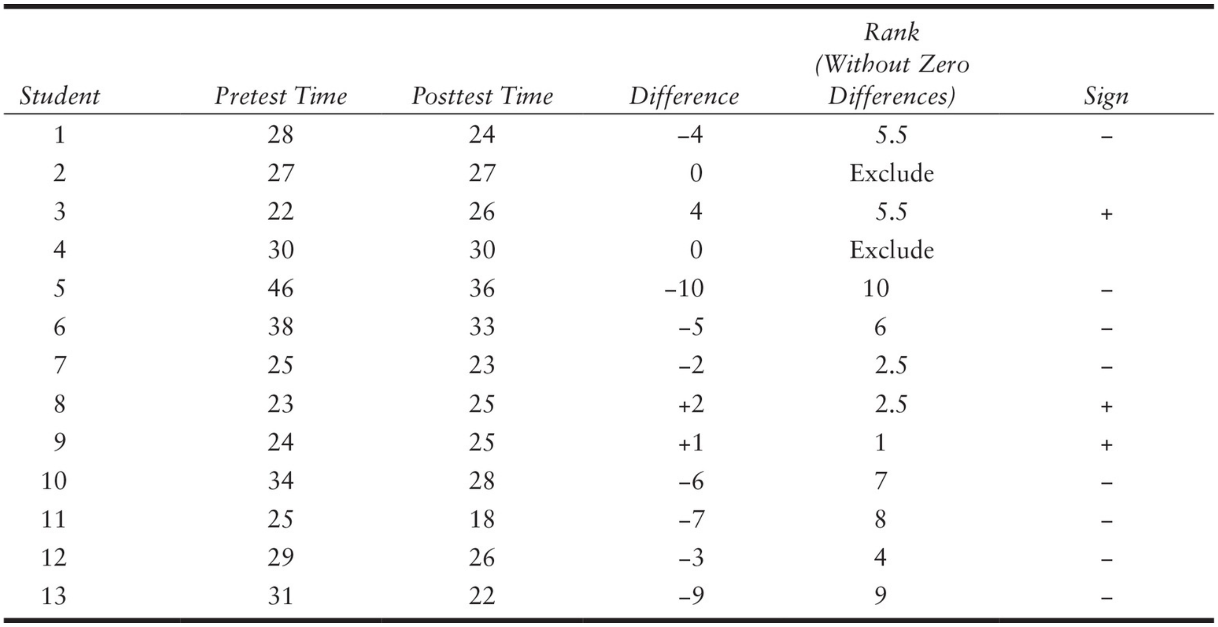
$$s_T = \sqrt{\frac{n(n+1)(2n+1)}{24}},$$

where *n* is the number of matched pairs included in the analysis. Recall, the value of *n* may not be equal to the total number of matched pairs, as the total subtracts the cases where the difference between the pairs is zero.

Example: Test Score Times

–

An example of test-taking times illustrates the calculation of the Wilcoxon signed ranks test using the steps in a *p* value hypothesis testing procedure. Suppose a cartography professor tests the time in minutes taken by introductory geography students to extract information from a set of maps prior to any instruction. At the end of the course, the same students take an identical test and the cartographer records their new times (see Table 1). Have the students learned how to use maps more effectively? In other words, have test-taking times for these students significantly decreased between the start and the end of the course?



Student	*Pretest Time*	*Posttest Time*	*Difference*	*Rank (Without Zero Differences)*	*Sign*
1	28	24	–4	5.5	–
2	27	27	0	Exclude	
3	22	26	4	5.5	+
4	30	30	0	Exclude	
5	46	36	–10	10	–
6	38	33	–5	6	–
7	25	23	–2	2.5	–
8	23	25	+2	2.5	+
9	24	25	+1	1	+
10	34	28	–6	7	–
11	25	18	–7	8	–
12	29	26	–3	4	–
13	31	22	–9	9	–

Because the cartography professor is monitoring test performance for the same set of students before and after instruction, the appropriate type of statistical test is matched pairs for dependent samples. However, the very small sample size, large outliers, and a likely nonnormally distributed population make the nonparametric Wilcoxon signed ranks test a logical alternative. The cartography professor also strongly suspects test times have decreased between the pre-and posttest; hence, the test is directional or one tailed.

After determining the appropriate test statistic and data requirements, the Wilcoxon signed ranks test for this example follows these steps:

Step 1: Statement of the null (H_0) and alternative (or research; H_A) hypothesis:

- H_0: δ = 0 (no differences in the test times between the pre-and posttests)
- H_A: δ < 0 (test times have decreased between the pre-and posttests)

Step 2: Calculate the Wilcoxon *T* (see Table 1):

- Find the difference between the matched pairs by subtracting the pretest time from the related posttest time.
- Based on the *absolute value* of the differences, rank each observation from the lowest difference (Rank 1) to highest for all *nonzero* values. Observations tied for specific rank position are allocated an average rank to each associated pair. In this data set, two sets of two students had the same absolute differences in their test-taking times and received an average of their ranks. Excluded from the analysis are two students (Students 2 and 4), who showed no differences in their test-taking times.
- Sum the ranks for the positive differences (T_p) and the negative differences (T_n):

$$T_p = 1 + 2.5 + 5.5 = 9,$$

$$T_n = 2.5 + 4 + 5.5 + 7 + 8 + 9 + 10 = 46.$$

- Select either T_p or T_n for the Wilcoxon *T* based on the smaller number of hypothesized differences. Because the negative differences are hypothesized to be the greatest (i.e., test-taking times have decreased between the pre-and posttest), then T_p should be selected as the Wilcoxon *T*.

Step 3: Compute the Wilcoxon signed ranks test:

- Determine the mean X_T based on the number of matched pairs included in the analysis (*n* = 11):

$$\overline{X}_T = \frac{n(n+1)}{4} = \frac{11(11+1)}{4} = 33.$$

- Find the standard deviation s_T based on the number of matched pairs included in the analysis (*n* = 11):

$$s_T = \sqrt{\frac{n(n+1)(2n+1)}{24} = \frac{11(11+1)(2(11)+1)}{24}} = 11.5.$$

- Calculate the test statistic:

$$Z_W = \frac{T - \overline{X}_T}{s_T} = \frac{9 - 33}{11.5} = -2.08.$$

- Determine the *p* value based on Z_W and a normal distribution table because *n* > 10. The corresponding *p* value for a Z_W is approximately equal to 0.018 (one tailed).

Step 4: Make a decision regarding the null and alternative hypothesis. Using a *p* value of 0.018 and a negative Z_W score, test times between the pre-and posttests significantly decreased based on a 0.05 significance level (α). In other words, only a 5% chance or less (in this case a 1.18% chance) exists of making a Type I error, a critical error whereby by the null hypothesis is falsely rejected. The results support the alternative hypothesis with a relatively large degree of confidence; therefore, the cartography professor should reject the null hypothesis.

Summary

The Wilcoxon signed ranks test is the nonparametric equivalent of the matched pairs test for dependent samples. As a nonparametric statistic, the test is often less powerful than the parametric version and makes achieving high confidence and statistical significance (e.g., *p* values less than .05) more difficult. Consequently, data sets are often transformed (e.g., logarithmic) to a normal distribution if possible in order to use parametric test procedures. On the other hand, the test utilizes ordinal-level data to test the differences between ranked values and is more robust against outliers and nonnormally distributed data sets.

Jill S. M. Coleman

See also Hypothesis Testing; Inferential Statistics; Ordinal-Level Measurement; *p* Value; *t* Tests; Type I Error; Z Scores

Further Readings

Corder, G. W., & Foreman, D. I. (2014). Nonparametric statistics: A step-by-step approach (2nd ed.). Hoboken, NJ: Wiley.

Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference (5th ed.). Boca Raton, FL: CRC Press.

Gravetter, F. J., & Wallnau, L. B. (2016). Statistics for the behavioral Sciences (10th ed.). Boston, MA: Cengage Learning.

Sprent, P., & Smeeton, N. C. (2007). Applied nonparametric statistical methods (4th ed.). Boca Raton, FL: CRC Press.

Bruce E. Blaine Bruce E. Blaine Blaine, Bruce E.

Winsorizing Winsorizing

1817

1818

Winsorizing

Winsorizing is a procedure that moderates the influence of outliers on the mean and variance and thereby creates more robust estimators of location and variability. The procedure is named for biostatistician Charles P. Winsor. Parametric inferential procedures that rely on the mean and variance (e.g., *t* test) become more robust when they incorporate Winsorized estimators. Winsorizing is an important tool for educational and social science researchers for two reasons. First, significance tests based on the mean and variance are very common procedures for significance testing in the social sciences. Second, surveys of the educational and psychological literature show that nonnormally distributed data are the rule rather than the exception, and even modest departures from normality disproportionately affect the mean and variance compared with other more robust estimators of location (e.g., median) and variability (e.g., median absolute deviation).

Winsorizing reassigns values to a percentage of cases in both tails of a distribution using the next highest (in the lower tail) and lowest (in the upper tail) value; the resultant variable is said to have been Winsorized. The Winsorized mean is the mean of the Winsorized values, and the Winsorized variance is the average squared deviation of the Winsorized values from the Winsorized mean. Consider, as an example, a variable with the following values:

2, 2, 3, 3, 3, 4, 5, 8, 15, 25.

The mean and variance of this variable are 7.0 and 55.6, respectively. Owing to the presence of at least one unusually large score, these estimators perform poorly in characterizing location and variability. A 20% Winsorizing procedure identifies the lowest (e.g., 2, 2) and highest (e.g., 15, 25) 20% of the cases. These cases reassigned the value of the adjacent upper (e.g., 3) or lower (e.g., 8) case, resulting in this Winsorized variable:

producing this Winsorized variable:

$$3, 3, 3, 3, 3, 4, 5, 8, 8, 8.$$

The 20% Winsorized mean and variance are 4.8 and 5.3, respectively. From intuitive and statistical perspectives, these statistics are better estimators of the location and variability of the original variable.

For moderating the influence of outliers on the mean and variance and creating better estimators from skewed or heavy tailed data, Winsorizing is an alternative to trimming. The most common levels of Winsorizing are 10% and 20%, although this decision is up to the researcher; greater levels of Winsorizing create more robust estimators. Relative to trimmed means and variances, Winsorizing does not cast aside data and create associated issues (e.g., reduced degrees of freedom). Functions for computing Winsorized means and variances, other Winsorized estimators (e.g., Winsorized product–moment correlation coefficient), and inferential procedures incorporating Winsorized estimators are available in some statistical software packages. Notably, there are numerous packages in R for doing robust descriptive and inferential data analysis, including Winsorized procedures.

Bruce E. Blaine

See also Missing Data Analysis; Robust Statistics

Further Readings

Erceg-Hurn, D., & Mirosevich, V. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. American Psychologist, 63(7), 591–601. doi:10.1037/0003-066X.63.7.591.

Wilcox, R., & Keselman, H. (2003). Modern robust data analysis methods: Measures of central tendency. Psychological Methods, 8, 254–274.

Christopher R. Niileksela Christopher R. Niileksela Niileksela, Christopher R.

Woodcock-Johnson Tests of Achievement Woodcock-johnson tests of achievement

1818

1821

Woodcock-Johnson Tests of Achievement

The Woodcock-Johnson Tests of Achievement (WJ ACH) is an individually administered, standardized test of academic achievement in the Woodcock-Johnson family of academic achievement tests. The 2014 revision, known as WJ IV ACH, was revised at the same time as the Woodcock-Johnson IV Tests of Cognitive Abilities and the Woodcock-Johnson IV Tests of Oral Language. The WJ achievement tests are designed to measure a variety of academic skills, primarily in reading, writing, and mathematics. These tests are designed to be clinically useful instruments that can help identify strengths and weaknesses in academic skills and may be especially useful for the identification of learning disabilities and determining educational needs for a variety of school difficulties. The academic skills measured by the WJ IV ACH are consistent with current understanding of academic skills in reading, writing, and math. For instance, reading tests have been designed to measure basic word decoding skills, reading rate and fluency, and reading comprehension. Understanding these different areas can help practitioners or researchers understand where individuals may be having specific difficulties with reading, writing, or mathematics and ultimately determine where instruction may need to occur based on these scores.

The WJ IV ACH includes 20 tests that are designed to measure a number of broad and specific academic skills. Test scores can be combined to create composite scores that represent broader representations of different academic skills. The following descriptions of the tests indicate the name of the test, the specific academic skills measured by the test (in parentheses), and what the examinee is asked to do on each test:

- *Letter–word identification* (*basic reading skills*)*:* Orally read increasingly difficult individual words.

- *Applied problems* (*math problem solving*): Solve mathematics word problems that require the application of mathematics knowledge through the identification of appropriate steps, operations, and accurate calculation.
- *Spelling* (*basic writing skills*): Spell single words provided orally by the examiner.
- *Passage comprehension* (*reading comprehension*): Read a short sentence or paragraph that has one word missing, and provide the missing word.
- *Calculation* (*math calculation skills*): Solve individual math calculation problems, ranging from simple addition problems to geometry and calculus.
- *Writing samples* (*written expression*): Write increasingly complex sentences based on an oral prompt.
- *Word attack (basic reading skills*): Orally read words that use typical English phonemes and morphemes, but are not real words (e.g., retabbered).
- *Oral Reading* (*reading fluency*): Orally read increasingly difficult sentences.
- *Sentence reading fluency* (*reading fluency*): Quickly read short sentences silently and determine if the sentence is true or false in a specified time limit.
- *Math facts fluency* (*math calculation skills*): Quickly solve as many single-digit addition, subtraction, and multiplication problems in a specified time limit.
- *Sentence writing fluency* (*written expression*): Write syntactically accurate sentences that use three target words in a specified time limit.
- *Reading recall* (*reading comprehension*): Silently read short passages then tell everything that is remembered without referring back to the passage.
- *Number matrices* (*math problem solving*): Provide a number that is missing in a matrix of other numbers.
- *Editing* (*basic writing skills*): Identify spelling or punctuation errors in short sentences or passages.
- *Word reading fluency* (*reading fluency*): Quickly read individual words silently and identify which two words conceptually go together in sets of four words in a specified time limit.
- *Spelling of sounds* (*phoneme–grapheme knowledge*): Spell nonsense words heard orally that use typical English phonemes but are not real words (e.g., plinger).
- *Reading vocabulary* (*reading comprehension*): Read individual words orally, and then provide synonyms and antonyms for those words.

- *Science* (*academic knowledge*): Answer questions about general science knowledge.
- *Social studies* (*academic knowledge*): Answer questions about general social studies knowledge.
- *Humanities* (*academic knowledge*): Answer questions about general humanities knowledge.

The tests from the WJ IV ACH can be combined into a number of composite scores that represent a range of broad and specific academic skills.

- *Broad achievement*: The Broad Achievement composite is designed to provide an overall picture of a person's academic skills. This composite includes tests from reading, writing, and mathematics. The tests included in broad achievement are letter–word identification, passage comprehension, sentence reading fluency, calculation, applied problems, math facts fluency, spelling, writing samples, and sentence writing fluency.
- *Brief achievement*: The brief achievement composite is designed to provide a short measure of general academic skills. This includes three tests from reading, mathematics, and
- writing, which include letter–word identification, applied problems, and spelling, respectively.
- *Reading*: This composite is designed to be a general measure of reading ability and includes tests of basic reading skills and reading comprehension. The tests included in reading are letter–word identification and passage comprehension.
- *Broad reading*: A composite meant to represent a person's general reading skills across all domains of reading, including decoding, reading comprehension, and reading fluency. The tests included in broad reading are letter–word identification, passage comprehension, and sentence reading fluency.
- *Basic reading skills*: This composite represents a person's ability to decode individual words using typical English conventions. The tests included in basic reading skills are letter–word identification and word attack.
- *Reading comprehension*: This composite that represents a person's ability to understand what he or she read. The tests included in this composite are passage comprehension and reading recall. An extended reading comprehension composite comprising three tests is available if reading vocabulary is also administered.
- *Reading fluency*: This composite represents a person's ability to read

quickly and accurately and includes tests of reading rate and reading accuracy. The tests included on reading fluency are sentence reading fluency and oral reading.

- *Reading rate*: This composite represents a person's silent reading fluency and includes two speeded reading tests, sentence reading fluency and word reading fluency.
- *Mathematics*: This composite is designed to measure a person's general math skills in basic calculation skills and the application of math skills to the real-world problems. The tests included in mathematics are calculation and applied problems.
- *Broad mathematics*: A composite meant to represent a person's general mathematics skills, which includes tests of calculation, word problems, and fact retrieval fluency. This composite includes calculation, applied problems, and math facts fluency.
- *Math calculation skills*: A composite that represents a person's ability with mathematics
- calculation, including a test of calculation and fact retrieval fluency. This composite includes calculation and math facts fluency.
- *Math problem solving*: This composite represents a person's ability to apply mathematics concepts to the real-world problems and to use quantitative reasoning skills. Math problem solving includes both applied problems and number matrices.
- *Written language*. The written language composite is a general measure of writing that includes tests related to spelling and writing syntactically and grammatically correct sentences. The tests included in written language are spelling and writing samples.
- *Broad written language*. This composite is meant to represent a person's general writing skills, which includes tests of written expression, spelling ability, and writing fluency. The tests included on broad written language are writing samples, spelling, and sentence writing fluency.
- *Written expression*: This composite is meant to represent a person's ability to write meaningful and syntactically correct sentences in English without penalizing for basic writing skills (e.g., spelling or punctuation). This composite includes writing samples and sentence writing fluency.
- *Basic writing skills*: This composite represents a person's basic skills related to writing, which include spelling and ability to identify errors in writing. Basic writing skills includes spelling and editing.
- *Academic knowledge*: The WJ IV ACH includes three tests designed to provide an estimate of general academic knowledge. This composite can be

used to compare a person’s general academic knowledge to his or her specific academic skills. The tests included in this composite are science, social studies, and humanities.

- *Academic skills*: This composite is designed to represent a person’s mastery of basic academic skills, including decoding of words, basic math calculation, and spelling. This composite can help understand if a person has some of the basic skills necessary for engaging in academic tasks. The tests included on academic skills are letter–word identification, calculation, and spelling.
- *Academic fluency*: This composite represents the ease with which an individual engages in academic skills across the different domains of reading, writing, and math. The tests included in this composite are sentence reading fluency, math facts fluency, and sentence writing f.
- *Academic applications*: This composite represents a person’s ability to apply academic skills appropriately in everyday situations and includes tests of reading comprehension, math problem solving, and written expression. This composite includes passage comprehension, applied problems, and writing samples.
- *Phoneme–grapheme knowledge*: This composite represents a person’s knowledge of English phonemes, including the decoding of phonemes in reading (word attack) and their ability to encode words that are heard using typical English conventions (spelling of sounds).

Reliability evidence for the WJ IV ACH tests and composite scores is strong. For individual tests, median reliability estimates across ages range from .84 to .96. For composite scores, median reliability estimates range from .92 to .99. Similar to the Woodcock-Johnson IV Tests of Cognitive Abilities, there is an extensive amount of validity evidence for the WJ IV ACH that is based on a range of statistical analyses, including exploratory and confirmatory factor analysis, cluster analysis, multidimensional scaling, correlations with other well-established tests of academic achievement, and scores for tests administered to groups of individuals with exceptionalities (e.g., intellectual disability, giftedness, and learning disabilities) to examine whether scores would be in the expected ranges. Overall, the WJ IV ACH is one of the premier tests that can be used to help identify academic strengths and needs in comparison to the general population.

Christopher R. Niileksela

See also Standardized Tests; Woodcock-Johnson Tests of Cognitive Ability; Woodcock-Johnson Tests of Oral Language

Further Readings

Schrank, F. A., Mather, N., & McGrew, K. S. (2014a). Woodcock-Johnson IV tests of oral language. Rolling Meadows, IL: Riverside publishing.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014b). Woodcock-Johnson IV tests of achievement. Rolling Meadows, IL: Riverside publishing.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV Tests of Cognitive Abilities. Rolling Meadows, IL: Riverside publishing.

Woodcock, R. W., & Johnson, M. B. (1977). Woodcock-Johnson psycho-educational battery. Hingham, MA: Teaching Resources.

Woodcock, R. W., & Johnson, M. B., (1989). Woodcock-Johnson psycho-educational battery—revised. Chicago, IL: Riverside publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III. Itasca, IL: Riverside publishing.

Christopher R. Niileksela Christopher R. Niileksela Niileksela, Christopher R.

Woodcock-Johnson Tests of Cognitive Abilities Woodcock-johnson tests of cognitive abilities

1821

1824

Woodcock-Johnson Tests of Cognitive Abilities

The Woodcock-Johnson Tests of Cognitive Abilities (WJ COG) is an individually administered, standardized test of cognitive functioning and is part of the Woodcock-Johnson, a flexible battery of tests designed to measure intellectual functioning in a number of relevant cognitive areas. The 2014 version, the WJ IV COG, was developed simultaneously as the Woodcock-Johnson IV Tests of Achievement and the Woodcock Johnson IV Tests of Oral Language. The WJ IV COG measures general intelligence and several other specific cognitive abilities, including comprehension knowledge, fluid reasoning, short-term working memory, long-term retrieval, visual processing, auditory processing, and cognitive processing speed. The measurement of these cognitive abilities may be useful for researchers interested in the measurement of these abilities, and they are also clinically useful for identifying a number of exceptionalities, including learning disabilities, intellectual disabilities, and giftedness.

The WJ IV COG is based on previous versions of the WJ cognitive tests, including the Woodcock-Johnson Psychoeducational Battery (1977), the Woodcock-Johnson Revised (1989), and the Woodcock-Johnson Third Edition (2007). The WJ tests are closely aligned with Cattell–Horn–Carroll (CHC) theory, an influential taxonomy of cognitive abilities based on extensive factor analytic research. CHC theory was used as a guiding structure for the WJ IV COG, but current neuropsychological theory and research were also used to help update the tests within the WJ IV COG to make them consistent with current understanding of cognitive processing.

The WJ IV COG can be used for individuals between the ages of 2 and 90+. The normative sample was representative of the U.S. population based on the 2010 Census and included 7,416 individuals. The test provides a number of different scores that can be used for interpretation, including age-based standard scores (with a mean of 100 and standard deviation of 15), percentile ranks, age equivalents, grade equivalents, the relative performance index (a criterion-referenced score), and *W* scores (an interval-level ability score used to calculate all other scores). Standard scores and percentile ranks are the most commonly used scores for test interpretation and may be used to compare test scores to each other to help determine an individual's cognitive strengths and weaknesses.

The WJ IV COG includes 18 individual tests that measure a variety of cognitive abilities. Scores from these tests are combined to create composite scores that represent broad and narrow cognitive abilities based on CHC theory. The following descriptions of the tests indicate the name of the test, the broad CHC ability measured by the test (in parentheses), and what the examinee is asked to do on each test:

- *Oral vocabulary (comprehension knowledge)*. Provide synonyms and antonyms for words.
- *Number series (fluid reasoning)*. Examine a sequence of numbers and identify a missing number that would logically fit into the sequence.
- *Verbal attention (short-term working memory)*. Listen to a series of numbers and animals and answer a question about one or more components of the series (e.g., name the first animal).
- *Letter–pattern matching (cognitive-processing speed)*. Quickly identify which two letters or sets of letters are the same in a row of distractors.
- *Phonological processing (auditory processing)*. This test comprises three short subtests. On word access, the examinee is asked to think of a word that includes a specific phoneme at the beginning, middle, or end of a word. On word retrieval, the examinee is asked to quickly come up with words that start with a specific phoneme. On substitution, the examinee hears a word and is asked to change one of the sounds in the word to create a new word.
- *Story recall (long-term retrieval)*. Listen to short stories and retell everything that is remembered from the stories.
- *Visualization (visualization)*. Visualization includes two short subtests. On spatial relations, the examinee is asked to determine which of a set of pieces go together to make a puzzle. On block rotation, the examinee is asked to

determine which shapes in a set of choices is the same as a target shape, but the shapes that the examinee must choose have been rotated in space.

- *General information (comprehension knowledge)*. Identify where common objects are found or how they are used.
- *Concept formation (fluid reasoning)*. View two groups of objects and identify a rule that separates those objects into different categories.
- *Numbers reversed (short-term working memory)*. Listen to a series of numbers and repeat the numbers in backward order.
- *Number–pattern matching (cognitive processing speed)*. Quickly identify which two numbers or sets of numbers are the same in a row of distractors.
- *Nonword repetition (auditory processing)*. Repeat a word that consists of common English phonemes but is not a real word (e.g., craffing).
- *Visual–auditory learning (long-term retrieval)*. Learn a set of symbols that represent words and read increasingly longer and complex sentences that use the symbols.
- *Picture recognition (visual processing)*. Briefly view a set of pictures and identify which pictures were previously seen among a set of distractors.
- *Analysis synthesis (fluid reasoning)*. Use a set of specified rules to solve picture puzzles.
- *Object–number sequencing (short-term working memory)*. Listen to a series of numbers and
- animals and repeat them in the order they were presented but in their respective categories (i.e., name the animals first and then the numbers).
- *Pair cancellation (cognitive processing speed)*. Quickly identify a specific sequence of pictures within a set of distractors.
- *Memory for words (short-term working memory)*. Listen to a series of words and repeat them in the same order they were presented.

The tests can be combined into different composite scores, which represent different cognitive abilities based on CHC theory. The composites on the WJ IV COG are measures of general intelligence (*g*), broad cognitive abilities, or more specific cognitive processing composites.

- *General intellectual ability* is a composite score that includes seven tests representing the seven broad cognitive ability domains from CHC theory that are measured on the WJ IV COG. The tests included in the g*eneral intellectual ability* are oral vocabulary, number series, verbal attention, phonological processing, story recall, visualization, and letter–pattern matching.

- *Gf–Gc composite* includes four tests, two that measure fluid reasoning and two that measure comprehension knowledge. These abilities are related to higher order thinking and can be used as an alternative measure of general intelligence. The tests included in the *Gf–Gc* composite are oral vocabulary, general information, number series, and concept formation.
- *Brief intellectual ability* is designed to be a brief measure of general intelligence and includes three tests, one each from comprehension knowledge, fluid reasoning, and short-term working memory. The tests included in the b*rief intellectual ability* are oral vocabulary, number series, and verbal attention.
- *Comprehension knowledge* represents an individual's breadth and depth of language development and general knowledge. The composite for comprehension knowledge includes two tests: oral vocabulary and general information. It is possible to obtain an extended comprehension knowledge composite based on three tests if picture vocabulary from the Woodcock-Johnson IV Tests of Oral Language is also administered.
- *Fluid reasoning* represents a person's ability to use inductive, deductive, or quantitative reasoning to solve novel problems. The WJ IV COG includes a fluid reasoning composite that includes two tests: number series and concept formation. An extended fluid reasoning composite that includes three tests is available if analysis synthesis is also administered.
- *Short-term working memory* represents individuals' ability to attend to and manipulate information in their short-term memory stores to achieve a goal. The WJ IV COG short-term working memory includes two tests: verbal attention and numbers reversed. It also includes an extended short-term working memory composite if object–number sequencing is also administered.
- *Visual processing* is the ability to use visual imagery and mental visualization to solve puzzles or problems. The WJ IV COG visual processing composite includes two tests: visualization and picture recognition.
- *Long-term retrieval* is the ability to store and retrieve information in long-term memory. Long-term retrieval refers to information that is held over a longer period of time (from seconds to years) than would be possible to keep in short-term memory. The WJ IV COG long-term retrieval composite includes two tests: story recall and visual–auditory learning.
- *Cognitive processing speed* represents the ability to complete cognitive tasks quickly and accurately. The WJ IV COG composite for cognitive processing speed includes two tests: letter–pattern matching and pair

cancellation.

- *Auditory processing* represents the ability to hear, understand, and manipulate sounds. The tests on the WJ IV COG primarily focus on the hearing and understanding of speech sounds. The auditory processing composite includes two tests: phonological processing and nonword repetition.
- *Cognitive efficiency* includes tests of processing speed and short-term working memory that represents the ease with which a person can actively take in and utilize information. This composite includes letter–pattern matching and numbers reversed. An extended cognitive efficiency composite is available if
- number–pattern matching and numbers reversed are also administered.
- *Perceptual speed* is the ability to quickly recognize similarities and differences in stimuli and includes two tests: letter–pattern matching and number–pattern matching.
- *Quantitative reasoning* is a measure of quantitative and sequential reasoning, a specific ability within fluid reasoning. This composite includes number series and analysis synthesis.
- *Number facility* is designed to measure the ease with which a person is able to utilize numbers in cognitive processing. This composite includes numbers reversed and number–pattern matching.
- *Scholastic aptitude* has six composites that are designed to provide combinations of cognitive ability tests that best predict a variety of academic skills.

Overall, the psychometric properties of the WJ IV COG are strong and the tests demonstrate excellent validity evidence. For individual tests, median reliability estimates across ages range from .74 to .97. For composite scores, median reliability estimates range from .86 to .97. All reliability estimates are strong for tests and composites, except for picture recognition, which has a median reliability of .74.

There is an extensive amount of validity evidence for the WJ IV COG that is based on a range of statistical analyses, including exploratory and confirmatory factor analysis, cluster analysis, multidimensional scaling, correlations with other well-established tests of intelligence, and scores for tests administered to groups of individuals with exceptionalities (e.g., intellectual disabilities, giftedness, learning disabilities) to examine whether scores would be in the expected ranges. Overall, the WJ IV COG represents one of the most comprehensive and psychometrically strong tests of cognitive abilities, making it comprehensive and psychometrically strong tests of cognitive abilities, making it useful for a variety of clinical and research purposes.

Christopher R. Niileksela

See also Standardized Tests; Woodcock-Johnson Tests of Achievement; Woodcock-Johnson Tests of Oral Language

Further Readings

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor analytic studies. New York, NY: Cambridge University Press.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical manual. Woodcock-Johnson IV. Rolling Meadows, IL: Riverside.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (3rd ed., pp. 99–144). New York, NY: Guilford.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014a). Woodcock-Johnson IV Tests of Achievement. Rolling Meadows, IL: Riverside.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014b). Woodcock-Johnson IV Tests of Oral Language. Rolling Meadows, IL: Riverside.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV Tests of Cognitive Abilities. Rolling Meadows, IL: Riverside.

Schrank, F. A., McGrew, K. S., & Mather, N. (2015). The WJ IV Gf-Gc Composite and its use in the identification of specific learning disabilities (Woodcock-Johnson IV Assessment Service Bulletin No. 3). Rolling Meadows, IL: Riverside.

Woodcock, R. W., & Johnson, M. B. (1977). Woodcock-Johnson psycho-educational battery. Hingham, MA: Teaching Resources.

Woodcock, R. W., & Johnson, M. B. (1989). Woodcock-Johnson psycho-educational Battery—Revised. Chicago, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). Woodcock-Johnson III. Itasca, IL: Riverside.

Christopher R. Niileksela Christopher R. Niileksela Niileksela, Christopher R.

Woodcock-Johnson Tests of Oral Language Woodcock-johnson tests of oral language

1824

1827

Woodcock-Johnson Tests of Oral Language

The Woodcock-Johnson Tests of Oral Language (WJ IV OL) is a new battery to the WJ family of tests (which previously only included batteries for cognitive assessment and academic assessment). However, all of the tests that are included on this individually administered, standardized test of oral language skills, often abbreviated as WJ IV OL, were a part of previous versions of the WJ cognitive or academic achievement batteries. With the 2014 revision of the WJ, several tests that had previously been a part of the Oral Language composite on the WJ III, along with others that were relevant for language assessment, were included on a separate battery of tests. The WJ IV OL was developed and normed at the same time as the Woodcock-Johnson IV Tests of Cognitive Abilities (WJ IV COG) and the Woodcock-Johnson IV Tests of Achievement (WJ IV ACH). All tests that are a part of the WJ IV are closely aligned with Cattell–Horn–Carroll theory, which provides a theoretical structure for categorizing the tests included on the WJ IV OL that has been based on extensive research.

The WJ IV OL includes a number of expressive and receptive language tests that are designed to measure a number of abilities related to language development, including vocabulary, listening comprehension, oral expression, phonological awareness, and speed of lexical access. Expressive language tests require the individual to verbally provide answers to items, whereas receptive language tests only require the individual to point to or provide minimal verbal responses that depend highly on the understanding of language. Contrasting skills on these two types of language may be clinically useful for identifying difficulties with the understanding or production of language. In addition, there are three corresponding English and Spanish language tests that can be used to help evaluate language skills for both languages.

The WJ IV OL can be used for individuals across the life span, from ages 2 to 90+ years. The normative sample included 7,416 people and was representative of the U.S. population based on the 2010 census. The WJ IV OL provides a number of different scores that can be used for interpretation that are consistent with the other tests in the WJ IV, including age-based and grade-based standard scores (with a mean of 100 and standard deviation of 15), percentile ranks, age equivalents, grade equivalents, the relative performance index (a criterion-referenced score), and *W* scores (an interval-level ability score used to calculate all other scores). In addition, the WJ IV OL provides a level of cognitive academic language proficiency, which refers to a person's ability to understand more complex and formal academic language, as opposed to basic interpersonal communicative skills, which is simpler, more conversational language.

The WJ IV ACH includes 12 tests that were developed to measure several different oral language skills. There are nine tests in English and three tests in Spanish. The three Spanish language tests are Spanish versions of three of the English language tests. The Spanish language tests on the WJ IV OL were developed as parallel forms to the English language tests; they are not simply translations of the English test forms. The use of parallel forms rather than translations allows for direct comparisons for the scores on the English and Spanish versions of the tests, which may be useful in identifying a dominant language for a student whose primary language is Spanish but who is currently learning English.

Similar to the WJ IV COG and WJ IV ACH, scores from individual tests can be combined to create composite scores that represent broader language constructs. The following test descriptions include the name of the test, the specific oral language skills measured by the test (in parentheses), and what the examinee is asked to do on each test:

- *Picture vocabulary* (*oral expression*): Identify the name of an object presented in a picture.
- *Oral comprehension* (*listening comprehension*): Listen to a sentence or short paragraph presented orally and provide a word that would finish the sentence.
- *Segmentation* (*phonetic coding*): Listen to a word and break the word down into individual syllables or sounds.
- *Rapid picture naming* (*speed of lexical access*): Quickly identify the names of pictures of common objects.

- *Sentence repetition* (*oral expression*): Listen to a sentence presented orally and repeat the sentence back exactly as it was presented.
- *Understanding directions* (*listening comprehension*): Listen to a series of oral directions and point to specific areas or pictures on a page as directed.
- *Sound blending* (*phonetic coding*): Listen to words presented one sound at a time and identify the whole word.
- *Retrieval fluency* (*speed of lexical access*): Name as many words in a specific category as quickly as possible (e.g., name as many games as you can in one minute).
- *Sound awareness* (*phonological processing*). Sound awareness includes two subtests that are designed to measures phonological awareness, which are rhyming (identifying or providing rhyming words) and deletion (removing a phoneme or syllable from a word), and is meant to be a screener for potential phonological issues. Sound awareness is not a part of any of the composites on the WJ IV OL.
- *Vocabulario sobre dibujos* (*Spanish language version of picture vocabulary*): Identify the names of pictures in Spanish.
- *Comprehensión oral* (*Spanish language version of oral comprehension*): Listen to a sentence presented in Spanish and provide a word that would complete the sentence.
- *Comprensión de indicaciones* (*Spanish version of understanding directions*): Listen to a series of directions presented orally in Spanish and act out the directions as specified (e.g., point to specific pictures on a page in a specified order).

The tests from the WJ IV OL can be combined into a number of composite scores that are designed to represent several broad and narrow oral language skills.

- *Broad oral language*: The broad oral language composite includes three tests and is designed to provide an overall estimate of a person's language skills. This composite includes two tests of listening comprehension (oral comprehension and understanding directions) and one test of oral expression (picture vocabulary).
- *Oral language*: This composite provides a brief estimate of oral language abilities and includes two tests: one of listening comprehension (oral comprehension) and one of oral expression (picture vocabulary).
- *Oral expression*: This composite measures a person's ability to express himself or herself using language and includes two tests: picture vocabulary

and sentence repetition.

- *Listening comprehension*: This composite measures a person’s ability to hear and understand language and includes two tests: oral comprehension and understanding directions.
- *Phonetic coding*: Phonetic coding is a composite that measures a person’s ability to take apart and put together individual speech sounds in words. This includes two tests: segmentation and sound blending.
- *Speed of lexical access*: This composite is designed to measure the efficiency with which a person can access linguistic information from long-term memory stores and includes rapid picture naming and retrieval fluency.
- *Lenguaje oral*: This composite is the Spanish language equivalent of the oral language composite and includes Vocabulario sobre diujos and Comprensión Oral.
- *Amplio lenguaje oral*: This composite is the Spanish language equivalent of the broad oral language composite and includes Vocabulario sobre dibujos and Comprensión Oral, and Comprensión de indicaciones.
- *Comprensión auditiva*: This composite is the Spanish language equivalent of the listening comprehension composite, and includes Comprensión Oral, and Comprensión de indicaciones.
- *Auditory memory span*: This composite is available if memory for sentences from the WJ IV OL and memory for words from the WJ IV COG are both administered and provide a measure of how much verbal information a person can hold in short-term memory and repeat it back exactly as it was presented.
- *Vocabulary*: This composite is available if picture vocabulary from the WJ IV OL and oral vocabulary from the WJ IV COG are both administered, and this provides a composite that measures vocabulary development and lexical knowledge.

It is important to note that many of the tests on the WJ IV OL are multidimensional and may not measure only language skills, although language is a significant part of performance on these tests. From a CHC perspective, picture vocabulary and oral comprehension are measures of comprehension knowledge (specifically, lexical knowledge and listening ability, respectively), understanding directions and memory for sentences are also measures of short-term working memory (working memory and memory span, respectively), retrieval Fluency and rapid picture naming both measure long-term retrieval (both measure speed of lexical access, rapid picture naming also measures naming facility), and segmenting, sound blending, and sound awareness measure auditory processing (specifically, phonetic coding). The technical manual provides information on what other cognitive abilities are measured on some of the tests on the WJ IV OL. An understanding of these other abilities is necessary to interpret the scores on the tests appropriately. For example, an individual with difficulties in short-term working memory may also have low scores on understanding directions and memory for sentences, but this may not be due to a language issue, rather it may be due to a short-term working memory issue.

Reliability evidence for the WJ IV OL tests and composite scores are strong and similar to those from the WJ IV COG and WJ IV ACH. For individual tests, median reliability estimates across ages range from .80 to .94. For composite scores, median reliability estimates range from .89 to .95. Similar to the WJ IV COG, there is an extensive amount of validity evidence for the WJ IV OL based on a wide variety of statistical analyses, including exploratory and confirmatory factor analysis, cluster analysis, multidimensional scaling, correlations with other well-established tests of oral language, and scores for tests administered to groups of individuals with exceptionalities (e.g., intellectual disability, giftedness, and learning disabilities) to examine whether scores would be in the expected ranges. Overall, the WJ IV OL, as an addition to the WJ family of cognitive and academic tests, can be used to help identify academic strengths and needs in comparison to the general population.

Christopher R. Niileksela

See also Standardized Tests; Woodcock-Johnson Tests of Achievement; Woodcock-Johnson Tests of Cognitive Ability

Further Readings

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor analytic studies. New York, NY: Cambridge University Press.

Mather, N., & Wendling, B. J. (2015). Essentials of WJ IV tests of achievement. Hoboken, NJ: Wiley.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical manual. Woodcock-Johnson IV. Rolling Meadows, IL: Riverside Publishing.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014a). Woodcock-Johnson IV tests of oral language. Rolling Meadows, IL: Riverside Publishing.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014b). Woodcock-Johnson IV tests of achievement. Rolling Meadows, IL: Riverside Publishing.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV tests of cognitive abilities. Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001, 2007). Woodcock-Johnson III. Itasca, IL: Riverside Publishing.

Nelson Cowan Nelson Cowan Cowan, Nelson

Working Memory

Working memory

1827

1829

Working Memory

The term *working memory* refers to a limited amount of information that is very easily kept in mind temporarily and is used to carry out mental tasks such as comprehending or producing language, solving problems, and making decisions. It has been important to understand working memory through research because of the key role of working memory in human cognition. For example, in language comprehension, one must keep in mind the sequence of words until a sentence makes sense. If one does not remember that a speaker said “*The man hoped the box …,” then when the sentence is completed by “… contained his missing tools,*” that second part of the sentence will not make sense. The working memory load is often increased by ambiguity or uncertainty. For example, if one could not tell if the word was *hoped* or *hopped*, both options have to be kept in mind until the second part of the sentence provides clarification. Because problem solving requires working memory (e.g., holding in mind the premises of a reasoning problem or partial products in a math problem), individual differences in intelligence and maturational level are highly correlated with working memory abilities. Various learning disabilities are often accompanied by working memory deficits.

The remainder of the entries discusses varieties, theories, and training of working memory.

Conceptualizations of Working Memory

Although researchers agree that working memory is important, different investigators seem to mean slightly different things when they refer to working memory, a point that has caused some confusion in the field. Some researchers include information that comes from any source, even guidelines that one gets from long-term knowledge. An example would be remembering which name goes with which face when one has met several new people. Some of this information will not be held in the conscious mind throughout, say, an hour-long event, but it is newly memorized information that may be easy to retrieve for the time being because one is still in the same context or situation as when the names and faces were first encountered. Other researchers restrict the term *working memory* to information that is in an active state, that is, in which one is currently thinking of the information, such as when one is currently looking for three kinds of fruit at the grocery store. For some researchers, working memory implies that one is actively doing processing while holding the information in mind. An example is a task in which one must remember the names of five individuals who were presented in a random order but must write the names in alphabetical order. These researchers use the term *short-term memory* when one is only holding the information, not also carrying out a process or manipulation of it. For other researchers, however, *short-term memory* and *working memory* are considered two labels for the same kind of memory, namely, any temporarily held information. This can also be called *immediate memory*.

Theories of Working Memory

In 1974, Alan Baddeley and Graham Hitch published a book chapter that has been seminal in this field. Previous to their work, authors found it sufficient to think of one mechanism for short-term memory, represented as a box in a flow diagram that represented the progression of information from sensory memory to short-term memory to long-term memory. If that were the case, however, then there should be severe interference if one had to use short-term memory in two ways at once, such as remembering a list of seven numbers while concurrently carrying out a reasoning problem that involved remembering premises and deducing some point from them. Instead, this kind of task produced only minimal interference. Baddeley and Hitch instead found that the conflict between such tasks was slight. They proposed that the term *working memory* should be used and that it involves multiple components working together. A phonological store and a visuospatial store (or buffer) were said to be involved in saving verbal and nonverbal and visual materials, respectively. A central executive component was needed to regulate the flow of information between the stores. In 2000, Baddeley added another component to the theory called the episodic buffer, needed to link two different kinds of information together or to hold meanings in working memory.

Other theories of working memory have been less committed to the notion of the separation of different components, as indicated, for example, in work by Akira Miyake and Priti Shah. It may be that each item held in working memory includes various kinds of features to represent the way the item looks, sounds, or feels and what it means. There may be interference in working memory between items with similar features (e.g., two red objects, two objects that both are means of transportation). In the theory suggested by Nelson Cowan, a few items can be held clearly in the focus of attention, whereas other items are maintained in a less clear form in terms of some of their features.

Research has begun to examine the brain correlates of working memory, contributing to the theoretical understanding of it. Different kinds of information (e.g., visual, phonological, meaning based) may be preserved in areas of the surface of the brain, or cortex, near where the information of each time comes in from the senses. The temporal areas over the ears would be heavily involved in retaining features of speech, whereas the occipital areas in the back of the brain would be heavily involved in retaining features of visual stimuli that cannot easily be verbalized. Some parietal lobe areas (in the upper part of the back of the brain not as far back as the occipital areas) seem involved in the focus of attention and frontal lobe areas seem involved in central executive functions such as controlling the focus of attention, updating information in working memory, switching between tasks, and inhibiting irrelevant responses. New techniques based on analyzing the pattern of brain activity in carefully designed experimentation allow a rough indication of what kind of information is held in working memory at a particular moment.

Training Working Memory

Given the importance of working memory for various cognitive activities, it has been hoped that intensive training on working memory tasks (in which information recently presented has to be recognized or recalled) could improve cognitive performance. So far, most of the evidence (like that from Randall Engle’s laboratory) suggests that training can improve the specific skills upon which the training is based, but that there is not much generalization of this improvement to other tests of intelligence. Most of this training, however, has been tried on typical individuals, and some remain hopeful that training working memory would be of particular use to individuals with deficient processing of information.

Nelson Cowan

See also Attention; Attention-Deficit/Hyperactivity Disorder; Cognitive Development, Theory of; Cognitive Diagnosis; Cognitive Neuroscience; Learning Disabilities; Long-Term Memory; Short-Term Memory

Further Readings

Baddeley, A. D. (2007). Working memory, thought, and action. New York, NY: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), The psychology of learning and motivation, Vol. 8 (pp. 47–89). New York, NY: Academic Press.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. Educational Psychology Review, 26, 197–223.

Cowan, N. (2016). Working memory capacity (Psychology Press and Routledge Classic Edition). New York, NY: Routledge.

D’Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. Annual Review of Psychology, 66, 115–142.

Gathercole, S. E., & Baddeley, A. D. (1993). Working memory and language. Hove, UK: Erlbaum.

Logie, R. H., & Cowan, N. (2015). Perspectives on working memory: Introduction to the special issue. Memory & Cognition, 43, 315–324. doi:10.3758/s13421-015-0510-x

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81–97.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960). Plans and the structure of behavior. New York, NY: Holt, Rinehart and Winston.

Miyake, A., & Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. Cambridge, UK: Cambridge University Press.

Mei Hoyt Mei Hoyt Hoyt, Mei

World Education Research Association

World education research association

1829

1830

World Education Research Association

The World Education Research Association (WERA) is a nonprofit association of national, regional, and international specialty education research associations committed to work together as a global community of organizations to advance education research as a scientific and scholarly field. According to WERA's website, the associations comprising WERA are resolved to communicate and collaborate "to address such issues as building capacity and interest in education research, advancing education research policies and practices, and promoting the use and application of education research around the world." WERA relates to education research worldwide, as it advances scholarship and internationalizes education research. This entry describes the establishment of WERA, its mission and goals in education research worldwide, and its programs and governance.

The Establishment of WERA

The establishment of WERA on April 18, 2009, in San Diego, CA, was the result of a 2-year effort of its funding members and representatives from worldwide educational research associations, including the American Educational Research Association, the Brazilian Black Researchers Association, the Educational Research Association of Singapore, and the Korean Educational Research Association, to name a few. WERA is chartered in the District of Columbia and includes three types of membership: (1) Governing Association Membership, (2) Association-in-Formation Membership, and (3) Individual Membership.

Mission and Goals

Mission and Goals

With a worldwide lens, WERA aims to undertake global initiatives and engage educational researchers who are cross cultural, international, and transnational in scope, conceptualization, and design. WERA hopes to build upon the diverse traditions of worldwide communities through sharing (e.g., skills) among its members.

Committees and Programs

Through initiatives such as International Research Networks, Capacity Development workshops and courses, WERA aims to facilitate international exchange and cooperation in education research, and disseminate education research across countries, regions, and diverse scholarly communities. In an effort to promote inclusiveness, WERA has promoted various initiatives through its WERA OUTREACH program that support education research efforts of scholars in developing nations in Africa, Asia, and the global South. In addition, WERA arranges symposia and keynote presentations to promote education excellence within and beyond this global group.

To share research and develop research capacities among the members, WERA holds an annual Focal Meeting, which consists of symposia sessions, papers, lectures, and other critical projects focusing on issues of significance to worldwide education research.

Governance

WERA’s governing structure includes an Executive Committee and a Council, which act in accordance with WERA’s purposes. The Executive Committee constitutes the Officers of WERA. In between meetings of Council, the Executive Committee acts on behalf of the Council and makes all necessary operational decisions for WERA.

WERA is still a young group, and it is not clear what its impact will be on the international education community. It also remains unclear in what ways different international groups share knowledge and which forms of educational research are valued among this diverse group.

Mei Hoyt

See also American Educational Research Association; Educational Research, History of; Educational Researchers, Training of

Further Readings

World Education Research Association Founded: Two-year effort leads to new international forum for collaboration. (2009). Educational Researcher, 38(5), 388–389. Retrieved from http://www.jstor.org/stable/20532570

Website

World Education Research Association (WERA): http://www.weraonline.org/

Lia Plakans Lia Plakans Plakans, Lia

Written Language Assessment Written language assessment

1830

1834

Written Language Assessment

Assessments of written language are used to gauge language and writing development for various formative and summative purposes. They have an important role in educational research, measurement, and evaluation, as they provide information about language and literacy development via analysis of writers’ use of language through performance. Moreover, analysis of written language is important in teaching language and literacy as well as in researching linguistic development and differences across writers, learners, and languages.

The assessment of written language falls into the category of *performance-based assessment*, wherein test takers respond in written form to prompts. For example, a written language assessment might prompt a test taker to compose a short story about a picture or to summarize content from a text the test taker has read. These performances would then be evaluated and scored by a qualified person trained to rate performances often using a rubric that describes features of a written language.

As a performance assessment, these constructed response test tasks are considered to be robust in terms of validity, as they can elicit processes of language use potentially similar to target language use situations. However, they can create challenges for reliability as scoring entails stages of judgment and interpretation. Performance assessments do not have readily objective scoring like selected response test items, such as multiple-choice questions, which have efficient and clearly identified correct answers. Given this trade-off, written language assessments require careful development in order to produce meaningful writing and useful outcomes.

In the past, written language was assessed indirectly and used as a vehicle to assess other aspects of language development. In the mid-20th century, multiple-assess other aspects of language development. In the mid-20th century, multiple choice items were employed to test knowledge of writing, perhaps better described as knowledge of language structure. Tests would include items that required test takers to select the correct structure of a sentence from several options, thus assessing writing indirectly. In foreign language teaching, writing was used frequently to show the ability to translate between a learner’s first language and the target language. These forms of writing assessment still occur and can be useful depending on purpose and the construct. However, in general, writing assessment has moved toward more direct methods of assessment and has expanded the construct well beyond the grammatical structure of language. This entry reviews the development of written language assessment, including identifying constructs of written language, creating prompts, and establishing a means of providing feedback. The entry concludes with a discussion of innovations in written language assessment.

Constructs of Written Language

The first step in developing an assessment of written language is to identify and define a construct on which to base the writing tasks and the evaluation of the performance. The construct of written language ability may vary depending on a number of variables about the test and test taker, including age, grade level, purpose, and whether the test is in a first or second language; yet some aspects of this construct are fairly stable. The foundational structures of a language are important in a construct of written language. In English, writing requires grammar and syntax as well as lexical depth and breadth. Structure may also include spelling and punctuation, conventions specific to the written form of the language (i.e., not in speaking). However, writing is much more than that. Constructs of writing also include ability to compose extended discourse, including the development and organization of ideas. Organizing writing includes the connection between sentences (i.e., cohesion) and logic of discourse through coherence. In addition, a construct of written language may include more nuanced aspects of language such as a writer’s ability to convey authorial voice, pragmatic competence, and style. Common to assessing writing, these aspects of the construct focus on the written product, with less attention to the processes of composing or knowledge about writing. Once a construct of writing has been defined for the test purpose and test takers, the next step is developing a task to elicit this construct.

Prompts for Assessing Written Language

Writing assessment tasks generally fall into two categories: (1) bare prompts for writing that is entirely writer generated and (2) prompts that provide some stimulus or content for writers to draw on. Both types are commonly used and have advantages and disadvantages.

Independent writing prompts allow writers to create their own content on a topic. These types of tasks often follow a genre-based approach to writing; for example, asking test takers to narrate a significant event in their lives or to argue a position on a controversial topic. These tasks allow considerable freedom to writers in terms of content, which can be seen as both positive and negative. In some cases, writers feel taxed and anxious to generate ideas in a timed writing situation. They may feel pressure when trying to think of a novel or engaging topic on which to write. Research shows that independent writing requires more planning time, given the need to brainstorm content to develop a topic. However, these tasks can be somewhat easier to design as stimulus texts do not need to be created nor is the writer’s level of reading comprehension a factor.

Writing tasks that are integrated with visuals or source texts provide writers with content and ideas on which to write. Research has shown that source material can provide writers with not just topic development but also support in organization and some language features such as key vocabulary. In these tasks, test takers are prompted to read, listen, or view a visual. In most cases, the prompt then guides the writer to use content from these sources in composing writing. For example, the integrated writing task in the Test of English as a Foreign Language asks writers to read a short passage, then listen to a lecture on the same topic but with an opposing view. The prompt then asks the test taker to summarize the lecture in terms of its differences with the reading. Integrated writing tasks require more development to assure that texts or visuals are clear and level appropriate. Another challenge with these tasks is the concern that source material muddies the construct of writing by including reading, listening, or other skills.

Evaluating Performances on Written Language Assessments

Once a test taker has responded to a prompt in a written language assessment, some process must be undertaken to arrive at a score or to provide feedback on the writing. A scoring rubric is often used for this purpose. A rubric can take many forms but typically consists of descriptions about certain writing features as well as a scale to judge the level of quality. Rubrics may be adopted or developed by the test user; however, they should always align with the construct of writing that was the basis for the assessment.

Rubrics and Data-Driven Measures for Writing

Two types of rubrics or scoring scales are common in writing: analytic and holistic. Analytic rubrics provide a separate score for each quality descriptor, such as 5 points for organization, 10 for language use, and 7 for idea development. These types of scales are potentially more useful in providing specific feedback to writers and can be more informative for raters. However, they are more time consuming and may not capture the overall quality of a piece of writing when added up. In contrast, a holistic scale gives a single score to the whole performance. Holistic scales may include descriptors of different qualities or features but do not try to separate them, viewing the writing as a whole rather than a composite. Holistic scales have been found to be more reliable and can be more efficient. They are, however, less useful in providing feedback on specific strengths and weaknesses and are better for decision-making assessment purposes rather than diagnostic or formative uses.

An alternative to rubrics often used in research on written language assessment is a data-driven approach to evaluating writing in which certain features of written language are operationalized in a way that can be reliably measured. For example, to evaluate fluency, a common measure is to count the words in a performance or the number of words in each sentence or clause. Accuracy, another common feature assessed in written language can be measured using several data-driven metrics such as (a) total number of errors, (b) errors per clause, (c) errors per 100 words, or (d) number of error-free clauses. Fluency and accuracy are commonly followed by evaluation of complexity. Complexity can be defined as the writer’s use of language beyond simple structures, showing efficient and sophisticated use of the language. While accuracy is fairly rule-governed, complexity is less so, making agreed-upon metrics less prevalent; complexity can appear through coordination, subordination, phrasal relationships, and length. These three linguistic features are often considered salient traits for distinguishing language proficiency at different levels. However, there are limitations to what this language features approach can evaluate.

Raters

The process of scoring performance assessments relies not only on rubrics but also on someone to use, interpret, and apply the rubric to the piece of writing. This individual may be a teacher providing feedback on a classroom assessment or a trained rater in the case of larger-scale assessments (automated scoring is discussed in the following section). Employing raters in scoring recognizes that writing is communication with a rater representing a reader. Training raters on scoring written language assessment is critical to achieve reliability and to support an assessment's validity. Usually training familiarizes raters with the rating rubric or scale including reading sample performances at each level of the scale. This is followed with practice rating as a group and individually with opportunities to discuss results and concludes with an individually scored set of ratings to assure rater consistency. Once trained, raters score writing individually with frequent brush-up sessions, which allows them to recalibrate to the intended scoring scale. Best practice entails having at least two ratings for each performance, and if these do not agree, a third can be conducted. The rating process in performance assessments is a necessary aspect of scoring in order to provide reliable and meaningful scores or feedback; however, this human element is also what makes performance assessment expensive and time consuming.

Future Directions for Written Language Assessment

The assessment of written language will likely continue to be enacted through performance assessments, yet there is room for innovation and new approaches within that format. One recent development is in rating—automated essay scoring (AES). In fact, automated scoring has been around since the 1960s, but it has recently become more pervasive in large-scale assessments (e.g., Test of English as a Foreign Language). AES systems are essentially developed by training a computer to score writing based on human ratings. There has been some backlash against AES, particularly from the field of rhetoric and composition, as it removes humans from written communication. AES scoring systems have been found to have high reliability, at least as high, if not higher, than human raters. However, these systems are limited in what they can judge in writing, as they cannot model more subjective or interpretive reading processes and may not recognize pragmatic competence, appropriate use of source material, or stylistic features of writing.

Another area for innovation is assessments of written language that focus on the processes in language production, as current performance assessment attends only to the resulting product of writing. The constructs presented early in this entry are product based; however, there are also constructs that look at how individuals compose writing, including aspects such as planning, drafting, revising, or editing. These are processes that are used by successful writers, thus being able to understand test takers' strengths and weaknesses therein could be useful for diagnostic testing and classroom assessment. However, gathering evidence of process is difficult and solving challenges in scoring make this direction formidable and perhaps less feasible for large-scale assessment. However, some options include allowing untimed writing assessment, requiring multiple drafts by writers, and using portfolio assessments, which entail multiple drafts and writing artifacts to show development of written language over time.

Lia Plakans

See also Formative Assessment; Performance-Based Assessment; Rubrics; Scales

Further Readings

Belanoff, P. (1997). Portfolios (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Crusan, D. (2010). Assessment in the second language writing classroom. Ann Arbor: University of Michigan Press.

Cumming, A. (1998). Theoretical perspectives on writing. Annual Review of Applied Linguistics, 18, 61–78.

Grabe, W., & Kaplan, W. (1996). Theory and practice of writing: An applied linguistic perspective. Harlow, UK: Longman.

Hout, B. (2002). (Re)articulating writing assessment for teaching and learning. Logan: Utah State University Press.

Knoch, U. (2007). “Little coherence, considerable strain for reader”: A comparison between two rating scales for the assessment of coherence. Assessing Writing, 12, 108–128.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. Assessing Writing, 13, 111–129.

Weigle, S. (2002). Assessing writing. Cambridge, UK: Cambridge University Press.

Yancy, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. College Composition and Communication, 50(3), 483–503.

Z

David Westfall David Westfall Westfall, David

Z Scores

Z Scores

1835

1837

Z Scores

A *Z* score, or standard score, is the number of standard deviations an observation is away from the mean of the corresponding reference population. It is a measurement of the value of a single observation in relationship to the scores of a group of observations (population). *Z* scores transform units of analysis into a standardized form, allowing for comparison of variables measured in different units. When population parameters are known, *Z* scores are powerful for locating an individual observation in relation to all observations. When population parameters are not known, the *Z* distribution (normal distribution) changes slightly. As such, *Z* scores are a fundamental concept in statistical analysis. The transformation of units of analysis into standardized units (units of standard deviation) can be seen in the formula:

$$z = \frac{X - \mu}{\sigma}.$$

The numerator of the equation calculates the difference between an individual observation and the mean (or average) of all observations. This difference is divided by the standard deviation of the population in the denominator. As a reminder, the standard deviation is the average distance, each observation in a population is away from the mean of all observations in the population.

For example, if a student in class scores 90 on an exam (*X*), the mean score of all exams in the class is 80 (μ), and the standard deviation for exam scores is 5 (σ).

$$z = \frac{X - \mu}{\sigma} = \frac{90 - 80}{5} = \frac{10}{5} = 2.$$

A *Z* score of 2 informs the researcher that the student’s test score was 2 standard deviations above the mean of all test scores in the class. *Z* scores can be positive, above the mean of the population, or negative, below the mean of the population.

A *Z* score of 0 indicates the observation has a score identical to the mean of the population. A *Z* score greater than 0 indicates an observation that scores higher than the mean of the population. A *Z* score less than 0 indicates an observation that scores lower than the mean of the population. The higher the absolute value of the *Z* score, the more extreme an observation is from the mean of the population. A *Z* score of 1 is 1 standard deviation above the mean of the population. A *Z* score of −1 is 1 standard deviation below the mean of the population.

In addition to standardizing scores, the *Z* score is a valuable calculation for several reasons. First, computing a *Z* score allows the researcher to locate a score in a distribution of scores, indicating if the score is good or bad and above the mean or below the mean, in relation to all scores in a distribution. A student may seem to score poorly on an exam, 50 out of 100 possible points; however, if the mean exam score is 40 and a standard deviation of 10, the student actually scored better than approximately 85% of the class—thus indicating a difficult exam.

Second, it allows the researcher to compare scores on variables that may appear quite different from each other or that are from two different normal distributions—for example, exam scores and time spent studying for an exam. The standard deviation of the number of minutes spent studying for an exam in a population is not directly comparable to the standard deviation of the exam scores of a population. A standard deviation of 1 hour for time spent studying and 5 points for exam scores are not comparable, one unit is hours and the other unit is points. By converting to a *Z* score, these calculations are standardized into units of standard deviation and are thus comparable. *Z* = 2 carries the same meaning in both populations, allowing the researcher to determine which is more extreme from their respective means, whereas a standard deviation calculated for each individually does not carry the same meaning. No matter what unit a variable is expressed in, when standardized to a *Z* score, the units become the same, that is, units of standard deviation. The value of the mean is 0 and the standard deviation is the same for each variable.

Third, *Z* scores allow for the calculation of the probability that a certain score will occur in a normal distribution. Given a few conditions, the central limit theorem states that if a sample is large enough, variance is finite, and expected values are well defined, the arithmetic mean of random independent variables will be approximately normally distributed. That is, mathematically most observations will cluster around the mean of all observations, and the higher the number of standard deviations away from the mean, the fewer the number of observations. The larger the sample drawn, the more the distribution of the sample will resemble a normal distribution. However, samples as small as 30 cases can be treated as normal distributions.

The normal distribution, expressed as a histogram, is a symmetrical distribution that takes the shape of a bell due to variation around the mean. The mean is a *Z* of 0. The median and mode are also found in the center of the distribution. The tails of the normal distribution, or curve, are asymptotic. This means that they will continuously draw nearer to each other but never quiet reach. Imagine holding two pencils with the sharpened tips facing each other one inch apart. By reducing the distance by half periodically, the points will continuously draw nearer to each other. The distance between them will become infinitesimally small but never touch as the distance is being reduced by half each time. The tails of a normal distribution feature this same characteristic. This allows for the possibility of extremely rare observations, no matter how unlikely they may seem (i.e., an individual who is 8 feet tall).

Of all the observations, 68.26% fall between a *Z* of +1 and −1, 95.44% of all observations fall between a *Z* of +2 and −2, and 99.72% of all observations fall between a *Z* of +3 and −3. Only 0.26% of all cases are more extreme than a *Z* of +3 and −3. Fifty percent of all observations are higher than the mean, or *Z* of 0, and 50% of all observations are lower than the mean. Ninety-five percent of all cases lie between *Z* scores of 1.96 and −1.96 (also known as the critical values). Due to these characteristics or properties of a normal distribution, researchers are able to make statements about sample data with known confidence levels. It also allows a researcher to answer questions such as what percentage of students on a standardized test scored better than a particular score, or less than a particular score, thus allowing for a percentile rank.

Z scores are used for normal distributions. If the distribution is not normal, or the parameters of the population are not known, the distribution changes to a *t* distribution or Student's *t* distribution. That is, if the researcher does not know the mean or standard deviation of the population, the mean and standard deviation of the sample can be used; however, the distribution changes from a normal distribution to a *t* distribution. The *t* distribution has many of the same characteristics of the normal distribution. Observations cluster around the mean. The tails of a *t* distribution are asymptotic (they draw nearer to each other but never touch); the mean, median, and mode are in the center; and the curve is symmetrical. Half of all observations are above the mean and the other half below the mean. In addition, all of the percentages remain the same as in a normal distribution. There is one major change in the distribution, however. When calculating a *Z* score, the parameters (characteristics about the population), standard deviation, and mean are known. When the parameters about the population are not known, statistics from the sample may be used (the sample standard deviation and mean). When doing so, the distribution becomes an estimate of the normal distribution. Only one sample of an infinite number of possible samples is obtained in a population. As such, the mean of the obtained sample is an estimate of the population mean. If another sample was drawn, the mean would likely be different but again an estimate of the population mean.

As such, the shape of the curve, percentage of cases between each standard deviation (or *t* score) and the mean, is impacted by the sample size; the smaller the sample size, the greater the number of cases in the tails of the distribution. The larger the sample size, the more the *t* distribution takes the shape of a normal distribution. This feature of the *t* distribution allows researchers to make more conservative observations or predictions. In a normal distribution, a 95% confidence level or 5% significance level is between the critical values of +1.96 and *Z* of −1.96.

The confidence level is the degree of confidence that the researcher has that he or she is not committing a Type I error or rejecting the null hypothesis when the null hypothesis is actually true. Conversely, the significance level is the amount of risk the researcher is willing to take by committing a Type I error. Traditionally, confidence levels are set at 90%, 95%, 99%, and 99.9%. The higher the confidence level, the more extreme an observation must be before the researcher is willing to say that the observation is different from the population and it is not due to chance. Higher confidence levels are traditionally used in medical research. In social science research, a 95% confidence is the norm.

In a *t* distribution, critical values for confidence (the point at which the researcher is willing to say the observations are statistically different from the population and it is not due to chance) levels are directly impacted by sample size (and degrees of freedom). As such, the critical values for a 95% confidence level cannot be determined for *t* scores in the same way that can be done in a normal distribution with *Z* scores. Degrees of freedom are the number of values that are free to vary in a calculation. Degrees of freedom provide another estimate of the parameters. For example, if the researcher knows that the sum of any three numbers is 10, the first number is free to be any number from negative infinity to positive infinity. The second number is free to vary and take the form of any number from negative infinity to positive infinity. However, when the third number is reached, it is locked or bound and cannot vary. It must be whatever value is needed to reach a sum of 10 for the number set. Thus, the three number set has 2 degrees of freedom. In this way, degrees of freedom are directly impacted by sample size and provide a conservative estimate. Degrees of freedom factor into the denominator of a calculation and artificially inflate the output, making the estimate more conservative.

David Westfall

See also Normal Distribution; Significance

Further Readings

Carlson, K. A., & Winquist, J. (2014). An introduction to statistics: An active learning approach. Thousand Oaks, CA: Sage.

Salkind, N. J. (2014). Statistics for people who (think they) hate statistics (5th ed.). Thousand Oaks, CA: Sage.

Schutt, R. (2017). Understanding the social world: Research methods for the 21st century. Thousand Oaks, CA: Sage.

Szafran, R. (2012). Answering questions with statistics. Thousand Oaks, CA: Sage.

Rana S. Hinman Rana S. Hinman Hinman, Rana S.

Zelen’s Randomized Consent Design Zelen’s randomized consent design

1837

1839

Zelen’s Randomized Consent Design

Zelen’s design (also known as the randomized consent, or prerandomization, design) is a type of randomized controlled trial in which randomization occurs *before* participants give informed consent to participate. In 1979, Marvin Zelen proposed the randomized consent design method for researchers and clinicians for easy enrollment of patients into clinical trials. This entry explains Zelen’s design and its purpose, single and double consent options, as well as advantages and disadvantages. The entry concludes with an overview of the ethical issues surrounding the use of Zelen’s design in clinical research.

Limitations of Conventional Informed Consent

In a conventional randomized controlled trial, participants provide informed consent before being randomly allocated to (usually) the intervention or control treatment groups. This can present a number of difficulties for researchers and for the clinicians who are oftentimes tasked with entering patients into the trial. Obtaining consent to participate in conventional trials requires full disclosure to participants about all potential benefits and risks associated with participating in the trial, including the intervention being tested and the alternative (control) treatments used for comparison. Participants may be concerned about being randomized to the control group (often seen as the inferior or less effective intervention) and/or not receiving their preferred intervention and refuse to participate. This can slow trial recruitment and may limit the feasibility of conducting research in some circumstances. It also has implications for the risk of bias. Participants who hold positive attitudes toward, and beliefs about, the intervention being studied are more likely to enroll than those who don’t. This may mean the sample ultimately recruited is not reflective of the broader clinical population. Importantly, knowledge about the intervention of interest and the alternative (control) can influence participants’ adherence to treatment as well as their self-reported response to treatment. It is widely acknowledged that patient pretreatment expectations of outcome influence treatment response on patient-reported domains such as pain, physical function, quality of life, and mental health. Put simply, participants who expect to get better with treatment at the outset tend to report better treatment outcomes than those who don’t. For participants allocated to an intervention that they perceive to be less effective than the alternative, resentful demoralization may occur. This is particularly relevant to participants allocated to control, placebo, or “usual care” trial arms. This increases the risk of increased dropouts and/or reporting of poorer outcomes and threatens the validity of the trial.

Randomized Consent

Zelen’s design can overcome some of the problems experienced with conventional consent procedures by randomizing participants before consent to participate has been obtained. Zelen proposed two types of randomized consent design termed single and double consent designs, respectively. Participants randomized to the control group (usually a standard treatment or usual care) with single consent are not asked for consent to participate in the trial. Instead, they are provided standard treatment (or usual care) for their medical disease or condition, without knowing they are participating in the trial and without knowledge of the other treatment arm. Their data are collected (usually) as part of standard treatment and used as a comparison for the intervention group. Only participants allocated to the intervention group are asked to consent, and if they decline the intervention of interest, they receive standard treatment (or usual care). With double consent designs, participants in the control group are also approached to provide consent to receive standard treatment. Those who decline receive an alternative treatment that can include the experimental intervention. With both designs, trial data are analyzed according to original random allocation, irrespective of the treatment received.

Advantages

Zelen’s design may be more attractive, when a trial requires clinicians to recruit participants from among their patients (as opposed to researchers advertising for volunteers from within the broader community). The process of obtaining informed consent is simplified with Zelen’s design. Clinicians may be more comfortable with the consent procedures and thus more willing to participate, because they only have to seek patient consent for treatment they know the patient will receive, without explaining the process and risks of randomization. Similarly, patients are not at risk of disappointment or resentful demoralization. Zelen’s design may be preferable for certain medical conditions (such as cancer) or therapies (such as surgery) where patients may hold strong treatment preferences or in settings where obtaining consent prior to randomization is more difficult (such as the emergency department). It can be most useful for “screening” trials where real-world population-based estimates of effect are desirable.

Disadvantages

With Zelen’s design, participants may refuse to consent after allocation. They may choose to withhold their data from analysis, which may introduce bias if nonconsent rates differ across trial arms. Participants may consent to data collection but refuse their allocated intervention. Risk of contamination increases when patients refuse their randomized intervention cross over to the alternative treatment arm, diluting observed treatment effects and making it harder for the trial to detect significant differences in outcomes between treatment groups. If a large number of patients refuse randomized intervention and cross over, a reduction in statistical power will occur, meaning that researchers planning a trial with Zelen’s design need to inflate the sample size a priori to account for anticipated refusals. An inflated sample size has spin-off disadvantages in increasing the duration of the trial as well as the costs involved in running it. Close (intrusive) monitoring of participants and data collection procedures (outside of routine clinical care) is not feasible with Zelen’s design, as participants allocated to the control group may be alerted to the trial and become unblinded.

Ethical Considerations

Zelen’s design is considered ethically controversial because participants do not provide consent to be randomized. Critics argue that the partial disclosure involved in obtaining consent after randomization has taken place is unethical. Although consent is sought from participants to receive their allocated treatment, many argue that it is unethical not to also disclose the process behind initial group assignment (i.e., that allocation was due to chance alone). Proponents argue that, in some cases, it is more ethical to employ Zelen's design than the conventional consent procedures. Many believe that the patient–clinician relationship may be compromised when treatment allocation is subject to chance alone and/or by forcing clinicians to disclose risks of standard treatment (or usual care). Ultimately, Zelen's design may be considered ethical or unethical depending on the circumstances. When employing Zelen's design, researchers must ensure that scientific advantages of its use outweigh the ethical concerns.

Rana S. Hinman

See also Causal Inference; Ethical Issues in Testing; Experimental Designs; Informed Consent; Random Assignment; Threats to Research Validity

Further Readings

Homer, C. S. E. (2002). Using the Zelen design in randomized controlled trials: Debates and controversies. Journal of Advanced Nursing, 38, 200–207. doi:10.1046/j.1365-2648.2002.02164.x

Zelen, M. (1979). A new design for randomized clinical trials. New England Journal of Medicine, 300, 1242–1245. doi:10.1056/NEJM197905313002203

Zelen, M. (1990). Randomized consent designs for clinical trials: An update. Statistics in Medicine, 9, 645–656.

Asmalina Saleh Asmalina Saleh Saleh, Asmalina

Joshua Danish Joshua Danish Danish, Joshua

Zone of Proximal Development Zone of proximal development

1839

1842

Zone of Proximal Development

The zone of proximal development (ZPD) is a core concept in sociocultural theories of learning, which build on the work of Lev Vygotsky. The ZPD is described as the difference between how learners can perform on their own and how the learners can perform with the help of a “more knowledgeable other.” The ZPD was conceptualized as both a theoretically driven form of assessment and a core theoretical construct to explain the relationship between learning and development. Specifically, Vygotsky argued that learning happens best when students are engaging with concepts at the edge of their competence—the “developing” psychological functions (e.g., perception, speech, thinking) rather than those that are already developed, which Vygotsky referred to as “fossilized.” The task of the researcher is therefore to identify the ZPD relevant to the learner’s psychological stage and to measure the learner’s current maturing functions. As an assessment, the ZPD is intended to focus less on what the learners can already do but, rather, to determine what they are currently ready to learn. Rather than focusing on whether students have already mastered core content, the ZPD identifies when they are ready to master the “next” content and thus can directly support instructional efforts.

Vygotsky’s characterization of the ZPD is also built on the assumption that in order to understand child development, one must take into account how the child as a whole interacts with his or her environment. This means that each maturing function—be it perception, memory, or speech—must be understood in relation to one another and not independently. Thus, understanding learners’ psychological structure requires attending to qualitative psychological changes in the learners as an entire being over specific time points as well as the learners’ interaction with their environment. In particular, work with the ZPD recognizes that each cultural and historical context has local expectations of what a child should be able to do within a given age range, referred to as the objective ZPD. For example, preschool children are typically expected to master basic social communicative skills and to do so through play. Once they enter school, children are expected to learn formal disciplinary content such as formal mathematics.

Individual learners, however, have their own maturing capabilities or their subjective ZPDs. In order for learners to move to the next development stage, they must confront contradictions between what they want to do and how their maturing capabilities affords or constrains them. To support this shift, leading activities, or activities that will encourage the learner to take part in actions to promote new psychological functions, must be analyzed. These activities are considered leading when an individual takes part in them and they significantly change the individual. It is critical to note that maturing functions, and not the starting point of each stage of development, are the end result of participating in leading activities. The ZPD thus refers both to the presumed process for how learning and development interact and how one might assess learning within that theoretical context.

The ZPD as a Lens for Research

While the ZPD is focused on development, researchers have often used it as a lens to support student learning. This is in part because the ZPD as a theoretical framework provides the mechanisms involved in supporting both learning and development. As a result, research has highlighted how a range of tasks can be used to support the learner, how an instructor can interact with the student, and, to some extent, how to measure students’ maturing functions. However, care must be taken to differentiate between learning and development. Vygotsky’s concern is not about specific skills but, as noted, a qualitative change in how the individual interacts with the world. For instance, even if a learner is at a stage ready for complex linguistic tasks, being able to perform certain undertakings such as listening and speaking does not mean that the learner has reached the next developmental stage. Rather, the researcher must take into account the objective ZPD.

The ZPD has inspired several new lines of assessment research including an explicit examination of scaffolding and dynamic assessments (DAs). In the scaffolding literature, considerable work has been done on the nature of support that instructors can provide to learners. Here, the focus is not necessarily on measuring the learner’s maturing faculties but rather on analyzing how the learner responds to instructional support, or scaffolds, by a more knowledgeable other. Another key characteristic of this work is the concept of fading, or analyzing when the learner is able to perform a task without support. Initial work focused on the effectiveness of certain forms of scaffolds in one-on-one settings with instructors but later expanded into the use of material and/or technological scaffolds such as cognitive tutors and computer-assisted prompts.

DA seeks to capture an individual’s potential for learning new concepts. DA models fall on an interventionist–interactionist spectrum; the former focuses on standardization of the protocol, whereas the latter is more responsive to the learner and ad hoc in nature. Interventionist DA or standardized dynamic tests are concerned with psychometric properties of the test and therefore tend to be more heavily standardized. Interactionist DA, however, is less focused on quantitative measurement of student performances and is more interested in interpreting and understanding the learner’s potential competencies qualitatively. Although the prompts provided to the learner can be somewhat standardized, the nature of the interaction is more responsive to the learner, so that the researcher is able to better understand the learner’s maturing functions. Research in this tradition has also expanded into group DA, where the collaborative performance of the group is attended to, rather than individual capabilities.

Asmalina Saleh and Joshua Danish

See also Learning Progressions; Scaffolding

Further Readings

Chaiklin, S. (2003). The zone of proximal development in Vygotsky’s analysis of learning and instruction. In A. Kozulin, B. Gindis, V. S. Ageyev, & S. Miller (Eds.), Vygotsky’s educational theory in cultural context (pp. 39–64). Cambridge, England: Cambridge University Press.

Griffin, P., & Cole, M. (1984). Current activity for the future: The Zo-ped. New Directions for Child and Adolescent Development, 23, 45–64.

Hedegaard, M. (2005). The zone of proximal development as basis for instruction. In H. Daniels (Ed.), An introduction to Vygotsky (pp. 227–254). New York, NY: Routledge.

Lauchlan, F., & Elliott, J. (2001). The psychological assessment of learning potential. British Journal of Educational Psychology, 71(4), 647–665.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1984). The zone of proximal development: Some conceptual issues. New Directions for Child and Adolescent Development, 23, 7–18.

Appendix A: Resource Guide

Agencies and Organizations

American Educational Research Association (AERA): http://www.aera.net/

A professional organization founded in 1916 comprised of educational researchers from the United States and around the world. AERA's mission is to advance knowledge about education, encourage scholarly inquiry related to education, and promote the use of research to improve education and serve the public good.
American Evaluation Association (AEA): http://www.eval.org/

AEA is a professional association of evaluators devoted to the application and exploration of program evaluation, personnel evaluation, technology, and many other forms of evaluation.
American Institutes for Research (AIR): http://www.air.org/

Founded in 1946, AIR is one of the largest nonprofit behavioral and social science research and evaluation organizations in the world.
American Psychological Association (APA): http://www.apa.org/

The primary professional organization for psychologists.
Council for the Accreditation of Educator Preparation (CAEP): http://caepnet.org/

First recognized in 2013, CAEP seeks to advance equity and excellence in educator preparation through evidence-based accreditation of programs at the certificate, licensure, associate's, bachelor's, master's, post-baccalaureate, and doctoral levels in degree-granting institutions of higher education. CAEP is a consolidation of the National Council for the Accreditation of Teacher Education (NCATE) and the Teacher Education Accreditation Council (TEAC).
Center for Research on Evaluation, Standards, and Student Testing (CRESST): http://cresst.org/

Founded in 1966, *CRESST conducts research focused on assessment,* *evaluation, methodology, and technology to improve student learning and educational outcomes*.

Council of Chief State School Officers (CCSSO): http://www.ccsso.org/

CCSSO is a national nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues.

Institute of Education Sciences (IES): https://ies.ed.gov/

IES, part of the U.S. Department of Education, is a leading source for rigorous, independent education research, evaluation, and statistics in the United States. IES includes several centers including the National Center for Education Statistics (NCES), National Center for Education Evaluation and Regional Assistance (NCEE), and National Center for Special Education Research (NCSER).

Joint Committee on Standards for Educational Evaluation (JCSEE): http://www.jcsee.org/

Created in 1975, the Joint Committee is a coalition of major professional associations in the United States and Canada concerned with the quality of educational evaluation.

National Board for Professional Teaching Standards (NBPTS): http://nbpts.org/

Established in 1987, NBPTS is an independent, nonprofit organization working to advance the quality of teaching and learning for all students.

National Center for Education Statistics (NCES): https://nces.ed.gov/

Located within the Institute of Education Sciences, NCES fulfills a Congressional mandate to collect, collate, analyze, and report complete statistics on the condition of American education; conduct and publish reports; and review and report on education activities internationally.

National Center for Education Evaluation and Regional Assistance (NCEE): https://ies.ed.gov/ncee

Part of the IES, NCEE helps educators and policy makers make evidence-based decisions about educational programming. It conducts large-scale

evaluations of federally funded education programs and practices, supports locally developed research projects and technical assistance through ten Regional Educational Laboratories (https://ies.ed.gov/ncee/edlabs), and the dissemination of research through the What Works Clearinghouse (https://ies.ed.gov/ncee/wwc), the National Library of Education, and the Education Resources Information Center (ERIC) online database.
National Center for Special Education Research (NCSER): https://ies.ed.gov/ncser/

An IES center, NCSER supports rigorous research on children and youth with and at risk for disabilities by advancing the understanding of and practices for teaching, learning, and organizing education systems.
National Council on Measurement in Education (NCME): http://www.ncme.org/NCME

The primary professional organization for psychometricians.
National Science Foundation (NSF): https://www.nsf.gov/

NSF is a US federal agency that supports fundamental research and education in all the non-medical fields of science and engineering.
Organisation for Economic Cooperation and Development (OECD): http://www.oecd.org/

The OECD exists to promote policies that will improve the economic and social well-being of all people worldwide. It enables governmental agencies to work together to understand what drives educational, economic, social, and environmental change. The Programme for International Student Assessment (PISA) is sponsored by the OECD.
Research Triangle Institute (RTI): https://www.rti.org/

RTI is an independent, nonprofit organization that provides research, development, and technical services to government and commercial clients worldwide on subjects that include education and workforce development.
WestEd: https://www.wested.org/

WestEd is a nonpartisan, nonprofit research, development, and service agency that resulted from a 1966 merger of the Southwest Regional Educational Laboratory (SWRL) and the Far West Laboratory for Educational Research and Development (FWL). WestEd provides consulting and technical assistance, evaluation, policy analysis, professional

development, and research to improve learning and promote healthy development.
What Works Clearinghouse: https://ies.ed.gov/ncee/wwc/

WWC reviews research on educational interventions and policies.

Books

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Widely used textbook and guide to basic educational measurement issues and item response theory.
Frey, B. B. (2015). There’s a stat for that!: What to do & when to do it. Thousand Oaks, CA: SAGE.

A comprehensive guide for matching the right statistic with the right research design.
Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Introduced the concept of research paradigms as a way of understanding scientific methodology.
Linn, R.L. (Ed.) (1987). Educational measurement (3rd ed.). New York: Macmillan.

Full of important chapters, including Samuel Messick’s classic article presenting the unitary view of validity.
Maxwell, J. A. (2012). *Qualitative research design: An interactive approach* (Vol. 41). Thousand Oaks, CA: SAGE.

A very understandable and approachable guide to qualitative research.
Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Classic analysis of the varied threats to research validity and what can be done to protect against them.
Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the*

behavioral sciences. New York: McGraw Hill.

Still one of the few comprehensive guides for the use of nonparametric statistics in social science research.

Journals

American Educational Research Journal: http://www.aera.net/Publications/Journals/American-Educational-Research-Journal

The flagship journal for the American Educational Research Association.
Educational Measurement: Issues and Practice: https://www.ncme.org/ncme/NCME/Publication/Educational_Measurement/

An NCME journal aimed at practitioners.
Educational Researcher: http://www.aera.net/Publications/Journals/Educational-Researcher

Publishes research on a wide variety of topics in educational science.
Journal of Educational and Behavioral Statistics: http://www.aera.net/Publications/Journals/Journal-of-Educational-Behavioral-Statistics

Focuses on new approaches to data analysis.
Journal of Educational Measurement: https://www.ncme.org/ncme/NCME/Publication/Journal_of_Educational_M hkey=6380e466-a3ec-4154-b06f-96888a76ec97

Produced by the National Council on Measurement in Education, this journal publishes original research and introduces new instruments.
Practical Assessment, Research & Evaluation: http://pareonline.net/

PARE is an online journal providing access to refereed articles that can have a positive impact on assessment, research, evaluation, and teaching practice.
Review of Educational Research: http://www.aera.net/Publications/Journals/Review-of-Educational-Research

Publishes major literature reviews of educational topics

Publishes major literature reviews of educational topics.

Policies and Standards

Individuals with Disabilities Education Act (IDEA): https://sites.ed.gov/idea/

First approved in 1975 as the Education for All Handicapped Children Act and renamed in 1990, IDEA, Public Law 94–142, mandates that all children ages 3–21 have a right to a free and appropriate education designed to meet their education needs. Education and related services must be provided in the most appropriate and least restrictive environment.
InTASC Model Core Teaching Standards: http://www.ccsso.org/Documents/2011/InTASC_Model_Core_Teaching_St

Developed in 2011 by CCSSO's Interstate Teacher Assessment and Support Consortium (InTASC), the Model Core Teaching Standards outline what teachers should know and be able to do to ensure every K-12 student reaches the goal of being ready to enter college or the workforce in today's world. These standards outline the common principles and foundations of teaching practice.

Web Resources

ERIC Document Reproduction Services: https://eric.ed.gov/

An Institute of Education Sciences service that accepts, stores, and provides scholarly reports and articles.
The Evaluation Center at Western Michigan University: https://wmich.edu/evaluation

This site provides guidelines and evaluation checklists for planning and conducting evaluations.
National Assessment of Educational Progress (NAEP): https://nces.ed.gov/nationsreportcard/

NAEP, known as "The Nation's Report Card," is the largest nationally representative and continuing assessment of what students in the United States know in various subject areas. In 2017, NAEP began administering

digitally based assessments (DBA) for mathematics, reading, and writing, with additional subjects to be added in 2018 and 2019.
Online Evaluation Resource Library: https://oerl.sri.com/home.html

Sponsored by the National Science Foundation, this site provides resources for creating evaluation plans, data collection instruments, and evaluation reports.
Organisation for Economic Cooperation and Development (OECD) Online Education Database: http://www.oecd.org/education/database.htm

Home to a variety of up-to-date international education statistics.
Programme for International Student Assessment (PISA): http://www.oecd.org/pisa/aboutpisa/

PISA is a triennial international assessment intended to evaluate education systems by testing the science, mathematics, reading, collaborative problem solving, and financial literacy knowledge of 15-year-old students worldwide. In 2015 over half a million students from 72 countries took the exam.
ProQuest Education Database: http://www.proquest.com/products-services/pq_ed_journals.html

Database focusing on the theory and practice of education with more than 1,000 full-text journals and 18,000 dissertations.
PsychINFO: http://www.apa.org/pubs/databases/psycinfo/index.aspx

Provides citations and references for social science researchers.
W.K. Kellogg Foundation Evaluation Handbook: http://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook

This site provides an evaluation framework for project directors and others interested in program evaluation.
Research Methods Knowledge Base: http://www.socialresearchmethods.net/kb/index.php

This is a web-based textbook that provides extensive information about various research methods and analyses used in social sciences.

Appendix B: Chronology

Date	*Event*
600	Around this time, the Chinese dynasties of Sui and, later, T'ang, begin relying on a standardized system of imperial examination called *Keju* to identify talented people for future positions in civil service.
1599	Edward Wright publishes a book on navigation suggesting the use of the median to summarize observations.
1733	While exploring coin flipping, Abraham de Moivre studies what becomes known as the normal curve.
1761	Thomas Bayes suggests a method of estimating likelihood of occurrences that becomes known as Bayes' Theorem.
1786	William Playfair introduces line charts and bar charts.
1792	British professor, chemist and engineer, William Farish, assigns quantities (grades) to evaluate his students' performance. This allows him to rank students.
1802	Pierre-Simon Laplace estimates the population of France using sample data.
1805	Adrien-Marie Legendre is the first of several who independently develop the method of least squares to make more precise estimates of population parameters.
1810	The Central Limit Theorem is introduced by Pierre-Simon Laplace.
1835	Adolphe Quetelet describes human traits as normally distributed, varying around a mean.
1845	Horace Mann advocates standardized essay testing for Massachusetts schools.
1852	Massachusetts enacts the first compulsory education law in the United States.
1857	The first teachers' union, the National Teachers Association (NTA) (now the National Education Association) is launched.
1866	New York Regents Examination Program begins.
1879	Sir Francis Galton uses the term *psychometric* to describe psycho-physical research.
1879	Wilhelm Wundt establishes the first psychology lab in Germany, which leads to the idea that psychology and related social sciences, such as education, could be studied as a science.
1896	The U.S. Supreme Court rules in *Plessy v. Ferguson*, upholding the doctrine of "separate but equal" that justified racial segregation in facilities that included schools.
1897	James Rice uses spelling tests in a research study. He uses these scores to compare the teaching of spelling across schools and presents what is considered the first educational program evaluation report.
1899	William James publishes "Talks to Teachers."
1900	The College Board is founded to develop college entrance exams.
1904	Edward Thorndike publishes *An Introduction to the Theory of Mental and Social Measurements*.
1904	Charles Spearman develops factor analysis, *g* theory, and classical test theory.
1905	The first standardized intelligence test, the Binet-Simon Intelligence Test, is introduced to identify students in French schools with intellectual disabilities.
1908	William Sealy Gosset, while working for a brewery, develops a simple method for estimating population means using small samples. Ronald Fisher later refined this formula into the *t* test.
1913	John Watson publishes an article introducing what he called behaviorism and suggests that as a social science, psychology should be studied objectively through observation of behavior, both human and animal.
1913	Henry Goddard begins using intelligence testing on immigrants at Ellis Island.
1914	Frederick Kelly introduces the Kansas Silent Reading Test, the first test to use the multiple-choice test format to reduce subjectivity in teacher scoring and grading.
1916	An American version of the Binet-Simon Intelligence Test, the Stanford-Binet, is introduced.
1916	The American Educational Research Association (AERA) is founded.
1916	Boston schools are evaluated based on locally developed tests.
1917	The United States enters World War I and the Army Alpha intelligence test is introduced to assess large numbers of enlistees in groups.
1918	Education is compulsory in all US states.
1921	The Psychological Corporation is founded to produce standardized testing materials.
1924	Ronald Fisher publishes what becomes known as the *F* distribution and begins developing the experimental design that allows for analysis of variance.
1925	The book *Statistical Methods for Research Workers* by Ronald Fisher is published. Among the statistical concepts introduced in the book are meta-analysis, using .05 as the level of significance, and using 1.96 standard deviations or standard errors to produce confidence intervals.
1926	The Scholastic Aptitude Test (SAT) for admission into higher education is first introduced. The test would later be known as the Scholastic Achievement Test and finally just as the SAT.
1928	Louis Thurstone publishes "Attitudes Can Be Measured" in the *American Journal of Sociology,* introducing an interval-level method for measuring attitude.

Date	*Event*
1932	Rensis Likert publishes “A Technique for the Measurement of Attitudes” in the Archives of Psychology, introducing an attitude assessment procedure that evolved into the very popular “Strongly Disagree to Strongly Agree” scaling format.
1932	Ralph Tyler suggests that evaluation should be objectives oriented. This remains the predominant approach.
1935	The Psychometric Society is formed.
1935	Ronald Fisher publishes *The Design of Experiments*, the first textbook on experimental design.
1938	The National Association of Teachers of Educational Measurement (now The National Council on Measurement in Education) is formed.
1942	Edward Guthrie publishes *A Theory of Contiguity* which suggests learning could not be explained simply by the idea of conditioning but that the events surrounding stimuli and responses were also important.
1942	Ralph Tyler publishes the “General Statement on Evaluation” in the *Journal of Educational Research*, describing the activities of evaluators.
1946	The American Institutes for Research (AIR) is founded.
1949	The Graduate Research Examination (GRE) is first used for admissions to master’s and doctoral degree programs.
1950	Abraham Wald publishes the book *Statistical Decision Functions*.
1954	The U.S. Supreme Court rules in *Brown v. Board of Education*. It reverses the 1896 decision in *Plessy v. Ferguson*, ruling that separate is not equal and outlawing segregation in schools.
1954	The National Council for Accreditation of Teacher Education (NCATE) is founded.
1956	Benjamin Bloom and others present a taxonomy of educational objectives that places learning objectives onto a six-level hierarchy ranging from low level memorized knowledge to high level understanding at the evaluation level.
1959	The ACT test is introduced. The organization that produces the test, American College Testing, is now known simply as ACT, Inc.
1960	Jerome Bruner publishes *The Process of Education,* which explores the role of structure in learning.
1960	Georg Rasch introduces what became known as the Rasch model, a measurement approach that weighs test items based partly on item difficulty.
1961	Albert Bandura publishes the first of his Bobo doll studies suggesting children learn vicariously through watching adult behaviors.
1962	Banesh Hoffman publishes his harsh criticism of standardized multiple-choice high stakes tests, *Tyranny of Testing*.
1962	Thomas Kuhn publishes *The Structure of Scientific Revolutions*, which introduces the ideas of research paradigms and paradigm shifts.
1962	Lev Vygostky’s work is first published in English. His idea of a *zone of proximal development*, referring to the amount of learning possible when guided by an adult or collaborating with “more capable peers,” eventually enters the mainstream.
1963	Robert Glaser uses the term *criterion-referenced* tests in an article about measuring learning outcomes. Criterion-referenced tests, as opposed to norm-referenced tests, produce scores that are meant to be interpreted against a set of performance standards.
1963	Stanley Milgram begins his *obedience to authority* studies in which subjects were instructed to give electrical shocks to other subjects. Controversy about the effects on subjects who followed the instructions informs modern human subject protection ethics.
1965	The Elementary and Secondary Education Act (ESEA) is signed into law by President Lyndon B. Johnson. It mandates equal access to education, high standards, and accountability.
1965	The Higher Education Act of 1965 is passed by Congress and signed into law by President Johnson.
1966	Center for Research on Evaluation, Standards, and Student Testing (CRESST) is founded.
1966	Donald Campbell and Julian Stanley distinguish between *experimental* designs that have random assignment to groups and *quasi-experimental* designs that do not.
1967	Michael Scriven distinguishes between summative evaluation (making a judgment about the effectiveness of a program after it is completed) and formative evaluation (providing feedback during the time a program is still underway).
1968	The Bilingual Education Act of 1968 is enacted, offering federal aid to local school districts in assisting them to address the needs of children with limited English-speaking ability.
1968	Benjamin Bloom uses Michael Scriven’s terms of *summative* and *formative* evaluation to describe summative and formative assessment.
1972	Carol Weiss publishes the book, *Evaluation Research*.
1972	Title IX of the Education Amendments of 1972 prohibits discrimination based on sex. It legislates equal access, opportunity, and treatment for female students in education, including in school athletics.
1973	George Box and George Tiao publish the book *Bayesian Inference in Statistical Analysis*, which reintroduces Bayesian analysis to the mainstream.
1975	Congress first enacted the Education for All Handicapped Children Act (Public Law 94-142), which was reauthorized in 1990 as the Individuals with Disabilities Education Act (IDEA) and in 2004 as the Individuals with Disabilities Education Improvement Act.
1975	The Joint Committee on Standards for Educational Evaluation (JCSEE) is created.
1983	Daniel Stufflebeam publishes a journal article presenting the CIPP model of program evaluation, which frames evaluations by context, inputs, process, and products.

Date	*Event*
1983	George Madaus, Michael Scriven, and Daniel Stufflebeam publish *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*.
1983	President Ronald Reagan's National Commission on Excellence in Education issues *A Nation at Risk*. The report sounds an alarm about what it indicates is severe underperformance of American schools compared to schools elsewhere.
1985	Yvonna Lincoln and Egon Guba publish the book *Naturalistic Inquiry*, which presents a qualitative framework for social science research and provides an alternative to the quantitative paradigm.
1986	The American Evaluation Association is formed from the merger of the Evaluation Research Society and Evaluation Network.
1987	Blaine Worthen and James Sanders publish *Educational Evaluation: Alternative Approaches and Practical Guidelines*.
1988	Robert Sternberg publishes *The Triarchic Mind* which defines intelligence as cognitive self-management.
1993	The Massachusetts Education Reform Act requires a common curriculum and statewide tests. Other states follow Massachusetts' lead and implement similar high-stakes testing programs.
1994	The Goals 2000: Educate America Act is signed into law by President Bill Clinton.
1994	The Improving America's Schools Act (IASA) is signed into law by President Bill Clinton, reauthorizing the ESEA of 1965.
1994	A chapter by Anselm Strauss and Juliet Corbin on *grounded theory* is included in the *Handbook of Qualitative Research*. Grounded theory emphasizes building theory that is faithful to the data.
1998	The Higher Education Act is amended and reauthorized, requiring institutions and states to produce "report cards" about teacher education.
2000	Diane Ravitch publishes the book, *Left Back: A Century of Failed School Reforms*, which criticizes progressive educational policies and argues for a more traditional, academically oriented education.
2002	The No Child Left Behind Act (NCLB) is signed into law by President George W. Bush. The reauthorization of the ESEA required schools and school districts to meet targets on annual statewide standardized tests and called for all students to be proficient in reading and math by 2014.
2004	The Individuals with Disabilities Education Act is reauthorized as the Individuals with Disabilities Education Improvement Act. The new law requires school districts to use the response to intervention (RTI) approach with the overall goal of reducing the need for special education services.
2004	*Ready or Not: Creating a High School Diploma That Counts*, the culminating report of the American Diploma Project (ADP), highlights discrepancies between employers' and colleges' academic demands for high school students and states' high school graduation expectations.
2009	The Common Core State Standards Initiative (CCSSI) coordinated by the National Governors Association and the Council of Chief State School Officers (CCSSO) is launched. By 2012, all but five states had adopted the Common Core standards in both English and math.
2009	Race to the Top initiative (RTTT), a $4.35 billion program designed to induce K–12 education reform is enacted as part of the American Reinvestment and Recovery Act.
2010	Partnership for Assessment of Readiness for College and Careers (PARCC) is formed to create and deploy a standard set of K–12 assessments in mathematics and English based on the Common Core State Standards.
2012	The Next Generation Science Standards are released.
2013	The Council for the Accreditation of Educator Preparation (CAEP) is formed from the consolidation of NCATE and TEAC.
2015	The Every Student Succeeds Act is passed by Congress and signed into law by President Barack Obama, replacing NCLB. ESSA gave states more flexibility than NCLB in setting targets for school and district performance.
2017	For the first time the National Assessment of Educational Progress (NAEP) begins administering digitally based assessments.

Index

NOTE: Page references to figures and tables are labelled as (fig.) and (table).