MULTIVARIATE APPLICATIONS SERIES



# Handbook of Ethics in Quantitative Methodology

A. T. Panter • Sonya K. Sterba

# Handbook of Ethics in Quantitative Methodology

**Multivariate Applications Series**

Sponsored by the Society of Multivariate Experimental Psychology, the goal of this series is to apply complex statistical methods to significant social or behavioral issues, in such a way so as to be accessible to a nontechnical readership (e.g., non-methodological researchers, teachers, students, government personnel, practitioners, and other professionals). Applications from a variety of disciplines such as psychology, public health, sociology, education, and business are welcome. Books can be single- or multiple-authored or edited volumes that (1) demonstrate the application of a variety of multivariate methods to a single, major area of research; (2) describe a multivariate procedure or framework that could be applied to a number of research areas; or (3) present a variety of perspectives on a topic of interest to applied multivariate researchers.

There are currently 17 books in the series:

- *What if there were no significance tests?* co-edited by Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger (1997)
- *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*, written by Barbara M. Byrne (1998)
- *Multivariate Applications in Substance Use Research: New Methods for New Questions*, co-edited by Jennifer S. Rose, Laurie Chassin, Clark C. Presson, and Steven J. Sherman (2000)
- *Item Response Theory for Psychologists*, co-authored by Susan E. Embretson and Steven P. Reise (2000)
- *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*, written by Barbara M. Byrne (2001)
- *Conducting Meta-Analysis Using SAS*, written by Winfred Arthur, Jr., Winston Bennett, Jr., and Allen I. Huffcutt (2001)
- *Modeling Intraindividual Variability with Repeated Measures Data: Methods and Applications*, co-edited by D. S. Moskowitz and Scott L. Hershberger (2002)
- *Multilevel Modeling: Methodological Advances, Issues, and Applications*, co-edited by Steven P. Reise and Naihua Duan (2003)
- *The Essence of Multivariate Thinking: Basic Themes and Methods*, written by Lisa Harlow (2005)
- *Contemporary Psychometrics: A Festschrift for Roderick P. McDonald*, co-edited by Albert Maydeu-Olivares and John J. McArdle (2005)
- *Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming*, Second Edition, written by Barbara M. Byrne (2006)

- *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology,* co-edited by Niels G. Waller, Leslie J. Yonce, William M. Grove, David Faust, and Mark F. Lenzenweger (2006)
- *Introduction to Statistical Mediation Analysis,* written by David P. MacKinnon (2008)
- *Applied Data Analytic Techniques for Turning Points Research,* edited by Patricia Cohen (2008)
- *Cognitive Assessment: An Introduction to the Rule Space Method,* written by Kikumi K. Tatsuoka (2009)
- *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Second Edition*, written by Barbara M. Byrne (2010)
- *Handbook of Ethics in Quantitative Methodology,* co-edited by A. T. Panter and Sonya K. Sterba (2011)

Anyone wishing to submit a book proposal should send the following: (1) author/title; (2) timeline including completion date; (3) brief overview of the book's focus, including table of contents and, ideally, a sample chapter (or chapters); (4) a brief description of competing publications; and (5) targeted audiences.

For more information, please contact the series editor, Lisa Harlow, at Department of Psychology, University of Rhode Island, 10 Chafee Road, Suite 8, Kingston, RI 02881-0808; phone (401) 874-4242; fax (401) 874-5562; or e-mail LHarlow@uri.edu. Information may also be obtained from members of the editorial/advisory board: Leona Aiken (Arizona State University), Daniel Bauer (University of North Carolina), Jeremy Biesanz (University of British Columbia), Gwyneth Boodoo (Educational Testing Services), Barbara M. Byrne (University of Ottawa), Scott E. Maxwell (University of Notre Dame), Liora Schmelkin (Hofstra University), and Stephen West (Arizona State University).

# Handbook of Ethics in Quantitative Methodology

**A.T. Panter**
The University of North Carolina, Chapel Hill

**Sonya K. Sterba**
Vanderbilt University

Routledge
Taylor & Francis Group
New York London

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
27 Church Road
Hove, East Sussex BN3 2FA

© 2011 by Taylor and Francis Group, LLC
Routledge is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-84872-854-7 (Hardback) 978-1-84872-855-4 (Paperback)

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

**Library of Congress Cataloging-in-Publication Data**

Panter, A. T.
Handbook of ethics in quantitative methodology / A.T. Panter, Sonya K. Sterba.
p. cm. -- (Multivariate applications series)
Includes bibliographical references and index.
ISBN 978-1-84872-854-7 (hbk. : alk. paper) -- ISBN 978-1-84872-855-4 (pbk. : alk. paper)
1. Quantitative research--Moral and ethical aspects. 2. Social sciences--Methodology--Moral and ethical aspects. I. Sterba, Sonya K. II. Title.

H62.P276 2011
174'.900142--dc22 2010045883

**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the Psychology Press Web site at**
**http://www.psypress.com**

*In memory of J. S. Tanaka – ATP*

*To my parents, Jim and Janet – SKS*

# *Contents*

Preface ..... xiii
Editors ..... xv
Contributors ..... xvii
Software Notice ..... xix

**1 Ethics in Quantitative Methodology: An Introduction ..... 1**
*A. T. Panter and Sonya K. Sterba*

## Section I Developing an Ethical Framework for Methodologists

**2 Ethics in Quantitative Professional Practice ..... 15**
*John S. Gardenier*

**3 Ethical Principles in Data Analysis: An Overview ..... 37**
*Ralph L. Rosnow and Robert Rosenthal*

## Section II Teaching Quantitative Ethics

**4 A Statistical Guide for the Ethically Perplexed ..... 61**
*Lawrence Hubert and Howard Wainer*

## Section III Ethics and Research Design Issues

**5 Measurement Choices: Reliability, Validity, and Generalizability ..... 127**
*Madeline M. Carrig and Rick H. Hoyle*

**6 Ethics and Sample Size Planning ..... 159**
*Scott E. Maxwell and Ken Kelley*

**7 Ethics and the Conduct of Randomized Experiments and Quasi-Experiments in Field Settings ..... 185**
*Melvin M. Mark and Aurora L. Lenz-Watson*

8 **Psychometric Methods and High-Stakes Assessment: Contexts and Methods for Ethical Testing Practice** ..... 211
*Gregory J. Cizek and Sharyn L. Rosenberg*

9 **Ethics in Program Evaluation** ..... 241
*Laura C. Leviton*

## Section IV Ethics and Data Analysis Issues

10 **Beyond Treating Complex Sampling Designs as Simple Random Samples: Data Analysis and Reporting** ..... 267
*Sonya K. Sterba, Sharon L. Christ, Mitchell J. Prinstein, and Matthew K. Nock*

11 **From Hypothesis Testing to Parameter Estimation: An Example of Evidence-Based Practice in Statistics** ..... 293
*Geoff Cumming and Fiona Fidler*

12 **Some Ethical Issues in Factor Analysis** ..... 313
*John J. McArdle*

13 **Ethical Aspects of Multilevel Modeling** ..... 341
*Harvey Goldstein*

14 **The Impact of Missing Data on the Ethical Quality of a Research Study** ..... 357
*Craig K. Enders and Amanda C. Gottschall*

15 **The Science and Ethics of Causal Modeling** ..... 383
*Judea Pearl*

## Section V Ethics and Communicating Findings

16 **Ethical Issues in the Conduct and Reporting of Meta-Analysis** ..... 417
*Harris Cooper and Amy Dent*

17 **Ethics and Statistical Reform: Lessons From Medicine** ..... 445
*Fiona Fidler*

**18 Ethical Issues in Professional Research, Writing, and Publishing** ..... **463**
*Joel R. Levin*

**Author Index** ..... **493**

**Subject Index** ..... **505**

# *Preface*

This Handbook provides the only available (a) interdisciplinary effort to develop a cohesive ethical framework for quantitative methods; (b) comprehensive and current review of ethical issues interfacing with quantitative social science; (c) set of case examples illustrating these issues; and (d) synthesized, practical guidance on these issues. As granting agencies, professional organizations, and universities are progressively recommending or requiring ethical perspectives to be incorporated in all stages of the research process, we hope that the material covered in this book will become increasingly relevant to practice.

We designed the Handbook to be of use for at least three types of audiences. One intended audience includes psychology and behavioral sciences graduate students enrolled in a core quantitative methods course sequence who may also be at the beginning stages of their own data analyses and reporting. This volume could easily be recommended or supplemental reading for a doctoral statistics sequence, where particular chapters could serve as the basis for class discussion. Chapter 4 most directly targets this audience. Students enrolled in research ethics seminars would benefit from the two framework chapters (Chapters 2 and 3). These chapters would be particularly helpful in bridging the contents of such a general ethics course with the quantitatively oriented contents of this book.

Given the more advanced quantitative topics in the design, data analysis, and reporting sections (Chapters 5–18), we intended another primary audience for this Handbook to be journal editors, journal reviewers, and grant reviewers, who routinely evaluate high-level quantitative analyses but who may not have taken specific coursework in quantitative methods in some time. Chapter 18 tackles specific issues that confront this audience on a daily basis, whereas Chapter 17 speaks to more general themes of policy relevance to this audience.

Finally, we expect an additional audience to be researchers and professionals with quantitative interests who share our concern with broad philosophical questions that determine our collective approach to research design, sampling, measurement, data collection, attrition, modeling, reporting, and publishing. The bulk of the Handbook, Chapters 5–16, targets this audience. In these chapters the level of prerequisite knowledge is at or above what graduate students would learn from a first-year statistics sequence. Still, these chapters remain accessible to a broad range of researchers with normative basic quantitative knowledge. Finally, although case examples in this Handbook are primarily drawn

from psychology, this audience could easily span related fields, including public health, education, sociology, social work, political science, and business/marketing.

We are grateful to the contributors of this volume, who carefully and creatively considered their established research programs in relation to the ethical frame of this Handbook. We are also grateful to the three reviewers whose constructive feedback shaped and improved this book.

We thank Senior Acquisitions Editor Debra Riegert for her guidance during all stages of this process and Multivariate Applications Series Editor Dr. Lisa Harlow (University of Rhode Island) for her enthusiastic support of these ideas. Erin Flaherty provided editorial support throughout the process.

A. T. Panter thanks her colleagues at the L. L. Thurstone Psychometric Laboratory (University of North Carolina at Chapel Hill), Dr. Lyle V. Jones, and Dr. David Thissen for fully appreciating the high-stakes nature of quantitative action from design to reporting. She also is grateful to her family members, especially Dr. Gideon G. Panter, Danielle Panter, Michaela Panter, and Jonathan Panter, for constant inspiration on this topic, and to Dr. Sarajane Brittis for her open and invaluable guidance. Dr. George J. Huba, Nechama, and Yaakov provide their own special form of inspiration.

Sonya K. Sterba thanks Dr. Erica Wise (University of North Carolina at Chapel Hill) for encouraging her study of linkages between quantitative practice and ethics.

# *Editors*

**A. T. Panter** is the Bowman and Gordon Gray Distinguished Professor of Psychology at the L. L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill. She develops instruments, research designs, and data-analytic strategies for applied research questions in health (e.g., HIV/AIDS, mental health, cancer) and education. Her publications are in measurement and testing, advanced quantitative methods, survey methodology, program evaluation, and individual differences. She has received numerous teaching awards, including the Jacob Cohen Award for Distinguished Contributions to Teaching and Mentoring from APA's Division 5 (Evaluation, Measurement, & Statistics). She has significant national service in disability assessment, testing in higher education, women in science, and the advancement of quantitative psychology.

**Sonya K. Sterba** is an assistant professor in the quantitative methods and evaluation program at Vanderbilt University. She received her PhD in quantitative psychology and her MA in child clinical psychology from the University of North Carolina at Chapel Hill. Her research evaluates how traditional structural equation and multilevel models can be adapted to handle methodological issues that commonly arise in developmental psychopathology research.

# *Contributors*

**Madeline M. Carrig**
Duke University
Durham, North Carolina

**Sharon L. Christ**
Purdue University
West Lafayette, Indiana

**Gregory J. Cizek**
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina

**Harris Cooper**
Duke University
Durham, North Carolina

**Geoff Cumming**
La Trobe University
Melbourne, Australia

**Amy Dent**
Duke University
Durham, North Carolina

**Craig K. Enders**
Arizona State University
Tempe, Arizona

**Fiona Fidler**
La Trobe University
Melbourne, Australia

**John S. Gardenier**
Emeritus, Centers for Disease Control and Analysis
National Center for Health Statistics
Vienna, Virginia

**Harvey Goldstein**
University of Bristol
Bristol, United Kingdom

**Amanda C. Gottschall**
Arizona State University
Tempe, Arizona

**Rick H. Hoyle**
Duke University
Durham, North Carolina

**Lawrence Hubert**
University of Illinois at Urbana-Champaign
Champaign, Illinois

**Ken Kelley**
University of Notre Dame
Notre Dame, Indiana

**Aurora L. Lenz-Watson**
The Pennsylvania State University
University Park, Pennsylvania

**Joel R. Levin**
University of Arizona
Tucson, Arizona

**Laura C. Leviton**
The Robert Wood Johnson Foundation
Princeton, New Jersey

**Melvin M. Mark**
The Pennsylvania State University
University Park, Pennsylvania

**Scott E. Maxwell**
University of Notre Dame
Notre Dame, Indiana

**John J. McArdle**
University of Southern California
Los Angeles, California

**Matthew K. Nock**
Harvard University
Cambridge, Massachusetts

**A. T. Panter**
University of North Carolina
at Chapel Hill
Chapel Hill, North Carolina

**Judea Pearl**
University of California, Los
Angeles
Los Angeles, California

**Mitchell J. Prinstein**
University of North Carolina
at Chapel Hill
Chapel Hill, North Carolina

**Sharyn L. Rosenberg**
American Institutes for Research
Washington, DC

**Robert Rosenthal**
University of California, Riverside
Riverside, California

**Ralph L. Rosnow**
Emeritus, Temple University
Radnor, Pennsylvania

**Sonya K. Sterba**
Vanderbilt University
Nashville, Tennessee

**Howard Wainer**
National Board of Medical
Examiners
Philadelphia, Pennsylvania

# *Software Notice*

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098 USA
Telephone: (508) 647-7000
Fax: (508) 647-7001
E-mail: info@mathworks.com
Web: http://www.mathworks.com

# 1

# *Ethics in Quantitative Methodology: An Introduction*

**A. T. Panter**
*University of North Carolina at Chapel Hill*

**Sonya K. Sterba**
*Vanderbilt University*

Social science researchers receive guidance about making sound methodological decisions from diverse sources throughout their careers: hands-on research experiences during graduate school and beyond; a small set of graduate quantitative or research methods courses (Aiken, West, & Millsap, 2008); occasional methodological presentations, proseminars, and workshops; professional reading; and participation in the peer review process. Even in these more structured contexts, the connections between design/analysis decisions and research ethics are largely implicit and informal. That is, methodology professors and research advisors convey ethical principles about research design, data analysis, and reporting experientially and by modeling appropriate professional behavior—without necessarily being cognizant that they are doing so and without labeling their behavior as such.

Some graduate programs and federal grant mechanisms require that students take a formal semester-long research ethics course. However, quantitative topics are largely absent from such ethics curricula. To obtain further knowledge, some researchers turn to professional organizations' ethics codes (e.g., American Psychological Association [APA] Code of Conduct, 2002) or to short online ethics training modules that are sanctioned by federal granting agencies and university institutional review boards (IRBs). However, these ethical principles and training materials also are limited and indirect in their guidance about specific areas of concern to quantitative methodologists (e.g., measurement, sampling, research design, model selection and fitting, reporting/reviewing, and evaluation).

Thus, a researcher has access to widely available ethical standards that are quantitatively vague and widely available quantitative standards that

are disseminated without an ethical imperative. It is not surprising then that there has been little explicit linkage between ethics and methodological practice in the social sciences to date.

A problem with this separation between ethics and methods is that published methodological guidelines for best practice in, for example, quantitative psychology, routinely go unheeded. It is well known that accessible statistical guidance given without an ethical imperative (e.g., Wilkinson & the Task Force on Statistical Inference, 1999) is painfully slow to infiltrate applied practice. We believe that the lack of attention to methodological guidelines may be because these guidelines lack an ethical imperative to motivate change, such as the identification of human costs related to quantitative decision making throughout design, data analysis, and reporting. Quantitative psychologists routinely lament that antiquated analytic methods continue to be used in place of accessible, preferred methods (e.g., Aiken et al., 2008; Cohen, 1994; Harlow, Mulaik, & Steiger, 1997; Hoyle & Panter, 1995; MacCallum, Roznowski, & Necowitz, 1992; MacCallum, Wegener, Uchino, & Fabrigar, 1993; Maxwell, 2000, 2004; Schafer & Graham, 2002; Steiger, 2001). We submit that quantitative-specific ethical guidance is valuable and necessary to motivate changes in quantitative practice for the social sciences (as has been already observed in the medical field; see Fidler, Chapter 17, this volume).

If establishing an ethics–methods linkage might be expected to have positive consequences for the quality of research design, analysis, and reporting in the social sciences, the obvious next question is: How might such an ethics–methods linkage be established? We began exploring this issue during periodic sessions of our lunchtime Quantitative Psychology Forum at the University of North Carolina at Chapel Hill. We considered topics such as how quantitative methodology is reflected in the APA ethics code and the place of null hypothesis significance testing in modern day quantitative approaches. We also considered dilemmas ranging from cost–benefit tradeoffs of increasingly extensive data retention and archiving practices to balancing pressures from consultees or reviewers to conduct simple analyses when more complex, statistically appropriate analyses are available. We eventually broadened the scope of the discussion through a symposium on "Quantitative Methodology Viewed through an Ethical Lens" presented to Division 5 (Evaluation, Measurement, & Statistics) at the 2008 APA convention (Panter & Sterba, 2008).

Ultimately, we realized that it was necessary to convene an even broader set of methodological experts to address issues along the quantitative research continuum (from research design and analysis through reporting, publishing, and reviewing). Such methodologists are ideally positioned to weigh in on whether and when decision making in their area of expertise should be recast in an ethical light. Because our selection of methodological issues could not possibly encompass all those that might

be encountered in practice, we believed it was critical to also hear from several ethics experts. These experts can offer overarching ethical frameworks that practitioners might use to guide research decisions on new and different methodological topics, independent of the included topics. This Handbook compiles the perspectives of these methodological experts and ethics experts into a unified source. Through our selected authors, we have sought broad coverage of the possible areas in quantitative psychology where ethical issues may apply. It is our hope that the consistent use of an ethical frame and the emphasis on human cost of quantitative decision making will encourage researchers to consider the multifaceted implications of their methodological choices. We also believe that this ethical frame can often provide support for methodological choices that optimize state-of-the-art approaches.

The remainder of this introductory chapter is structured as follows. First, we further develop the motivation for the Handbook. Specifically, we review the extent to which available research ethics resources for social scientists can be practically used to aid decision making involving data analysis and study design. Further, we review the extent to which available methodological teaching materials establish linkages with ethics, as well as the extent to which ethics teaching materials establish linkages with methods. We briefly consider whether the degree of intersection between ethics and methods in these textbooks is sufficient to provide an ethical imperative to guide decision making on the variety of quantitative issues that routinely confront researchers. On finding neither research ethics resources nor available textbooks up to this task, we turn to the goals and structure of this Handbook.

## Lack of Quantitative Specifics in Formal Ethics Resources

Our focus on ethical issues in quantitative psychology mirrors an emergent emphasis over the past several decades. The U.S. federal government has advocated the explicit interjection of ethical thought into all stages of scientific research, not only in the form of ethical standards and guidelines but also the form of education in "Responsible Conduct of Research" (RCR) (Code of Federal Regulations, 2009; Department of Health and Human Services, Commission on Research Integrity, 1995; Institute of Medicine, 1989; National Academy of Sciences Panel on Scientific Responsibility and the Conduct of Research, 1992; Office of Science and Technology Policy, 2000). Recently, the federal government has also begun soliciting large-scale primary research on "methodology and measurement in the behavioral and social sciences" and "issues of ethics in research" (e.g.,

National Institutes of Health, 2008). The APA has responded by implementing ethical guidelines (Ethical Principles of Psychologists and Code of Conduct, 2002), as well as RCR workshops and educational materials (e.g., see http://www.apa.org/research/responsible/index.aspx).

However, the sections of the Code and RCR training materials devoted to research provide little guidance and discussion in the way of specific design, analytic, and reporting guidance for choices confronting quantitative psychology researchers, consultants, and consumers (reviewed in Sterba, 2006). For example, the general directives of the APA code are to maintain competence (Section 2.03), not to fabricate data (Section 8.10), to "document … professional and scientific work … [to] allow for replication of research design and analyses" (Section 6.01), to share research data so that others can "verify the substantive claims through reanalysis" (Section 8.14), to "use assessment instruments whose validity and reliability have been established for use with members of the population tested" (Section 9.02), and to construct tests using "appropriate psychometric procedures and current scientific or professional knowledge for test design, standardization, validation, [and] reduction or elimination of bias" (Section 9.05). It is left to the discretion of the psychologist to determine what quantitative competencies are needed, how to define and present evidence for reliability and validity, what comprises an adequate tested population, what constitutes current scientific knowledge for eliminating biases, and so forth. Furthermore, although APA's RCR training materials include many sections that are highly relevant to quantitative methodologists (e.g., collaborative science, conflicts of interest, data acquisition and sharing, human protections, lab animal welfare, mentoring, peer review, responsible authorship, and research misconduct), they once again lack practical guidance on research design, analysis/modeling, and reporting of results.

In sum, research ethics resources sanctioned by federal granting agencies or professional societies simply lack the specificity to guide day-to-day decision making on difficult methodological topics. We next review whether available textbooks on research methods contain sufficient ethics training to provide such guidance and/or whether available research ethics textbooks provide sufficient methods training to offer such guidance.

## Lack of Cross-Pollination Between Methods and Ethics Textbooks

Students can be reasonably expected to integrate their thinking on ethics and methods to the extent that their textbooks do so. However, to this end,

existing textbooks and edited volumes that combine approaches to psychology research methods and ethics have only integrated the two topics to a minimal extent. For example, some textbooks about research methods include a single ethics chapter, and some textbooks about ethics include a single research methods chapter.

First, consider the content of single ethics chapters contained within psychology research methods texts. Some of these ethics chapters provide a general discussion of concepts and research-related principles in the most recent APA code, sometimes including case study examples (Dunn, Smith, & Beins, 2007; Kazdin, 2003) and sometimes including historical and philosophical background (Davis, 2003; Kimmel, 2003; Rosenthal & Rosnow, 2008). Other research methods texts focus specifically on the ethical treatment of human and animal study participants (Breakwell, Hammond, & Fife-Schaw, 2000; Davis, 2003; Haslam & McGarty, 2003). Although a useful starting point, these chapters largely lack attention to ethical aspects of specific quantitative topics (the discussion of Rosenthal and Rosnow, 2008, on causality, outlier detection, and multiple comparisons is a major exception). In addition, these chapters never devote more than a page or two to a specific topic.

Second, consider the content of single methods chapters contained within psychology research ethics texts. These methods chapters exclusively pertain either to assessment (Bersoff, 1999; Ethics Committee, 1987; Fisher, 2003; Koocher & Keith-Spiegel, 1998; Steininger, Newell, & Garcia, 1984) or to recruitment (Kimmel, 2007; Sales & Folkman, 2000). But their focus is narrower still, as within either topic, ethical implications are mainly mentioned when design or analysis decisions interface *directly* with human (or animal) participants. Design and analysis decisions have ethical implications when they *indirectly* affect human (or animal) welfare as well, through their influence on how federal grants are allocated, what future research is pursued, and what policies or treatments are adopted. For example, ethics texts whose methodological chapter concerned recruitment largely focused on direct human subject concerns such as vulnerability of participants recruited from captive groups and the effects of incentives. Concerns about the effects of recruitment strategies on the validity of statistical inferences were not discussed, nor were methods for accounting for selection bias at the modeling stage and methods for preventing selection bias at the design stage. These additional recruitment choices minimally have indirect ethical implications, if not direct ethical implications, as well (see Sterba, Christ, Prinstein, & Nock, Chapter 10, this volume). As another example, ethics texts whose methodological chapter concerned assessment primarily discussed competence of the test administrator and interpreter, and conditions for disclosure of test data to human participants (Bersoff, 1999)—possibly accompanied by case examples on these topics (Ethics Committee, 1987; Fisher, 2003;

Koocher & Keith-Spiegel, 1998). However, these sources did not devote more than an elementary definitional overview of psychometric topics such as test validity, reliability, standardization, and bias (Fisher, 2003; Koocher & Keith-Spiegel, 1998; Steininger et al., 1984). Competency regarding the latter topics minimally has indirect ethical implications, insofar as it encourages researchers to construct or choose measures appropriately at the design phase and then to apply and evaluate a statistical model at the analysis stage (see Carrig & Hoyle, Chapter 5, this volume; Cizek & Rosenberg, Chapter 8, this volume).

In sum, students and professionals could not be expected to internalize an ethical imperative guiding their day-to-day methodological decision making from a single ethics chapter in a methods text or a single methods chapter in an ethics text. To extend the material available in single ethics chapters of methods texts, this Handbook more evenly distributes emphasis between ethics and methods. For example, we have an initial section providing overarching theoretical frameworks with which to guide quantitative practice in general, followed by a series of chapters providing linkages between specific methodological topics (e.g., missing data, multilevel data structures, statistical power) and the ethical implications of alternative ways of addressing these topics. To extend the material already available in single methods chapters of ethics texts, this Handbook more broadly construes the kind of methodological issues and decisions that can potentially have ethical implications and discusses these issues in considerably greater depth.

## General Handbook Goals

The overarching goal of this Handbook is to achieve a shift in thinking among methodologists in social science similar to that which has already taken place among statisticians. Where the APA code and APA RCR materials currently stand with respect to ethics and quantitative psychology mirrors where the American Statistical Association (ASA) stood with respect to ethics and statistics in 1983. The 1983 ASA code (Ad Hoc Committee on Professional Ethics, 1983) was deemed purely aspirational without being educational, in that it was backed up neither by case studies nor practical guides to action (Gardenier, 1996). However, in 1983, statisticians opened a dialogue in *American Statistician* with 16 expert commentaries published on ethics and statistics. Statisticians concluded that an ethical imperative was helpful to guide individual and collective statistical practice by (a) documenting and disseminating changing norms of appropriate practice; (b) exposing "inherent conflicts" while providing

tactics for their resolution; and (c) assisting consultants "in maintaining ethical standards in confronting employers, sponsors, or clients in specific instances" by enabling them to reference a source that provides a "considered statement of appropriate behavior that has been accepted by the profession" (Martin, 1983, pp. 7–8; see also Gardenier, 1996; Mosteller, 1983; Seltzer, 2001). This discussion culminated in the 1999 revision of the ASA code, which included more specific, practical guidance, as well as accompanying case studies for handling specific situations (ASA, 1999).

We agree with the above aims (a) through (c). The purpose of this Handbook is to open up a similar dialogue among quantitative psychologists—not for the purpose of revising the APA code or RCR materials per se, but for the purpose of fulfilling the above three aims directly.

The Handbook chapters further these three aims in diverse ways. For example, chapter authors often draw a distinction between ethical matters and purely technical statistical controversy for a given topic (Seltzer, 2001). As another example, chapter authors provide ethically motivated methodological education on a given topic alongside concrete demonstrations and explanations—without resorting to overly literal standards that can be quickly outdated and/or appear doctrinaire.

## General Structure of the Handbook

The Handbook focuses on articulating and then illustrating ethical frames that can inform decision making in the research process. The first section of this Handbook is devoted to developing and disseminating two proposed ethical frameworks that cross-cut design, analysis, and modeling. One framework is supplied by the former chair of the ASA board responsible for the 1999 ethics code (Gardenier, Chapter 2, this volume), and the other framework is supplied by two longtime contributors to ethical theory as it interfaces with statistical practice in psychology and the behavioral sciences (Rosnow & Rosenthal, Chapter 3, this volume).

Our next section focuses on teaching the next generation of quantitative students. Hubert and Wainer (Chapter 4, this volume) consider ways to connect these ethical principles to statistics in the classroom. They provide a diverse assortment of pedagogical strategies for disseminating ethical training on methods in graduate and undergraduate statistics courses.

The order of chapters in the remaining three sections is intended to mirror the chronology of the research process starting with research design and data collection, moving to data analysis and modeling, and then concluding with communicating findings to others. Each chapter provides a brief introduction to its particular methodological topic and then

identifies prevailing methodological issues, misunderstandings, pitfalls, or controversies, and their potential direct or indirect ethical aspects or implications. Each chapter includes concrete example(s) and application(s) of an ethical imperative in deciding among approaches and solutions for a design or analysis problem within the purview of that methodological topic. These concrete examples often concern topics at the center of high-stakes scientific or policy controversies (e.g., value-added educational assessments in Goldstein's Chapter 13, this volume; employment testing in Fidler's Chapter 17, this volume; evaluation of the Cash and Counseling Program for disabled adults in Leviton's Chapter 9, this volume; and the implicit association test for detecting racial, age-related, or other biases in Carrig & Hoyle's Chapter 5, this volume).

In the ethics and research design section (Section 3), the chapter authors consider ethical aspects of selecting and applying behavioral measurement instruments (Carrig & Hoyle, Chapter 5, this volume) and consider ethical implications of alternative approaches to sample size planning that are, for example, specific to detecting the existence versus magnitude of effects (Maxwell & Kelly, Chapter 6, this volume). Section 3 chapter authors also compare the defensibility of experimental versus quasi-experimental research designs in various contexts (Mark & Lenz-Watson, Chapter 7, this volume) and consider potential ethical issues that could arise when designing high-stakes tests (Cizek & Rosenberg, Chapter 8, this volume), in the production of program evaluations, and in negotiations with stakeholders (Leviton, Chapter 9, this volume). These topics frame and subsequently affect the entire research process, including data collection activities, analyses, and reporting. Uninformed choices in these areas can lead to significant costs for participants, researchers, and taxpayers and, importantly, deny benefits to those whose health, social, occupational, emotional, or behavioral outcomes might otherwise have been aided by the research findings.

The chapters in the data analysis section (Section 4) directly address key decision points that shape researchers' approach to data analysis and the models that they might evaluate. These analytic decisions drive how data may be interpreted by the scientific community and ultimately how the findings are implemented in subsequent treatments, policy decision making, and future research planning in the field. These chapters discuss when and why analysts need to statistically account for the manner in which the sample was selected (Sterba et al., Chapter 10, this volume), how analysts can evaluate tradeoffs of hypothesis testing versus estimation frameworks (Cumming & Fidler, Chapter 11, this volume), and how analysts can increase awareness and transparency about approaching modeling from exploratory versus confirmatory frames of reference (McArdle, Chapter 12, this volume). Chapters in Section 4 also cover ethical issues arising when nested data structures are present but are not captured

by simple multilevel models (Goldstein, Chapter 13, this volume), when choosing among available methods for handling missing data under a variety of real world contexts (Enders & Gottschall, Chapter 14, this volume), and when causal inferences are desired from the statistical analysis (Pearl, Chapter 15, this volume).

Finally, our last section (Section 5) comprises chapters that emphasize reporting. Cooper and Dent (Chapter 16, this volume) review how experts in meta-analysis suggest that findings should be reported and the ethical implications of such reporting decisions. Fidler (Chapter 17, this volume) reflects on how widespread ethically motivated changes in analysis and reporting practices in the medical field have advanced medical science and practice, and what this means for social scientists. Finally, Levin (Chapter 18, this volume) evaluates a range of ethical dilemmas that emerge during the publication process from the perspective of a long-term journal editor.

## Conclusion

In sum, this Handbook provides a unique resource for applied journal editors who often lack methodological reviewers, research methods instructors who often lack formal ethics training, research ethics instructors who often lack formal methodological training, and granting agency project officers and IRB members who may lack training in one field or the other. This Handbook will be useful for faculty and graduate student statistical consultants who need to educate themselves and/or their clients on ethical practices involving difficult methodological issues. Finally, we hope this Handbook will serve as an impetus for informal (e.g., departmental brown bags) and formal consideration of how we collectively can better link ethical imperatives with quantitative practice.

## References

Ad Hoc Committee on Professional Ethics. (1983). Ethical guidelines for statistical practice. *The American Statistician, 37,* 5–6.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

American Statistical Association. (1999). *Ethical guidelines for statistical practice*. Retrieved from http://www.amstat.org/about/ethicalguidelines.cfm

Bersoff, D. N. (1999). *Ethical conflicts in psychology* (2nd ed.). Washington, DC: American Psychological Association.

Breakwell, G. M., Hammond, S., & Fife-Schaw, C. (2000). *Research methods in psychology* (2nd ed.). London: Sage.

Code of Federal Regulations, Title 45, Chapter 46. (2009). Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Davis, S. F. (2003). *Handbook of research methods in experimental psychology.* Malden, MA: Blackwell.

Department of Health and Human Services, Commission on Research Integrity. (1995). Integrity and misconduct in research. Washington, DC: U.S. Department of Health and Human Services.

Dunn, D. S., Smith, R. A., & Beins, B. C. (2007). *Best practices for teaching statistics and research methods in the behavioral sciences.* Mahwah, NJ: Erlbaum.

Ethics Committee. (1987). *Casebook on ethical principles of psychologists.* Washington, DC: American Psychological Association.

Fisher, C. B. (2003). *Decoding the ethics code: A practical guide for psychologists.* London: Sage.

Gardenier, J. (1996). What and where are statistical ethics? In *Proceedings of the Section on Statistical Education, 1996* (pp. 256–260). Alexandria, VA: American Statistical Association.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Haslam, A., & McGarty, C. (2003). *Research methods and statistics in psychology*. London: Sage.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Institute of Medicine. (1989). *The responsible conduct of research in the health sciences*. Washington, DC: National Academy Press.

Kazdin, A. E. (2003). *Methodological issues and strategies in clinical research* (3rd ed.). Washington, DC: American Psychological Association.

Kimmel, A. J. (2003). Ethical issues in social psychological research. In C. Sansone, C. Morf, & A. T. Panter (Eds.), *The Sage handbook of methods in social psychology* (pp. 45–70). Thousand Oaks, CA: Sage.

Kimmel, A. J. (2007). *Ethical issues in behavioral research: Basic and applied perspectives* (2nd ed.). Malden, MA: Blackwell.

Koocher, G. P., & Keith-Spiegel, P. (1998). *Ethics in psychology: Professional standards and cases.* New York: Oxford University Press USA.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199.

Martin, M. (1983). [Ethical Guidelines for Statistical Practice: Report of the Ad Hoc Committee on Professional Ethics]: Comment. *The American Statistician, 37,* 7–8.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5,* 434–458.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Mosteller, F. (1983). [Ethical Guidelines for Statistical Practice: Report of the Ad Hoc Committee on Professional Ethics]: Comment. *The American Statistician, 37,* 10–11.

National Academy of Sciences Panel on Scientific Responsibility and the Conduct of Research. (1992). *Responsible science: Ensuring the integrity of the research process*. Washington, DC: National Academy Press.

National Institutes of Health. (2008). Program Announcement PAR-08-212 for Methodology and Measurement in the Behavioral and Social Sciences (R01). Retrieved from http://grants.nih.gov/grants/guide/pa-files/PAR-08-212.html

Office of Science and Technology Policy. (2000). *Federal research misconduct policy*. Retrieved from http://ori.dhhs.gov/policies/fed_research_misconduct.shtml

Panter, A. T., & Sterba, S. (Chairs). (2008, August). *Quantitative psychology viewed through an ethical lens*. Symposium presented to Division 5 (Evaluation, Measurement, & Statistics) at American Psychological Association Meetings, Boston, Massachusetts.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). Boston: McGraw-Hill.

Sales, B. D., & Folkman, S. (2000). *Ethics in research with human participants.* Washington, DC: American Psychological Association.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Seltzer, W. (2001). *U.S. federal statistics and statistical ethics: The role of the American Statistical Association's Ethical Guidelines for Statistical Practice.* Washington, DC: Washington Statistical Society, Methodology Division Conference.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96,* 331–338.

Steininger, M., Newell, J. D., & Garcia, L. T. (1984). *Ethical issues in psychology.* Homewood, IL: Dorsey.

Sterba, S. K. (2006). Misconduct in the analysis and reporting of data: Bridging methodological and ethical agendas for change. *Ethics & Behavior, 16,* 305–318.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

# Section I

# Developing an Ethical Framework for Methodologists

# 2

# *Ethics in Quantitative Professional Practice*

**John S. Gardenier**
*Emeritus, Centers for Disease Control and Analysis,*
*National Center for Health Statistics*

Invited to write this introductory chapter because of my background in statistical ethics, let me explain how that came about. As an undergraduate philosophy major, I found ethics to be very interesting and relatively easy, but not very practical. All the great philosophers made well-reasoned arguments. They all would have us make positive contributions to society and avoid misbehavior. Still, none showed me *how to apply* ethical philosophy as opposed to logic, theory of knowledge, and philosophy of science (physics).

From my experience in science and with the American Association for the Advancement of Science (AAAS), I perceived science as inherently dependent on ethics but subject to corruption. Mark Frankel at AAAS organized a Professional Society Ethics Group in the 1980s. I joined initially as the representative of the Society for Computer Simulation (now the Society for Modeling and Simulation International) and later as the representative of the American Statistical Association (ASA). The AAAS Group encouraged ethics codes that provided practical guidance.

On becoming Chair of the ASA Committee on Professional Ethics, I resolved to revise the document *Ethical Guidelines for Statistical Practice* from a brief set of general principles to a more definitive guidebook for professional practice. This took a 4-year effort by the Committee and an informal advisory panel, repeated outreach to the membership generally, and the transparent posting of all written suggestions along with the Committee's documented response. After formal review by major governance committees, the document was unanimously approved by the ASA Board of Directors in 1999. It is still in effect as of this writing.

In Gardenier (2003), I analyzed the requirements for fully honest error in statistical applications. The claim of honest error can be a subterfuge for unprofessional carelessness or even for deliberate misuse of statistics. My goal in this chapter is to offer a demanding but practical introduction to general ethics and then relate that to concerns in quantification generally, in statistics, and in scientific research. This can provide a practical path

for students and others to follow toward becoming confident, competent, and highly ethical quantitative professionals. It also offers an option for a practical lifelong pursuit of happiness, philosophically defined.

It is important to understand that there are common elements to all professional ethics, whether medical, legal, engineering, or quantitative. Albert Flores (1988) summarized the societal expectations as follows:

> Professionals are expected to be persons of integrity whom you can trust, more concerned with helping than with emptying your pockets; they are experts who by the use of their skills contribute to the good of society in a variety of contexts, for a multitude of purposes; and they are admired and respected for the manifold ways they serve the growth of knowledge and advance the quality of human existence. (p. 1)

## General Frameworks for Ethics

Robert Fulghum's seminal essay "Credo" (1994) lists some key ethical maxims: "Play fair. Don't hit people. Put things back where you found them. Clean up your own mess. Don't take things that aren't yours. Say you're sorry when you hurt someone" (p. 2). Other items in the same list also suggest a philosophy of life. "Live a balanced life—learn some and think some and draw and paint and sing and dance and play and work every day some…. When you go into the world, watch out for traffic, hold hands, and stick together." "Share everything" (p. 2). This is a *rooted-in-kindergarten* ethics framework.

Fulghum's framework leaves some important questions unanswered. How does one make judgments of what is ethical or otherwise? Was Alexander the Great a good conqueror and Attila the Hun a bad conqueror? People of different cultures and ethnicities can have dramatically different assessments, as can different historians.

> The fear Attila inspired is clear from many accounts of his savagery, but, though undoubtedly harsh, he was a just ruler to his own people. He encouraged the presence of learned Romans at his court and was far less bent on devastation than other conquerors. (Attila, 2008)

In parts of central Europe and Asia, he is as greatly admired as Alexander. "After the soldiers were weary with the work of slaughtering the wretched inhabitants of the city, they found that many still remained alive, and Alexander tarnished the character for generosity and forbearance for which he had thus far been distinguished by the cruelty with

which he treated them. Some were executed, some thrown into the sea; and it is even said that two thousand were crucified along the sea-shore" (Abbott, 2009, p. 164).

My point is neither to denigrate Alexander nor to praise Attila. Rather, it is to illustrate some common ways that people make ethical judgments. It appears those judgments may depend less on the facts and more on relative self-identification with a social or ethnic culture. This creates a bias that our people are good, and anyone opposed to our people must be bad. Accepting that belief would reduce ethics to mere tribal or sectarian loyalty rather than a concept of principled behavior. Ethics, whatever we perceive them to be, must apply to both friend and foe if they are to have any credibility as representing moral values.

Bringing that thought down to the mundane, who are the unethical drivers? We may say those who drive while impaired are unethical, but who has not driven while excessively fatigued, overstressed, medicated, distracted, or injured? For ourselves, we tend to balance the perception of our impairment with the necessity or desirability of some motivation that induces us to drive. We may claim that our own impaired driving is not unethical if it is done "for a good reason" or "with careful assessment of the risks and benefits involved." When we excuse ourselves, do we also consider whether independent judgments by reasonable peers would agree with our excuse? Should we? When other drivers cut sharply in front of us, especially when they hit their brakes just after doing so, that is clearly unethical, is it not? Yet, when other drivers honk angrily at us for harmless maneuvers we have made, do they have any right to be judging the ethics of our driving?

Most people seem to find it uncomfortable to think through things like this, so they simply do not. Most people, I suspect, would prefer to keep this whole "ethics thing" simple. It helps that goal if we disparage "overthinking." We are not limited to what we learned in kindergarten, of course. We learn about ethics in many other ways: in other schooling, in religious settings, in athletics or drama or band, or in any group effort. We learn on the job to give a day's work for a day's pay, to be on time, to keep the boss informed, and so on. In fact, we have ethics pushed at us so much for so many years that we occasionally feel like saying, "Enough already about what I should sacrifice for other people; let me be the beneficiary of someone else's good deeds for a change!" Let this be our second framework. It encourages us to think about ethics beyond the simplest maxims but not to overdo that thinking. I call it the *common sense* ethics framework.

Most people do not receive formal training in ethical philosophy, so they may have little choice but to operate in the common sense framework. To the extent that they behave responsibly to family, friends, neighbors, colleagues, and to the public at large, they are clearly ethical. To the extent

that they go beyond normal expectations and devote effort and resources to those in need, to the environment, to the community, and to society at large, we should and usually do admire their ethics. People do not need a deep understanding of ethical philosophy to be basically good and decent. Still, there are some dangers lurking in the common sense framework that pose problems for individuals, for those who interact with them, and for society.

A key feature of the common sense framework is that it typically involves making ethical judgments about how badly other people behave. We can then declare ourselves ethical merely for not behaving that same way. When associated with strong emotions, this may be referred to as *moral outrage*. Moral outrage not only allows us a very strong self-pat on the back, but it also supports the idea that other people who feel that same moral outrage are, like us, good people. People who may disagree with us or argue for a more nuanced assessment of the actual issues then become ethically suspect. Moral outrage is easy, requires very little thought, and brooks no discussion. Thus, it engenders no reasoned debate. It typically does nothing to resolve the issue causing the outrage. Being angry, it certainly does not contribute to anyone's personal happiness. Above all, such moral outrage and emotionalism is beside the point. Our ethics is not a function of other people's behavior; it is about maintaining and improving our own good behavior.

Brooks (2009) very much approves of emotional ethics and disparages the ancient Greek approach to morality as dependent on reason and deliberation. He observes that moral perception is in fact emotional, which he considers "nice." After dismissing centuries of philosophical history and "the Talmudic tradition, with its hyper-rational scrutiny of texts," he states:

> Finally, it should also challenge the very scientists who study morality. They're good at explaining how people make judgments about harm and fairness, but they still struggle to explain the feelings of awe, transcendence, patriotism, joy and self-sacrifice, which are not ancillary to most people's moral experiences, but central.

A defining characteristic of maturity is having acquired the ability to guide our emotions with reason. Thus, Brooks is wrong even in the common sense framework. His title is far off the mark, however, because the philosophical tradition of ethics is not about that framework at all. In fact, emotionalism can be the enemy of ethics—not least because it allows us to praise identical behaviors in ourselves, our team, our coreligionists, our political party, and our armed forces that we simultaneously condemn in opposing bodies. That does not offer any principled structure applicable to quantitative professional ethics.

Earlier, I mentioned that studying the literature but not the application of philosophical ethics left me unsatisfied. Let us look at a particular collegiate *applied ethics* framework. The Association for Practical and Professional Ethics holds the final competition of the annual Intercollegiate Ethics Bowl in conjunction with its annual meeting. Teams of four students compete to analyze and resolve a set of preannounced complex ethical scenarios in a debate-like format. The annual Ethics Bowl is contested by the top 32 college teams from regional competitions involving many colleges and universities (Ethics Bowl, 2010). It is a great pleasure and honor to serve as a judge at such competitions, as I (in some previous years) and other applied philosophers have, alongside academic professors. The judges get to challenge each team during the debate by questioning specifics of their reasoning. No team can win unless its members are highly disciplined, well prepared and coached, capable of rapid and effective ethical reasoning, and highly articulate. They must be ready to address multiple issues involved in the ethical dilemma. Can they exhaustively name all of the affected stakeholder groups? What are the competing claims of fairness, justice, respect, and kindness? What are the impacts, if any, on the environment, on future generations? What ethical principles are most relevant? How can they draw on specific advice from noted philosophers to aid their reasoning in this case? How do politics and law condition the ethical reasoning? This is an *applied ethics* framework. It is not free from emotional involvement by any means, but it is definitely ruled by reason. Quantitative professional ethics definitely requires applied ethical thinking.

Ideally, all graduate-level courses (at least) in statistical and other analytic methods should include explicit attention to the associated ethical implications. However, there is no universal standard for such ethical thinking. Some ethics training is required for predoctoral grants from agencies such as the National Institutes of Health (NIH) and the National Science Foundation (NSF). Beyond such external requirements, some professors believe that ethical issues are just as fundamental as the definitions of terms, the algorithms taught, and the ability to solve textbook problems satisfactorily. Whatever quantitative ethics are taught and, more importantly, inculcated by example is highly dependent on the specific faculty involved.

In the professions, including the quantitative and scientific professions, various organizations and institutions promulgate ethics documents. These may be called codes of ethics, ethical guidelines, standards of practice, or other terminology. We need to know whether any such document applies to a particular profession or practitioner and what generic type of ethics document it is. Frankel (1989) addressed three types: aspirational, educational, and regulatory.

Aspirational ethics codes exhort people to "be good" in generic but vague ways. These are not intended to be comprehensive sets of instructions

dictating how to deal with any specific issue. Rather, they urge one to be honest, competent, trustworthy, just, respectful, loyal, law-abiding, and perhaps even collegial and congenial. The implicit assumption is that one has already learned ethics adequately. These codes basically say, "You know what is right; just do it." Ethics documents that are short lists of simple sentences are aspirational.

Educational ethics documents try to spell out some specific issues facing people in certain disciplines, occupations, or within specific companies or institutions. They lay out, in greater or lesser detail, certain types of behavior that are specifically favored by the promulgating body. Any significant variance from those guidelines is ethically suspect. These are considerably more detailed than aspirational codes, but they are not intended to be exhaustive. Just two of many examples of these ethics documents for quantitative professions include the ASA's *Ethical Guidelines for Statistical Practice* (1999) and the Association for Computing Machinery's *Code of Ethics and Professional Conduct* (1992).

Regulatory ethics codes are more definitively legalistic. They tend to use the verb "shall" to indicate ethical obligations. They may involve standards of evidence and proof and, over time, actual case histories. There may be specific punishments for specific sorts of violations. In case of an accusation of ethical violation, one may be subject to an initial inquiry and further measures: formal investigation, appointment of a prosecuting authority, entitlement to defensive support such as one's own attorney, a ruling authority, and an appellate process. They are exhaustive in the sense that no one is subject to the defined penalties except for specific proven violations of defined rules. An example is the National Society of Professional Engineers' *NSPE Code of Ethics for Engineers* (2007).

Taken together, these institutional documents or codes can be summarized as an *ethics from documents* framework. By this, I mean that one may approach ethics from any of the other frameworks and add relevant ethics documents for an important set of added values. Even if you use quantitative methods only to support nonquantitative projects, you should be equally as aware of the ethical principles involved as you are aware of the mathematics. Otherwise, you will be at risk of seriously misusing those methods to the detriment of your work. Typically, quantitative professionals need to follow both the ethics documents of the subject matter field, such as biomedical research or psychology, and those of the quantitative discipline involved, such as statistics.

This Handbook is part of the ethics from documents framework in the sense that it supplements whatever documented ethical guidance may otherwise be applicable to one's own professional work. A pervasive problem with such ethics documents is that their brevity precludes discussion of methodological means to resolve ethical problems. As you will see in subsequent chapters, this Handbook addresses such needs.

The most philosophical and demanding one is the *pursuit of happiness* framework. It involves making ethical service to society at large the central guiding principle of one's entire life, aided importantly by logic and science. This concept comes from ancient Greek and Roman schools of philosophy. Different philosophers, such as epicureans, stoics, and cynics, among others, competed to attract paying students (Irvine, 2009). They marketed their instructional programs as "schools of life." They taught students how to live well, that is, how to engage over a lifetime in a continuing and successful pursuit of happiness. Many of these schools thrived, having alumni who did in fact enjoy better lives. Terms associated with these philosophies have been corrupted to the definitions in today's language. A true Epicurean is not "epicurean" in the sense of being devoted to enjoying good food and wine. A true Stoic is not "stoic" in today's sense of someone who suppresses all emotion. One must consult authoritative sources to understand the true principles of these ancient philosophies, for example, Axios Institute (2008).

Key elements of instruction might include duty to society, control of one's desires, mastery of applied logic, and understanding the natural world. Means to pursue happiness include rigorous mental discipline, some greater or lesser degree of voluntary deprivation of comfort, refusal to worry about things one cannot control (such as the past), refusal to desire objects and goals that are either deleterious to self or society or are too remote to be practically attainable, and effective means to master one's personal emotional reactions to adverse natural or political events and to conflicts with other individuals.

For the most part, the instructional manuals for these philosophies of life have been lost to history. There are key elements of the Stoic philosophy in some surviving writings of the Romans Epictetus, Seneca, and Marcus Aurelius. One can find a number of books that explore and summarize their thinking, for example, Fitch (2008) and Marcus Aurelius (trans. 1963). Interestingly, George Washington considered himself a stoic based on his readings of Seneca's letters. That helped him face death fearlessly (Grizzard, 2002). "Washington had developed a stoical attitude toward death at an early age …" (p. 74). For those who may be interested, we are fortunate that a modern how-to book on Stoicism and its joy (Irvine, 2009) is now available. It provides us with the relevant history, philosophy, modern interpretation, and how-to principles.

Does it really make sense for quantitative professionals to opt for pursuit of happiness ethics? The *Wall Street Journal* (2009) published a list of the 200 best and worst jobs in the United States. The happiest three were mathematician, actuary, and statistician. The fifth and sixth were software engineer and computer systems analyst. The ranking is based on a combination of working conditions, pay, and interesting work (Needleman, 2009). Thus, choosing a quantitatively oriented career is a good first step

toward happiness. To expand that to a pervasive lifetime experience requires serious study and adaptation of the relevant principles to one's own culture, religion, and environment. It also requires disciplined practice to control one's reactions to adverse circumstances, to other people's offensive behavior, and to the disappointments that are inevitable if one does not gain effective control over one's desires. Finally, the pursuit of happiness framework is based on a lifelong ethical commitment to bettering society—in our case, by steadfast application of competent and ethical quantitative professionalism.

This concludes the topic of general ethics. To review, I have described the following alternative ethical frameworks. The first two may suffice for everyday life, but you will have to practice applied ethical thinking and learn the ethics documents relative to your work to become proficient in quantitative ethics.

1. Rooted-in-kindergarten ethics
2. Common sense ethics (applied in a principled manner)
3. Applied collegiate ethics
4. Ethics from documents
5. Pursuit of happiness ethics

## Assumptions in Mathematics and Statistics

All quantitative professionals depend on sound use of mathematics, so they should understand the dependence of mathematics on the underlying definitions and assumptions involved. Mathematics can be viewed as an elaborate and often useful a priori construction. "Whereas the natural sciences investigate entities that are located in space and time, it is not at all obvious that this is also the case with respect to the objects that are studied in mathematics" (Horsten, 2007). Mathematics can constitute a game or pastime (Bogomolny, 2009). It is a basis for understanding all physical, biological, and social science. Finally, it offers a tool kit for practical pursuits such as engineering and policy analysis. Like other tools, mathematics can be used well or badly, knowledgeably or ignorantly. It can benefit individuals or society; it can also be used to deceive, to cheat, and to destroy. Thus, it implicitly needs some ethical controls.

Realizing that numbers do not necessarily relate to realities of interest, it is important to communicate clearly what sort of mathematics is involved in work one has done (or will do). What justifies a claim that the quantification one is using is relevant to the realities of interest to the

reader, employer, or client? For example, the number 100 may represent a quantity of 100 (in base 10), a quantity of 4 (in binary), or many other quantities. It may be stated in units of thousands or millions or billions; it may be in a logarithmic or some more elaborate scale; or it may represent an ordinal relationship such as the 100th largest U.S. city in population. It may simply refer to the (arbitrarily) 100th item in a nonordered list. If it is a percentage, what is the base with which it is to be compared? How many digits are appropriate to express it? The bottom line is that we must constantly be aware of the assumptions inherent in the mathematics we use and clearly communicate any limits on our results imposed by those assumptions. Throughout this Handbook, the reader will encounter many examples of the dependence of particular statistical methods on specific, although not always stated, formal assumptions.

The alternative frameworks for mathematics, statistics, and quantification generally are as follows:

1. Mathematics as theoretical discipline requiring imagination, strict logic, and formal methods of proof, but not necessarily any empirical referents
2. Mathematics as a game or pastime for personal pleasure
3. Mathematics as a descriptor of the real world
4. Mathematics as a set of tools useful for good or bad purposes

Clearly, we are mainly interested here in Frameworks 3 and 4 and only in their use for good purposes. Frameworks 1 and 2 represent self-referential systems that can offer their devotees great amounts of insight, enlightenment, pleasure, awe, beauty, and fascination. They also underlie most, if not all, of applied mathematics. They are not to be dismissed but are merely set aside for our current purpose.

One subdiscipline of mathematics is statistics. Hays (1973) defines *descriptive statistics* as "the body of techniques for effective organization and communication of data" (p. v). He goes on to define *inferential statistics* as "the body of methods for arriving at conclusions extending beyond the immediate data" (p. v). Like other forms of mathematics, statistics can be viewed theoretically as the study of probabilistic relationships among variables with no necessary relationship to the world. People can also choose to play statistical games. Those are mathematical Frameworks 1 and 2. Applied statistics in mathematical Frameworks 3 and 4 seeks to describe aspects of the real world, to understand nature and society better, and to arrive at useful predictions about those. It is possible to arrive at entirely new formulations of statistical methods that expand known theory in Framework 1 simultaneously with solving practical problems in Frameworks 3 and 4.

Necessary concepts for statistical work include probability, randomness, error, distributions, variables, variance, and populations. For applied statistics, we must add data sampling as a necessary concept. Often useful but not always necessary concepts include independence, outliers, exactness, robustness, extrapolation, bivariate and multivariate correlation, causality, contingency, and risk. Although it is tempting simply to apply statistical software to real world problems without any training in these underlying concepts, that is misleading, error prone, and certainly far from professional. It is also necessary to bear in mind that, in statistical usage, the words describing these concepts have more precise and technical definitions than the same words in everyday usage.

For our purposes, statistics only incidentally touches on lists of facts. It is primarily about mathematical methods of estimation that seek to account explicitly for some degree of error in estimates. It rests importantly on the concept that any particular sample of data is a single instance of many possible data samples that might have been obtained to describe the phenomena of interest. As such, it is merely an anecdotal estimate of the true properties of those phenomena. In general we can obtain a more accurate estimate of the properties by taking more readings (larger samples) of the same phenomena. The estimation process depends importantly on assumptions we make about the phenomena and assumptions implicit in the specific statistical methods used. In practical applied statistics, it is common to relax these assumptions somewhat, but part of the competence of anyone who uses statistics professionally lies in the ability to understand when and to what degree such assumptions can be relaxed while maintaining validity.

One of the most important assumptions underlying statistical methods is randomness in the sense of lack of bias. Random sampling is a procedure designed to ensure that each potential observation has an equal chance of being selected in a survey. The same principle applies to other means of collecting observations besides surveys. Much of statistical theory rests on the assumption that the data sample is random—either overall or within defined boundaries such as strata. If the sample is biased with respect to any of the properties we seek to estimate, then our estimates will be incorrect, as will our estimates of the surrounding error. Another basic concept is that we cannot "shop around" for statistical estimates of any property from one given sample of data. That is, we cannot subsequently decide that we do not like the result we have calculated and then reperform the estimate on the same sample using a different method in hope of getting a more pleasing result.

Similarly, we have to accept that the data are what they are. An issue that may be overlooked is how do we know what the data are? They do not simply appear out of nowhere, neatly organized and arrayed for analysis. Statisticians need to be cautious when being asked or told to analyze

a data set someone has handed to them. How and why were the data obtained? Were they collected expressly for the study at hand or for some other purpose? Why are these specific variables and these specific observations or experimental points collected and not others? Were all data points collected with a common protocol and in a disciplined manner? What was considered to be the "population" (totality of possible observations) of interest? What process defined the portion of the population (by time, location, or characteristics) from which a statistical sample was to be drawn—sometimes called the sampling frame? What sampling plan was used to ensure a random sample from that frame? (See Sterba, Christ, Prinstein, & Nock, Chapter 10, this volume.) How were the sampled data reviewed, quality controlled, organized, arrayed, and transcribed? Were any collected data discarded or set aside and, if so, why? Are there elements in the data set that were not directly collected but rather were derived from the collected data and, if so, how? Many data sets have missing values for good and practical reasons. How are those missing data to be treated in the analyses? Ethical issues with missing data are discussed by Enders and Gottschall, Chapter 14, this volume. Above all, what does the requestor seek to learn from the data? What uses are likely to be made of the results? Answers to all of these questions will influence how a diligent, competent, and ethical statistician will proceed. As ASA President Sally Morton stated in her commencement address to the 2009 graduating class of North Carolina State University, "… don't trust complicated models as far as you can throw them—protect yourself and examine those data every which way."

In routine applications, such as quality control calculations for a standard industrial process, the procedures to be followed may be very straightforward and well established. Let us look at the opposite case, where no similar data set has ever been analyzed for the current purpose. In such a case, the statistical practitioner would be wise to proceed very cautiously. First, one should perform a thorough exploratory data analysis (Tukey, 1977). One seeks first to understand the distributional characteristics of each variable separately. What does a scatter plot look like? Is this a continuous variable like weight or time, or is it instead a discrete variable like gender or nationality? Does it vary smoothly like temperature, or does it jump in discrete steps like military or professorial rank? Does it have a single area of concentration, or does it resolve into two or more groups? Do most of the data cluster together, or do many observations trail apart in small bunches? After understanding each variable separately, one can explore their relationships one to another and one to several others. Only with the knowledge gained from exploratory data analysis can a diligent statistician determine which of many possible statistical approaches offers the best chance of obtaining a methodologically valid set of analytic outputs for the task at hand.

With the analytic output—charts, graphs, tables, and computational results—in hand, the user of statistics then proceeds to interpret the data. Mindful of all that is known about the data characteristics and the sample, the goals of the analysis, the results of the exploratory stage, and the assumptions implicit in the final analytic method, a competent and ethical practitioner can then translate the mechanical outputs into a readable narrative with supporting illustrations.

The previous sentence raises another important ethical problem—communication. Ethics demands consideration of the intended reader(s). If directed to an employer or client in response to a specified requirement, the output must focus on the issue raised, the approach taken, the solution arrived at, and any caveats regarding the credibility and applicability of the result. If that document is to be reviewed by other experts, it may be appropriate to attach a methodological appendix. If it is to be a work for a technical audience, the methodology will usually be in the body of the document and, where appropriate, should contain means of accessing the data for peer review and replication purposes. This is subject to limitations of confidentiality, proprietary secrets, and any contractual provisions, of course. Finally, if the communication is to a general readership, the primary consideration is that the readers will gain an accurate picture of the results of the study (and the limitations of those results) in the absence of technical jargon or detailed justification. This is difficult and demands special attention. Above all, any communication of quantitative results must be credibly capable of resulting in an accurate understanding by the reader without distortion, misdirection, or confusion.

Statistical methods are often used to predict future behaviors or events based on descriptive observations of the past. This generally involves time series analysis. The predictions are not "data" in the sense of some characteristic or measurement drawn from the actual world. Rather, these are extrapolations of observed patterns into the future. They require an assumption that we understand the causes or at least some strong correlations between "independent" variables as event drivers and "dependent" variables in the present and as predicted outcomes. We may also make assumptions about the future changes in the independent variables to predict the effects on the dependent variables of interest. Prediction is even more fraught with complications than statistical description. It is well beyond the scope of this introductory discussion. Problems of statistical prediction tend to be discussed in specific contexts such as process quality control (Wilks, 1942).

Both observations and predictions are subject to two basic types of error: sampling error and nonsampling error, sometimes called "measurement error." Sampling error arises, as noted above, from the basic assumption that any data sample is but one random sample of a much larger set of possible observations or predictions. Measurement error addresses the

possibility that we have mischaracterized the variables of importance to the task at hand, that we have used deficient methods to measure or record those variables, or that we have mischaracterized the relationships among the variables. An excellent discussion of ethical issues with measurement error can be found in Carrig and Hoyle, Chapter 5, this volume. Caution and diligence must be regularly used to avoid measurement error and minimize sampling error.

There are two basic ways of obtaining quantitative data: by controlled experiment or by observation. Observations are drawn from social, economic, or physical settings without any attempt to alter the processes being observed. Controlled experiments are importantly different. They artificially manipulate some determinant of behavior, such as mazes for rats to solve, medical interventions to assess treatments, or controlled crashes of test vehicles to study the efficacy of safety features. Sometimes ethical problems in experimentation can be addressed by comparing different sets of observations in structured quasi-experiments (see Mark & Lenz-Watson, Chapter 7, this volume).

## Understanding Scientific Research

Scientific "research" differs from the everyday usage. Schoolchildren conduct research when they consult a dictionary, library book, or some web sites. In that sense, research means consulting a supposedly authoritative source to obtain the truth about some concept, person or group, or set of events that are known to others but not to the person doing the research.

Science is very different in that, in the strictest sense, there are no authorities in science. Scientific facts can only come from observation or experiment or computation. Even then, the "facts" do not constitute "truth." The fact of a scientific observation is a combined product of the underlying true state of nature, the means of observation and data recording, the skill of the personnel, the adequacy of the equipment used, and finally the concepts and assumptions that underlie the investigation at hand. As a social construct, a consensus among respected scientists is sufficient "authority" to accept a body of knowledge without each individual having to reproduce every observation and experiment involved.

Scientific studies may or may not embody the concept of random error. In the many cases in which this is the case, statistical concepts and methods are crucial to the science. Where repeated observations have established that the random error is trivial it can be ignored. This is not as simple as it sounds; good science requires exhaustive investigation of any

conceivable cause for the observations to appear to be precise when that is not actually true.

One of the most important assets to science is the *scientific record*, which refers to the accumulated body of knowledge contained in books, journals, conference proceedings, transcripts of scientific meetings, data banks, data sets, and online libraries. The ideal is a record that describes only research that was meticulously designed and conducted, recorded in a complete and accurate manner, that is properly described and interpreted, and is backed up by the availability of the underlying observational or experimental data for use in peer review or replication studies. Ethical scientists seek to approach that ideal.

The scientific record generally contains only reports of studies that were successful in adding to the state of knowledge in the field. Some scientists have argued that it could be equally valuable in many cases to document well-designed and -executed studies that were unexpectedly unsuccessful. The World Health Organization (2009) notes, "Selective reporting, regardless of the reason for it, leads to an incomplete and potentially biased view of the trial and its results." They are referring to biomedical research generally and clinical trials specifically, but the same principle applies across the entire spectrum of scientific research.

There is also a problem with common judgmental shortcuts, such as conventional reliance on a statistical $p$ value of .05 or less as indicative of a gold standard of scientific acceptability. The $p$ value is intended to represent a suitable degree of confidence that a given observational or experimental result is unlikely to have resulted from mere random variation among possible samples. This becomes a systemic problem for the scientific record when journal editors and peer reviewers require evidence of a .05 $p$ value regardless of whether it provides a reasonable level of confidence for the issue at hand. It is even worse when the stated $p$ value is accepted based only on the normative science but with no competent review of the credibility of the statistical methods. Because statistics can readily be fudged, setting a criterion $p$ value for acceptance while failing to assess the validity of the value claimed creates an implicit motivation for failures of diligence or even deliberate misuses of statistics. That appears to be unprofessionally careless if not definitively unethical. It is also common. For a discussion about understanding and countering this problem, see Gardenier (2003). To understand preferable alternatives to hypothesis testing with $p$ values in some situations, see Cumming and Fidler, Chapter 11, this volume.

The scientific record can also be problematic in terms of intellectual property law. Science can only progress by building freely on previous science. Publishers of online and print journals of record and of some crucial databases are tempted to keep their material out of the public domain so that each user institution or individual has to pay subscription fees for

access. The scientific community would prefer that all science and related databases be freely available in the public domain as soon as possible. Still, much of the publication is controlled by scientific societies or institutions that have compelling financial needs themselves. The result is a search for balancing the good of the institutions that serve science with the best interests of the scientific enterprise at large. See, for example, Esanu and Uhlir (2003).

It is essential to understand the demands made on scientists that detract to some extent from their preferred work of "doing science." Many academic scientists cannot and should not limit themselves to their own research. They must also teach and mentor students endeavoring to follow in their footsteps. Some of them must devote time to administrative chores or else their department, school, or university could not function. Peer reviewers and journal editors are usually not paid positions. These crucial tasks, without which the scientific enterprise cannot function, are often performed by volunteers. How much volunteer time is reasonable for a working scientist to devote pro bono to the benefit of "the scientific enterprise" instead of spending that time on his or her own research? Also, scientific research is usually expensive. Often the money needed for it must come from grants funded by governments or foundations. The grants are competitive to ensure funding only the best scientific proposals, which means that applications for grants tend to be voluminous, highly technical, and sometime gargantuan in the workload demanded. Scientists are also asked by the granting organizations to perform voluntary peer review of other scientists' grant proposals. Inherently, then, scientists contribute ethically to society by helping to facilitate the scientific enterprise as well as by any benefits resulting from the knowledge they produce.

In 2002, the Institute of Medicine Committee on Assessing Integrity in the Research Environments stated, "Attention to issues of integrity in scientific research is very important to the public, scientists, the institutions in which they work, and the scientific enterprise itself. [Yet] No established measures for assessing integrity in the research environment exist" (p. 3). The U.S. Office of Research Integrity has delved into this complex problem through five biennial research conferences on research integrity. No direct path to a solution has emerged.

Instead, an attempt has been made to infer research integrity from an absence of research misconduct. The National Science and Technology Council (2000) defined research misconduct essentially as including fabrication, falsification, and plagiarism (known as FF&P) plus any retribution taken against those who in good faith report or allege misconduct. (These people are popularly known as "whistleblowers.")

> *Fabrication* is making up data or results and recording or reporting them. *Falsification* is manipulating research materials, equipment,

> or processes, or changing or omitting data or results such that the research is not accurately represented in the research record. *Plagiarism* is appropriation of another person's ideas, processes, results, or words without giving appropriate credit. Research misconduct does not include honest error or differences of opinion.

However, the honesty of scientific error depends on the competence and diligence of the work as well as the lack of any intent to deceive.

Many scientists consider this approach to research ethics to be minimalist and inadequate. It fails to address the concept that scientists should not engage in research for which they lack the required scholarly and experiential background. It fails to address malfeasance in peer review. It ignores a practice by which science supervisors or advisors are listed as authors on scientific papers to which they made no relevant independent contribution. It ignores conflicts of interest as when a scientific study supports a commercial product or company without disclosing that the lead scientist is regularly a paid consultant to that same company. There are many more possible examples, but these are illustrative of the problem. Among sources that advocate a much broader view, see Commission on Research Integrity (1995). Above all, one can claim that the minimalist attempt to equate research integrity with a lack of research misconduct "will not work." An experienced assistant provost officer noted that colleagues will never report one of their own for "research misconduct" even when their colleagues exhibit behavior they recognize as immoral, disgusting, and scientifically wrong (Gardenier, 2003). This is an example of reducing ethics to mere sectarian loyalty rather than any concept of principled behavior.

As a practical matter, scientists and their institutions achieve success only to the extent that they are perceived to be not only competent and ethical but also exemplary in their performance. Academic scientists and some in industry are rewarded with tenure, promotions, opportunities for continuing study, awards, prestigious offices of professional societies, membership on important study groups, and election to academies such as the National Academy of Science, National Academy of Engineering, or Institute of Medicine. They also compete for awards such as the Nobel Prize and the Field Medal in mathematics. Scientists perceived by their peers to be ethically suspect do not and should not have access to such rewards.

The bottom line in scientific research ethics is that the individual scientists, the lab directors, and other supervisors and teachers of research, as well as the institutions that employ them, are the ultimate arbiters and enforcers of scientific ethics. They can get useful information and insights from publications in their fields and also from journals such as *Science and Engineering Ethics*. We will revisit this reliance on individual and group ethics when we address the real world next.

## Ensuring Your Right to Professional Integrity

The *Ethical Guidelines for Statistical Practice* (1999) devotes an entire section to "Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners." This was required because of widespread experience of employers trying to use quantitative methods as window dressing to bolster preconceived notions or special interests. Ethically, we as quantitative professionals cannot engage in such dishonesty. There is a relevant cartoon with the caption, "You are right, Jenkins, these numbers do not lie. Get me some that do" (Coverly, 2009). In fairness, not all such attempts at manipulation are blatantly dishonest. Some employers simply are not aware of or do not understand the concept of quantitative professionalism. We may have to explain it.

Earlier, I pointed out that quantitative careers tend to have a number of advantages, including generally safe and pleasant work environments, interesting projects, good pay, and opportunities to contribute to societal well-being. Among other things, this depends on a climate that respects and promotes professionalism and ethics. Thus, I would urge anyone evaluating a job opportunity, or a potential client in the case of an independent consultant, to consider and even discuss explicitly your own understanding of the ethical demands of your profession. Only agree to work for and with other people who have compatible values.

Let us look at some of the general principles of applied professional ethics. Trustworthiness should come first. No one should claim professional status who does not have all the skills needed for any task one may undertake. One must be totally honest also about the terms of employment or engagement. One will conduct and report the work as thoroughly and well as resources permit with a focus on the most methodologically valid result, which could entail a result different from what the client or employer would prefer. The advantage of course is that the client or employer can trust the result and proceed accordingly.

The next most important ethical principle has to be respect for persons as autonomous individuals. A professional sometimes is seen as an authority figure to people who lack the specific expertise and skill set. That power must be used with caution so as to avoid having people rely on you for matters they rightly should decide for themselves. This is a particularly prevalent consideration in biomedical research where a vast body of policy exists for the protection of human subjects of research (Office for Human Research Protections, n.d.). Those whose work is affected by those rules must learn and follow them. Even outside the areas where such policies apply, one should consistently treat every person with respect for his or her personal safety, dignity, and autonomy.

It is imperative to note, however, that treating people ethically does not imply letting down our guard or failing to protect our own legitimate interests when some people around us behave badly or inconsiderately. It is not "ethical" to be a hapless victim of arrogance, negligence, or mistreatment; that would be both unwise and unnecessary.

There is a lot of overlap between ethics, law, and politics. All three would have us avoid lying, stealing, taking credit for someone else's work, or causing harm to people or their work, to name just a few principles. When matters go the other way, however, and such things are done to us, there may or may not be an ethical lapse on another person's part. Usually we do not know what was in their hearts or minds. As to whether there may have been a breach of law or regulation may not be something we have the ultimate right to determine. We may form an opinion, but until the responsible authorities have made a duly considered determination, our opinion may have to remain only that. What we have to fall back on in such a case is politics. We must deal with friends, adversaries, and authorities in as effective a manner as we can. Typically one cannot start such a process only when an issue arises such that we need help. People need preventively to form networks of friends, colleagues, patrons or mentors, supporters, and administratively skilled experts. In times of need, we may have to mobilize these people's support.

One source of networking is your specific personal connections within your professional discipline. It is wise to join one or more professional societies in your field even while a student. One may start by attending meetings, submitting posters or contributed papers, and volunteering to serve on a committee or section. As you get to know more people in your field and gain experience in professional service, your reputation and your access to potential professional advice will grow.

The most important type of connectedness you should seek is within your own organization if you are an employee rather than an independent consultant. Your own ethics should of course cause you to work diligently, competently, and amiably. You should be dedicated to the success of your organization, your group, and your superiors. Such behavior is usually rewarded. It is wise also to make friends in administrative areas such as human resources, finance, and legal. One way to do that is through volunteering to help with organizing charity functions, work-related sport or social functions, or to serve in volunteer areas such as equal employment opportunity counseling or providing short courses in your specialty to others in the organization whose job performance could benefit from what you can teach. The possibilities are endless, so you have to be careful to keep your own assignments as your primary focus. Select your additional activities where you can make the greatest contribution with the least demand on your scarce

resources, like money and time. In addition to the rewards of service and the resulting larger circle of friends, your additional activities may occasionally earn you some favorable recognition by senior managers. If at all possible, one should also cultivate a senior, highly experienced person who is not in one's own chain of command as a mentor or advisor. Many senior professionals are happy to assist younger professionals.

You may ask what all this politicking has to do with ethics. There are several answers. For one thing, your overall ethical aim should be to contribute to the greater social welfare. The activities recommended above give you additional opportunities to do just that. It is also true that one can benefit from not only being an ethical person but also by being recognized as contributing to one's profession, one's organization, and one's community. There is downside potential as well, of course. You must always exercise personal humility and avoid any implicit or explicit braggadocio about your connections, your productivity, or your ethics. Avoid provoking your superiors or coworkers. If they are to admire you, it must be because you are viewed as benefiting their interests and their self-regard instead of diminishing those.

The more that people you associate with admire you, the less reputational harm a jealous and less ethical person can do to you. If you should need to find new employment, then the greater your network of amiable and admiring colleagues, and the greater your chances of finding desirable opportunities. Throughout one's professional career, there will be many situations in which one may be uncertain as to how to handle an opportunity, a perceived threat, an exceptionally tricky assignment, or a difficult coworker or superior. Some useful guidance may be available from books such as Whicker and Kronenfeld (1994). Beyond that, the greater your personal circle of wise, knowledgeable, and trustworthy friends, the better the ethical advice that will be available to you to handle such situations.

This becomes exceptionally necessary when one is faced with a serious ethical problem. If one were to observe a clear violation of law or professional ethics, one may not be able to avoid a personal responsibility to do something about it. Yet, stepping into such situations can involve tricky intersections of ethics, law, politics, and personalities. In general, one should seek the solution that rectifies the problem with the least number of people involved and the least amount of formal action. A relatively young or otherwise vulnerable employee must have wise senior friends as advisors on the handling of such situations. In the worst case, you may face an unavoidable personal responsibility to make a formal complaint against a coworker or supervisor. This is often called "blowing the whistle." Despite many protections for whistleblowers that are built into law and organizational practice, it is not

uncommon for whistleblowers to suffer severe personal, professional, and economic damage as a direct result of that action (Research Triangle Institute, 1995). An essential reference on handling such situations is Gunsalus (1994).

## Conclusion

This chapter has attempted to provide general practical guidance for ethical conduct of any quantitative professional work or career. It exhorts you to devote time and study to ethics with an emphasis on applications in real life quantitative practice. You must come to understand all ethics guidance relevant to your position or to your work. I also recommend adopting a definitive philosophy of life, perhaps based on the teachings of ancient philosophers and adapted to other important influences in your life.

You must gain a fundamental understanding of the dependence of practical mathematics on the definitions, assumptions, and basic principles that underlie their theory. When applying any statistical methods, you must account for—and possibly adjust for—the key assumptions underlying them. If your career takes you into science, you must assume responsibility for facilitating and protecting the scientific enterprise overall; you cannot limit yourself simply to doing your own science. Ethical quantitative and scientific work does not depend only on competence and honesty. Consistent diligence, collaborative association, and meticulous thoughtful communication are additional constant challenges. Finally, your own ethics will not be sufficient if you are not in a work environment that values and supports ethical professionalism. Even in such an environment, it is wise to form networks of friends and advisors who can help you in case ethical or other sorts of problems arise.

The bottom line to all the considerations in this chapter is that you have the opportunity to become a confident, competent, ethical, and well-respected quantitative professional. You may even become, with some experience and lots of connectedness, virtually invulnerable to any attack by rivals or disaffected superiors, to economic downturns, or to other events that may adversely impact your employment or client base. You will find then that your rigorous sense of ethics with respect to professional practice and your principled regard for all persons you deal with, combined with the satisfactions inherent in a quantitatively oriented career, will contribute importantly to your successful lifelong pursuit of happiness.

## References


Abbott, J. (2009). *Alexander the Great*. Chapel Hill, NC: Yesterday's Classics. (Original work published 1902.)

American Statistical Association. (1999). *Ethical guidelines for statistical practice.* Retrieved from http://www.amstat.org/about/ethicalguidelines.cfm

Association for Computing Machinery. (1992). *ACM code of ethics and professional conduct.* Retrieved from http://www.acm.org/about/code-of-ethics

Attila. (2008). In *Columbia Encyclopedia* (6th ed.). Retrieved from http://www.encyclopedia.com/topic/attila.aspx#1e1-attila

Aurelius, M. (1963). *The meditations* (G. M. C. Grube, Trans.). Indianapolis, IN: Bobbs-Merrill.

Axios Institute (Ed.) (2008). *Epicureans and stoics.* Mount Jackson, VA: Axios Press.

Bogomolny, A. (2009). *Interactive mathematics miscellany and puzzles.* Retrieved from http://www.cut-the-knot.org

Brooks, D. (2009, April 6). The end of philosophy. *New York Times.* Retrieved from http://www.nytimes.com/2009/04/07/opinion/07brooks.html?scp=1&sq=&st=nyt

Committee on Assessing Integrity in Research Environments. (2002). *Integrity in scientific research: Creating an environment that promotes responsible conduct.* Washington, DC: National Academies Press.

Commission on Research Integrity. (1995). *Integrity and misconduct in research.* Washington, DC: Department of Health and Human Services.

Coverly, D. (2009). *Speed bump.* Retrieved from http://www.americanprogress.org/cartoons/2009/04/040109.html

Esanu, J. M., & Uhlir, P. F. (Eds.). (2003). *The role of scientific and technical data and information in the public domain.* Washington, DC: National Academies Press.

Ethics Bowl. (2010). *Center for the Study of Ethics in the Professions at IIT, Illinois Institute of Technology*. Retrieved from http://ethics.iit.edu/index1.php/programs/ethics%20bowl

Fitch, J. (Ed.). (2008). *Seneca*. New York: Oxford University Press.

Flores, A. (Ed.). (1988). *Professional ideals.* Belmont, CA: Wadsworth.

Frankel, M. S. (1989). Professional codes: Why, how, and with what impact? *Journal of Business Ethics, 2,* 109–115.

Fulghum, R. (1994). Credo. In *All I really needed to know I learned in kindergarten.* New York: Ballantine.

Gardenier, J. S. (2003). Best statistical practices to promote research integrity. *Professional Ethics Report, 16,* 1–3.

Grizzard, F. E., Jr. (2002). *George Washington: A biographical companion.* Santa Barbara, CA: ABC-CLIO.

Gunsalus, C. K. (1994). How to blow the whistle and still have a career afterwards. *Science and Engineering Ethics, 4,* 51–64. Retrieved from http://www.indiana.edu/~poynter/see-ckg1.pdf

Hays, W. L. (1973). *Statistics: For the social sciences* (2nd ed). Austin, TX: Holt, Rinehart, and Winston.

Horsten, L. (2007). Philosophy of mathematics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Retrieved from http://plato.stanford.edu/entries/philosophy-mathematics

Irvine, W. B. (2009). *A guide to the good life: The ancient art of stoic joy.* New York: Oxford University Press.

Morton, S. (2009). ASA president delivers NC State commencement speech. *AMSTAT NEWS*, *385*, 13.

National Science and Technology Council. (2000). *Federal policy on research misconduct.* Retrieved from http://www.ostp.gov/cs/federal_policy_on_research_misconduct

National Society of Professional Engineers. (2007). *NSPE code of ethics for engineers*. Retrieved from http://www.nspe.org/ethics/codeofethics/index.html

Needleman, S. E. (2009, January 26). Doing the math to find the good jobs. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB123119236117055127.html

Office for Human Research Protections. (n.d.). *Policy guidance*. Washington, DC: U.S. Department of Health and Human Services. Retrieved from http://www.hhs.gov/ohrp/policy

Research Triangle Institute. (1995). *Consequences of whistleblowing for the whistleblower in misconduct in science cases.* Retrieved from http://ori.hhs.gov/documents/consequences.pdf

Tukey, J. W. (1977). *Exploratory data analysis.* New York: Addison-Wesley.

Whicker, M. L., & Kronenfeld, J. J. (1994). *Dealing with ethical dilemmas on campus*. Thousand Oaks, CA: Sage.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Annals of Mathematical Statistics, 13,* 400–409**.** Retrieved from http://projecteuclid.org/handle/euclid.aoms

World Health Organization. (2009). *Reporting of findings of clinical trials.* Retrieved from http://www.who.int/ictrp/results/en

# 3

# *Ethical Principles in Data Analysis: An Overview*

**Ralph L. Rosnow**
*Emeritus, Temple University*

**Robert Rosenthal**
*University of California, Riverside*

This chapter is intended to serve as a conceptual and historical backdrop to the discussions of particular ethical issues and quantitative methods in the chapters that follow in this Handbook. Before we focus more specifically on modern-day events that fired up concerns about ethical issues, it may be illuminating to give a sense of how the consequences of those events, as indeed even the need for this Handbook, can be understood as a piece in a larger philosophical mosaic. In the limited space available, it is hard to know where to begin so as not to oversimplify the big picture too much because it extends well beyond the quantitative footing of modern science. If we substitute "mathematical" for statistical or quantitative, and if we equate the development of modern science with the rise of experimentalism (i.e., *per demonstrationem*), then we might start with Roger Bacon, the great English medieval academic and early proponent of experimental science. In his *Opus Majus*, written about 1267, Bacon developed the argument that:

> If … we are to arrive at certainty without doubt and at truth without error, we must set foundations of knowledge on mathematics insofar as disposed through it we can attain to certainty in the other sciences, and to truth through the exclusion of error" (quoted work reprinted in Sambursky, 1974, p. 154).

Given the sweep of events from Bacon to Galileo and Newton, then to the 20th century and to our own cultural sphere, it is hardly a revelation to point out that the idea of "certainty without doubt" and the notion of "truth without error" were an illusion. The scientific method is limited in some ways that are specifiable (e.g., ethical mandates), in ways that are

"unknowable" because humans are not omniscient, and in symbolic ways that we "know" but cannot communicate in an unambiguous way (cf. Polanyi, 1966).

Suppose we begin not at the beginning but instead we back up just a little to give a glimpse of how we think a *Handbook of Ethics in Quantitative Methodology* fits into this big philosophical picture. For generations, the philosophical idealization of science as an unfettered pursuit of knowledge, limited only by the imagination and audacity of scientists themselves, remained relatively intact. By the mid-20th century, that old image had faded noticeably as challenges were directed against foundational assumptions and concepts not only in the philosophy of science but also in many other academic disciplines as well. Positivism, which had reigned supreme in Europe since the 1920s, no longer prevailed in philosophy, although there were (and continue to be) residual remnants in many disciplines.[1] In their absorbing account of that period, Edmonds and Eidinow (2001) described the work of the British philosopher A. J. Ayer, who had become an instant celebrity when he popularized positivism in the 1930s. Later asked about its failings, Ayer was quoted as replying: "Well I suppose that the most important of the defects was that nearly all of it was false" (Edmonds & Eidinow, p. 157). Following in the wake of Wittgenstein's (1921/1978) influential work, the very concept of "knowledge" was regarded as dubious. As Russell (1948/1992) explained: "All knowledge is in some degree doubtful, and we cannot say what degree of doubtfulness makes it cease to be knowledge, any more than we can say how much loss of hair makes a man bald" (p. 516). In *Principia Mathematica*, published in 1910–1913, Whitehead and Russell formulated a systematic codification that reduced mathematics to formal logic, suggesting that the arithmetical axioms of mathematical systems were all ultimately provable by logical deduction alone. In 1931, using

[1] Positivism, a philosophical movement inspired by the idea that scientific empiricism was a foundation for all intelligible knowledge (called *verificationism*), initially gained prominence in Auguste Comte's six-volume *Cours de Philosophie Positive* (1830–1842). Comte's work experienced a resurgence of interest in European philosophy beginning in the 1920s, inspired by the earlier philosophical work, the impressive observational developments in natural science, and the periodic discussions of a group of prominent philosophers and scientists (the Vienna Circle). Karl Popper was among those who disputed the positivist position, which he equated with the primitive naive empiricist notion that knowledge of the external world is like "a self-assembling jigsaw puzzle" where the parts take the form of sense experiences that, over time, fit themselves together (Popper, 1972, p. 341). In a famous 1948 lecture, Popper also caricatured that view as the "bucket theory of science" because it reminded him of a container in which raw observations accumulated like patiently and industriously gathered ripe grapes, which, if pressed, inexorably produced the wine of true knowledge (lecture reprinted in Popper, 1972, pp. 341–361). Positivism and related issues that are relevant to ethics and methodology are discussed in our recent book (Rosenthal & Rosnow, 2008, Chapters 1–3, 7) and, in the context of social psychology, in an earlier book (Rosnow, 1981).

an adaptation of Whitehead and Russell's system, Gödel's uncertainty theorems punctured the idea of limitless logical possibilities in what Gauss, a century earlier, famously called "the queen of the sciences," mathematics. Gödel demonstrated that there were axiomatic statements even in mathematics that could be neither proved nor disproved within the system of formal logic (cf. Franzén, 2005; Gödel, 1992).

This thumbnail sketch provides at least a quick look at the big philosophical picture, in which the development of ethical principles in data analysis is the most recent part. In 1945, Vannevar Bush, then Director of the Office of Scientific Research and Development, submitted a report to the President of the United States. The title of the report, "Science the Endless Frontier," reflected an idealization of science that would inevitably be replaced by what Gerald Holton (1978) later called "the notion of science as the suspected frontier" (p. 227). In the human sciences—including biomedical, behavioral, and social science—the illusion of limitless possibilities has metamorphosed into what Holton (1978) called an "ideology of limits" because of cascading ethical mandates regulating the conduct of scientists engaged in research with human subjects. In the following chapters, the authors mention ethical imperatives and codes of conduct in the context of quantitative methodological issues. Continuing with our objective of providing a perspective on these specific discussions, we turn in the next section to the circumstances leading to the development and evolution of the American Psychological Association (APA) code for researchers as a case in point (see APA, 1973, 1982, 1998).

Some major concerns when the idea of a research code for psychologists was initially proposed were the widespread use of deception in certain research areas, the potentially coercive nature of the subject pools used, and the protection of the confidentiality of subjects' responses (Smith, 2000). From the point of view of the most recent iterations of the APA code, it is evident that a wider net of moral concerns has been cast since the APA's 1973 and 1982 *Ethical Principles in the Conduct of Research with Human Participants*. As Sales and Folkman (2000) observed: "Dramatic shifts have taken place in the context in which research occurs … [including] changes in research questions, settings, populations, methods, and societal norms and values" (p. ix). In light of such changes, and the seemingly pervasive distress about moral issues in general, it is not surprising that nearly every facet of research has been drawn into the APA's wider net of concerns, from the statement of a problem, to the research design and its empirical implementation, to the analysis and reporting of data and the conclusions drawn (cf. Sales & Folkman, 2000). Next, we turn to the conventional risk–benefit assessment when proposed research is submitted to a review board for prior approval. After pointing out some limitations of this traditional approach, we describe an alternative model representing the cost–utility assessment of the "doing" and "not doing" of

research. Because the following chapters assume the research is going to be done, and thus focus specifically on whether or not a particular analysis and/or design should be used, we extend our perspective to the cost–utility of adopting versus not adopting particular data analytic or design techniques and reporting practices. Inasmuch as basic technical requirements further guide the investigatory process, there are also occasionally conflicts between technical and ethical standards, but there are often scientific opportunities as well (Blanck, Bellack, Rosnow, Rotheram-Borus, & Schooler, 1992; Rosenthal, 1994; Rosnow, 1997). With those ideas in mind, we conclude by sketching a framework for approaching opportunities for increasing utilities when ethical principles and data analytic standards intersect.

## Moral Sensitivities and the APA Experience

Perhaps the single historical event in modern times that is accountable for galvanizing changes in the way that scientists think about moral aspects of science was World War II. For atomic scientists, Hiroshima was an epiphany that vaporized the idyllic image of the morally neutral scientist, replacing it with a more nuanced stereotype. For scientists engaged in human subject research, the postwar event that set them on an inexorable path to the development of professional codes of conduct was the code drafted in conjunction with expert testimony against Nazi physicians and scientists at the Nuremberg Military Tribunal. Among the tenets of the Nuremberg (or Nuernberg) Code were specific principles of voluntary consent "without the intervention of any element of force, fraud, deceit, duress, over-reaching, or other ulterior form of constraint or coercion"; beneficence and nonmaleficence ("the experiment should be such as to yield fruitful results for the good of society" and "…avoid all unnecessary physical and mental suffering and injury"); and the assessment of the degree of risk to ensure that it "never exceed that determined by the humanitarian importance of the problem to be solved by the experiment" (*Trials of War Criminals Before the Nuernberg Military Tribunals*, 1946–1949, Vol. II, pp. 181–182). To be sure, the kinds of "risks" now appraised by review boards pale in comparison to the brutal "experiments" conducted in the name of science by Nazi physicians on civilian prisoners in concentration camps. Nonetheless, most of the modern guidelines codified into federal regulations can now be understood as having taken their lead from the philosophy of the Nuremberg Code of 1947. For example, as noted in the preamble of the 1979 Belmont Report (discussed below), the Nuremberg Code "became the prototype of many later codes intended to

assure that research involving human subjects would be carried out in an ethical manner."[2]

In 1966, as part of U.S. Public Health Service (USPHS) policy, ethical guidelines were formulated to protect the rights and welfare of human subjects in biomedical research. Three years later, after revelations of shocking instances in which the safety of subjects had been ignored or endangered (see Beecher, 1966), the Surgeon General extended the USPHS safeguards to all human research. A notorious case, not made public until 1972, involved a USPHS study, conducted from 1932 to 1972, of the course of syphilis in more than 400 low-income African-American men in Tuskegee, Alabama (Jones, 1993). The men in the study, recruited from churches and clinics, were not informed they had syphilis but instead were told they had "bad blood." They received free health care and a free annual medical examination, but they were warned that they would be dropped from the study if they sought medical treatment elsewhere, and local physicians were told not to prescribe antibiotics to the men in the study (Fairchild & Bayer, 1999). When the issue of scientific misconduct was publicly aired in hearings conducted in 1973 by the Senate Health Subcommittee (chaired by Senator Edward Kennedy), the time was ripe for an open discussion of misconduct in biomedical research.

Going back to the 1960s again, emotions about invasions of privacy were also running high as a result of publicized reports of domestic wiretapping and other clandestine activities by federal agencies. In sociology, concerns were raised about the legality and morality of the use of unobtrusive observation in field research. In the mid-1960s, sociologists had no code of ethics stipulating the need for informed consent or the subject's right to privacy. In psychology, another sign of the times was a multivolume handbook of more than 4,700 pages, entitled *Psychology: A Study of a Science*, published from 1959 to 1963 under APA sponsorship with funding from the National Science Foundation. The chapters, written by prominent researchers, documented the progress made by psychologists "in attempting to find a way, or ways, to the attainment of the explanatory power that we like to think of as characteristic of science" (Koch, 1959, p. v). Only once was *ethics* cited in the subject indexes: Asch (1959) concluded his chapter by remarking on the need for a "psychology of ethics" (p. 381). A few years later, others who were then caught up in the temper of the times added

[2] Most codes for human subject research now emphasize the use of volunteer subjects who were informed of the nature and risks of the research before deciding whether to participate, the avoidance of significant physical and psychological risks when possible, the use of highly qualified persons to conduct the research, every participant's right to withdraw from the research without penalty, and the scientist's responsibility to terminate the research if there is cause to believe that continuation could cause injury, disability, or death to participants (cf. Kimmel, 1996; Koocher & Keith-Spiegel, 1998; Schuler, 1981).

their voices to a growing debate over ethical issues and the attendant need for a code of conduct in psychological research (cf. Rosnow, 1981, for historical review). The APA appointed a task force to prepare such a code, which was formally adopted by the APA in 1972 (Cook et al., 1971, 1972). In Canada and in Western and Eastern Europe, similar codes drawing on the law, philosophy, and the APA experience were crafted as well (cf. Kimmel, 1996; Schuler, 1981).

There was not much in the way of enforceability or any significant penalties for noncompliance with the professional codes. The penalty for violation of the APA code was censure or expulsion from the APA, but many psychologists who were engaged in productive, rewarding research careers did not belong to the APA. However, by the end of the 1970s, enforceability was no longer in question because *accountability* had become the watchword of the decade (National Commission on Research, 1980). In 1974, the guidelines developed by the U.S. Department of Health, Education, and Welfare 3 years earlier were codified as federal regulations. The National Research Act of July 12, 1974 (Pub. L. 93-348) required institutions that received federal funding to establish institutional review boards (IRBs) for the express purpose of evaluating the risks and benefits of proposed research and monitoring ongoing studies. Also created by this federal act was the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Commission held hearings over a 3-year period and eventually published *The Belmont Report* of April 18, 1979, which set out the basic principles of (a) respect for persons, (b) beneficence, and (c) justice, and their applications in terms of (1) informed consent, (2) the assessment of risks and benefits, and (3) the selection of subjects. Inasmuch as states had legislated their own limits on the information that could be requested of people, and the finding (not in an IRB context, however) of individual biases in ethical decision making (Kimmel, 1991), it is perhaps not surprising that a proposal approved without alterations in one institution might be substantially modified, or even rejected, by a review board at another institution participating in the same research (cf. Ceci, Peters, & Plotkin, 1985).

In 1993, the APA's Board of Scientific Affairs sunset what was then known as the Committee on Standards in Research (CSR), with the idea of delegating the CSR work to a series of task forces. In a final report, the CSR partly attributed the inconsistent implementation of ethical standards by IRBs to the expanded role of such review boards over the previous 2 decades (Rosnow, Rotheram-Borus, Ceci, Blanck, & Koocher, 1993). Noting Kimmel's (1991) findings regarding individual biases in ethical decision making of a group of psychologists, the CSR speculated that some of the inconsistent implementation in IRBs may have resulted as well from the composition of IRBs and raised the possibility of systematically expanding

IRBs to "include people who are sensitive to nuances in the interface of ethical and technical aspects of behavioral and social research" (Rosnow et al., p. 822). The CSR also recommended that IRBs be provided "with a book of case studies and accompanying ethical analyses to sensitize members to troublesome issues and nuances in the behavioral and social sciences" (p. 825). Still, the CSR cautioned that "considerable variability would exist among IRBs in deciding whether issues of design and methodology fall within their responsibility" (pp. 822–823). For researchers, a critical issue is that they may have little recourse to press their claims or to resolve disagreements expediently.

In 1982, the earlier APA code was updated and formally adopted, and there were some changes and phrases ("subject at risk" and "subject at minimal risk") that reflected changes that had occurred in the context in which human subject research was evaluated for prior approval by IRBs. Called by some "the APA's 10 commandments," the 10 major ethical principles of the 1982 APA code are reprinted in Table 3.1. They are of interest not only historically but also because subsequent iterations of the APA ethics code echo the sensitivities in the 10 principles in Table 3.1. All the same, given the precedence of federal and state statutes and regulations, psychological researchers (whether they were APA members or not) were probably more likely to take their ethical cues from the legislated morality and its enforcement by IRBs. In the late 1980s, there was a fractious splintering of the APA, which resulted in the creation of the rival American Psychological Society, now called the Association for Psychological Science (APS). For a time in the 1990s, a task force cosponsored by APA and APS attempted to draft an updated code, but the APS withdrew its collaboration after an apparently irresolvable disagreement. In 2002, after a 5-year revision process, the APA adopted a reworked code emphasizing the following five general principles: (a) beneficence and nonmaleficence, (b) fidelity and responsibility, (c) integrity, (d) justice, and (e) respect for people's rights and dignity.[3] The tenor of this revised code of conduct reflects both the majority practitioner constituency of APA as well as the constituency of psychological scientists in APA and the wide variety of contexts in which research is conducted.

Still, there are relentless conflicts between specific ethical and scientific standards, which have become a source of consternation for some researchers. For example, principle (c) above (integrity) calls for "accuracy, honesty, and truthfulness in the science, teaching, and practice of psychology."

[3] The full code, including "specific standards" that flesh out each of the five general principles, was published in the *American Psychologist* in 2002 (*57*, 1060–1073). More recent changes and itemized comparisons are available at http://apa.org by searching "ethics code updates."

**TABLE 3.1**

Ethical Standards Adopted by the American Psychological Association in 1982: Research With Human Participants[a]

The decision to undertake research rests on a considered judgment by the individual psychologist about how best to contribute to psychological science and human welfare. Having made the decision to conduct research, the psychologist considers alternative directions in which research energies and resources might be invested. Based on this consideration, the psychologist carries out the investigation with respect and concern for the dignity and welfare of the people who participate and with cognizance of federal and state regulations and professional standards governing the conduct of research with human participants.

A. In planning a study, the investigator has the responsibility to make a careful evaluation of its ethical acceptability. To the extent that the weighing of scientific and human values suggests a compromise of any principle, the investigator incurs a correspondingly serious obligation to seek ethical advice and to observe stringent safeguards to protect the rights of human participants.

B. Considering whether a participant in a planned study will be a "subject at risk" or a "subject at minimal risk," according to recognized standards, is of primary ethical concern to the investigator.

C. The investigator always retains the responsibility for ensuring ethical practice in research. The investigator is also responsible for the ethical treatment of research participants by collaborators, assistants, students, and employees, all of whom, however, incur similar obligations.

D. Except in minimal-risk research, the investigator establishes a clear and fair agreement with research participants, before their participation, that clarifies the obligations and responsibilities of each. The investigator has the obligation to honor all promises and commitments included in that agreement. The investigator informs the participants of all aspects of the research that might reasonably be expected to influence willingness to participate and explains all other aspects of the research about which the participants inquire. Failure to make full disclosure before obtaining informed consent requires additional safeguards to protect the welfare and dignity of the research participants. Research with children or with participants who have impairments that would limit understanding and/or communication requires special safeguarding procedures.

E. Methodological requirements of a study may make the use of concealment or deception necessary. Before conducting such a study, the investigator has a special responsibility to (a) determine whether the use of such techniques is justified by the study's prospective scientific, educational, or applied value; (b) determine whether alternative procedures are available that do not use concealment or deception; and (c) ensure that the participants are provided with sufficient explanation as soon as possible.

F. The investigator respects the individual's freedom to decline to participate in or to withdraw from the research at any time. The obligation to protect this freedom requires careful thought and consideration when the investigator is in a position of authority or influence over the participant. Such positions of authority include, but are not limited to, situations in which research participation is required as part of employment or in which the participant is a student, client, or employee of the investigator.

G. The investigator protects the participant from physical and mental discomfort, harm, and danger that may arise from research procedures. If risks of such consequences exist, the investigator informs the participant of that fact. Research

**TABLE 3.1 (Continued)**

Ethical Standards Adopted by the American Psychological Association in 1982: Research With Human Participants[a]

procedures likely to cause serious or lasting harm to a participant are not used unless the failure to use these procedures might expose the participant to risk of greater harm or unless the research has great potential benefit and fully informed and voluntary consent is obtained from each participant. The participant should be informed of procedures for contacting the investigator within a reasonable period after participation should stress, potential harm, or related questions or concerns arise.

H. After the data are collected, the investigator provides the participant with information about the nature of the study and attempts to remove any misconception that may have arisen. Where scientific or human values justify delaying or withholding this information, the investigator incurs a special responsibility to monitor the research and to ensure that there are no damaging consequences for the participant.

I. Where research procedures result in undesirable consequences for the individual participant, the investigator has the responsibility to detect and remove or correct these consequences, including long-term effects.

J. Information obtained about a research participant during the course of an investigation is confidential unless otherwise agreed upon in advance. When the possibility exists that others may obtain access to such information, this possibility, together with the plans for protecting confidentiality, is explained to the participant as part of the procedure for obtaining informed consent.

[a] Copyright © 1982 by the American Psychological Association. Reproduced [or Adapted] with permission. American Psychological Association. (1982). *Ethical principles in the conduct of research with human participants* (pp. 5–7). Washington, DC: No further reproduction or distribution is permitted without written permission from the American Psychological Association.

However, the use of (active or passive) deception in certain research is frequently viewed by social psychological researchers as essential to ensure the scientific integrity of the results (cf. Behnke, 2009; Kimmel, 1998). The issue of deception is further complicated by the fact that active and passive deceptions are far from rare in our society. Trial lawyers often manipulate the truth in court on behalf of clients; prosecutors surreptitiously record private conversations; journalists often get away with using hidden cameras and undercover practices to get their stories; and police investigators use sting operations and entrapment procedures to gather incriminating information (Bok, 1978, 1984; Saxe, 1991; Starobin, 1997). Not that the use of deception in society excuses or justifies its use in research, but the fact is that applications of the APA principles often require that researchers read between the lines (e.g., Behnke, 2009).

Later, we mention scandalous examples of the deceptive marketing of pharmaceuticals to doctors and patients on the basis of biased data or the withholding of critical research data (see Spielmans & Parry, 2010). Deceptive practices like these are a minefield of immediate problems in evidence-based medicine and of potential problems for clinical psychologists who lobby for the option of training and subsequent authorization for prescription privileges.

## Weighing the Costs and Utilities of Doing and Not Doing Research

We turn now to the risk–benefit appraisal process. The Belmont Report defined the term *risk* as "a possibility that harm may occur." The report also referred to "small risk" rather than "low risk," while noting that expressions such as "small risk," "high risk," and "balanced risks and benefits" were typically used metaphorically rather than precisely. Nonetheless, the risk–benefit ideal that was advocated in the report was to consider enough information to make the justifiability of doing the research as precise and thorough as possible (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Figure 3.1 is an idealized representation of the usual risk–benefit assessment (Rosnow & Rosenthal, 1997), where the predicted risk of doing the research is plotted from low (C) to high (A) on the vertical axis and the predicted benefit of the research is plotted from low (C) to high (D) on the horizontal axis. Studies falling closer to A are unlikely to be approved; studies falling closer to D are likely to be approved; and studies falling along the B–C diagonal of indecision may be too difficult to decide without more information. On the assumption that methodological quality in general is ethically as well as technically relevant in assessing research (see Mark & Lenz-Watson, Chapter 7, this volume), it follows that studies that are well thought out (technically speaking), that are low risk, and that address important questions will be closer to D, whereas studies that are not well thought out, that present risks, and that address trivial questions will be closer to A.[4]

We have become convinced that the conventional risk–benefit model is insufficient because it fails to consider the costs (and utilities) of *not* conducting a specific study (Rosenthal & Rosnow, 1984, 2008). Suppose a review board rejected a research proposal for a study of an important health topic because there was no guarantee that the privacy of the participants would be protected. On the other side, the researchers argue that there is no alternative design that does not compromise the validity of the results. Depriving the community of valid scientific information with

[4] Review boards deal first and foremost with proposed studies, but their responsibility to monitor ongoing research suggests that they almost certainly judge risks and benefits in real time as well. Uncovering "research misconduct" is an extreme instance, defined by the Office of Research Integrity (ORI) as "fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results" (http://ori.dhhs.gov). A recent case involved an anesthesiologist who fabricated the results that he reported in 21 journal articles about clinical trials of pain medicine that was marketed by the very company that funded much of the doctor's research (Harris, 2009). The fabrication of data in biomedical research is high risk not only because it jeopardizes the treatment of future patients, but it also undermines the credibility of the journal literature on which the advancement of medical science depends.

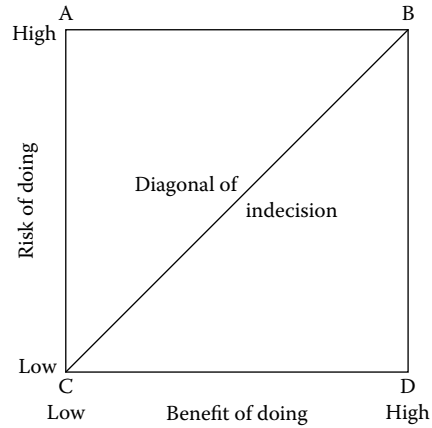



**FIGURE 3.1**
Decision-plane model representing the relative risks and benefits of research submitted to a review board for approval. (After Rosnow, R. L., & Rosenthal, R., *People Studying People: Artifacts and Ethics in Behavioral Research*, W. H. Freeman, New York, 1997.)

which to address the important health problem did not make the ethical issue disappear. It merely traded one ethical issue for another, and the cost in human terms of the research not done could also be high. Thus, we have proposed the alternative models shown in Figure 3.2 (Rosenthal & Rosnow, 1984). In Figure 3.2a, the decision-plane model on the left represents a cost–utility appraisal of the *doing of research*, and the decision-plane model on the right represents a cost–utility appraisal of the *not doing of research*. We refer to *costs* and *utilities* in a collective sense, where the *collective cost* is that incurred by the subjects in the study, by other people, and by wasted funds, wasted time, and so on, and the *collective utility* refers to the benefits (e.g., medical, psychological, educational) accruing to the subjects, to other people in the future, to the researchers, society, science, and so on. Figure 3.2b shows a composite model that reduces the three dimensions of Figure 3.2a back to two dimensions. Imagine an A–D "decision diagonal" in each of the decision-planes in Figure 3.2a (in contrast to B–C and B′–C′, the diagonals of indecision). For any point in the plane of "doing," there is a location on the cost axis and on the utility axis. In the composite model in Figure 3.2b, points near D tell us that the research *should* be done, and points near D′ tell us the research should *not* be done.[5]

As a data analytic application of the underlying reasoning in Figure 3.2, consider the options that researchers mull over when confronted with outliers and are trying to figure out what to do next. Having identified

[5] Adaptations of the models in Figures 3.1 and 3.2 have been used in role-playing exercises to cue students to ethical dilemmas in research and data analysis (Bragger & Freeman, 1999; Rosnow, 1990; Strohmetz & Skleder, 1992).

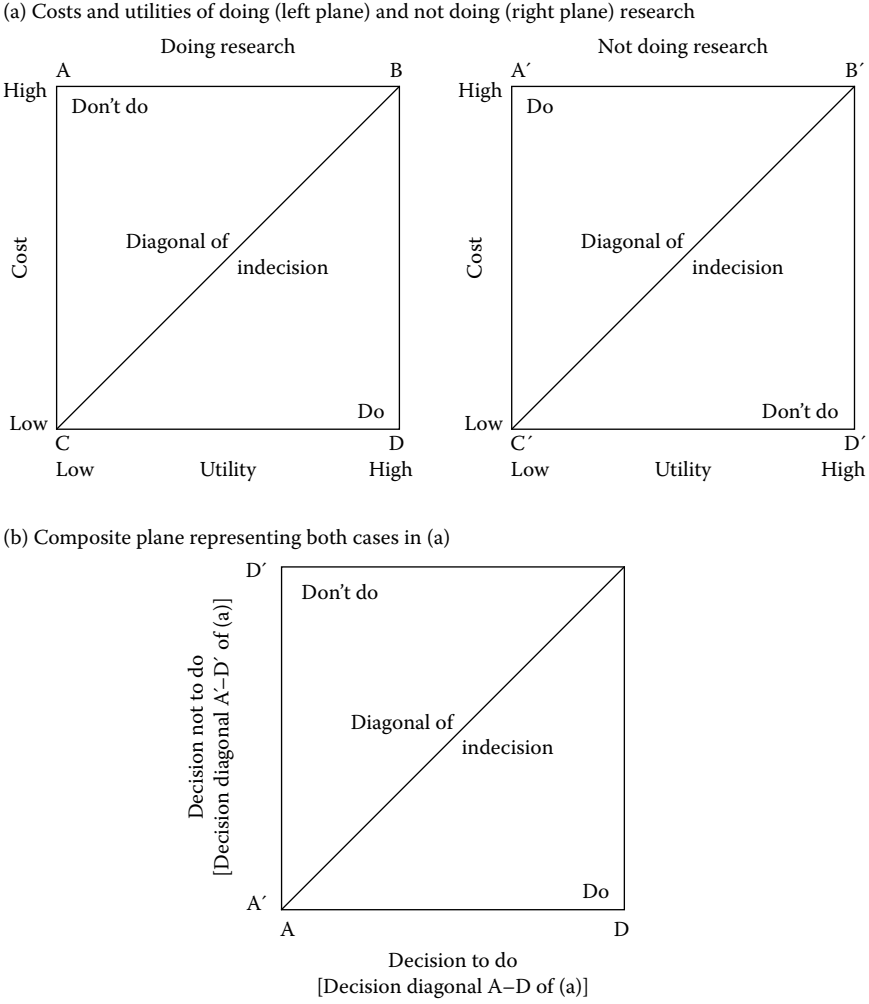



**FIGURE 3.2**
Decision planes representing the ethical assessment of the costs and utilities of doing and not doing research. (a) Costs and utilities of doing (left plane) and not doing (right plane) research. (b) Composite plane representing both cases in (a). (After Rosenthal, R., & Rosnow, R. L., *Am. Psychol.*, 45, 775, 1984 and Rosenthal, R., & Rosnow, R. L., *Essentials of Behavioral Research: Methods and Data Analysis*, 3rd ed., McGraw-Hill, New York, 2008.)

all the outliers (e.g., Iglewicz & Hoaglin, 1993), an option that they may wrestle with is "not to keep the outliers in the data set," which is analogous to the "not doing of research." For example, they may be tempted just to drop the outliers, but this is likely to result in a biased index of central tendency for the remaining data when knowing the mean or the median is important. Alternatively, they might be thinking about using

equitable trimming, which is a less biased procedure than dropping the outliers and usually works well when the sample sizes are not very small. Another possibility might be to "reel in" the outliers by finding a suitable transformation to pull in the outlying stragglers and make them part of the group. For example, Tukey (1977) described some common transformations for pulling in scores that are far out, such as the use of square roots, logarithms, and negative reciprocals. An outlier that is an error (e.g., a scoring or recording mistake) may be dealt with by not keeping it in the data set, but outliers that are not errors are a signal to look further into the data. Until one knows whether outliers are errors, any indecision about how to deal with them cannot be confidently resolved. Once we know for sure that any outliers are not errors, instead of thinking of them as a "nuisance," we might instead think of them as an opportunity to unlock the analysis of the data with the objective of searching for a plausible moderator variable for further investigation. As Tukey (1977) put it, "To unlock the analysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act" (p. viii). In this illustration, the outliers are the key, and the creative act will be a combination of exploratory data analysis, a keen eye, and an open mind.

## When Ethical and Technical Standards Crisscross

In this final discussion, we sketch a framework for thinking some more about cost–utility implications when ethical and technical standards intersect in quantitative methodology. Other frameworks are discussed in this volume, whereas here we approach this problem from the perspective of the matrix shown in Table 3.2. When there is a conflict between ethical and technical standards in science, the scientist must resolve the

**TABLE 3.2**

Five General Ethical Standards Crossed by Five Data Analytic and Reporting Standards in Quantitative Methodology

| | Data Analytic and Reporting Standards | | | | |
|---|---|---|---|---|---|
| **Ethical Standards** | **1. Transparent** | **2. Informative** | **3. Precise** | **4. Accurate** | **5. Grounded** |
| A. Beneficence | | | | | |
| B. Nonmaleficence | | | | | |
| C. Justice | | | | | |
| D. Integrity | | | | | |
| E. Respect | | | | | |

conflict because scientists are accountable for both the ethical and scientific merit of their work (cf. Scott-Jones & Rosnow, 1998). Conflicts may provide opportunities to expand knowledge and develop a stronger science, which implies that the resolution of ethical conflicts can serve a scientific and a moral purpose (Blanck et al., 1992; Rosenthal, 1994; Rosnow, 1997). In the remaining discussion, we briefly define the ethical standards in the rows and the technical standards in the columns of Table 3.2 and then give two illustrations of a way in which we can increase utilities in primary and secondary data analyses. The ethical and technical standards in Table 3.2 should not be viewed as exhaustive or mutually exclusive, but rather as illustrative of contemporary values in our cultural sphere. The five ethical standards are an amalgamation of certain ideals enunciated in the Belmont Report, the general principles of the 2002 APA code, and the 1999 code of the American Statistical Association (ASA, 1999).[6]

Starting with the five row headings, *beneficence* is the aspirational ideal to do good. Although it is generally conceded that the mere intention to do good will not always have a beneficial outcome, it is an ethical objective to aim for nonetheless. Second, *nonmaleficence* is the concomitant obligation to "do no harm," as in the Hippocratic Oath that physicians take and, of course, as in many other areas as well. For example, Cizek and Rosenberg (Chapter 8, this volume) discuss the potentially harmful consequences of the misuse of psychometrics in "high stakes assessment" test results, and the ASA code cautions us about the harm that can result from false or misleading statistics in medicine. Third, *justice* was defined in the Belmont Report as "fairness in distribution," implying that the burdens and benefits of research should be distributed equitably. In the Tuskegee study, none of the men could have benefited in any way, and they alone bore the appalling burdens as well. In the 2002 APA code, justice is equated with the exercise of reasonable judgment and with ensuring that potential biases, the boundaries of one's competence, and the limitations of one's expertise "do not lead to or condone unjust practices." Fourth, by *integrity* we mean the character of a particular action or claim (e.g., the reporting of data), as well as the honesty and soundness of those responsible or who take credit for the action or claim, is free from moral corruption. In the 2002 APA ethics code, integrity is equated with "accuracy [also a technical standard in Table 3.2], honesty, and truthfulness in the science, teaching, and practice of psychology." Integrity further implies the prudent use of research funding and other resources and, of course, the disclosure of any conflicts of interest, financial or otherwise, so as not to betray public trust. Fifth, in the Belmont Report, *respect* was stated to incorporate "at

[6] The ASA code, *Ethical Guidelines for Statistical Practice*, approved August 7, 1999 by the ASA Board of Directors, is available from http://www.amstat.org/committees/ethics/index.cfm

least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection." In the APA code, respect is equated with civil liberties: "privacy, confidentiality, and self-determination."

Turning to the column headings in Table 3.2, first, by *transparent*, we mean that the presentation of statistical results is open, frank, and candid, that any technical language used is clear and appropriate, and that visual displays are crystal clear.[7] Second, by *informative*, we mean there is enough information reported. How much information is sufficient? There should be enough for readers to make up their own minds based on the primary results or by performing secondary analyses using the summary results reported (see, e.g., Rosenthal, 1995; Rosenthal & DiMatteo, 2001; Rosnow & Rosenthal, 1995, 1997, 2008; Wilkinson & Task Force on Statistical Inference, 1999). Third, the term *precise* is used here not in a statistical sense (i.e., the likely spread of estimates of a parameter) but in a more general sense to mean that quantitative results should be reported to the degree of exactitude required by the given situation, which in many cases will require striking a balance between being vague and being needlessly or falsely precise.[8] Fourth, by *accurate*, we mean a conscientious effort is made to identify and correct any mistakes in measurements, calculations, or the reporting of numbers. Accuracy also means not exaggerating results by, for example, making claims that future applications of the results are unlikely to achieve (e.g., misleading claims of causal relationships where none have been established by the data). Fifth, by *grounded*, we mean that the methodology is logically and scientifically justified, the question addressed is appropriate to the design, and the data analysis addresses the question of interest as opposed to going off on a tangent or mindlessly having a computer program frame the question (see also Rosenthal, Rosnow, & Rubin, 2000; Rosnow & Rosenthal, 1989).

The five column headings of Table 3.2, the five data analytic and reporting standards, tend to come precorrelated in the real world of quantitative research methodology. Research results reported clearly (transparent) tend to give readers more of the information needed to understand the findings (informative), reported with appropriate exactness (precision), with quantitative and interpretive errors minimized (accuracy), and with design and analysis both appropriate to the conclusions drawn (grounded). When a particular investigation meets these standards to a high degree, there is a better chance that the research will "do more good"

[7] Basic elements of graph design have been illustrated elegantly by Tufte (2001) and Wainer (1984, 2000, 2009) and also explored insightfully by Kosslyn (1994) from the perspective of how the brain processes visual information.

[8] For example, reporting the scores on an attitude questionnaire to a high level of decimal places is psychologically meaningless (false precision), and reporting the weight of mouse subjects to six decimal places is pointless (needless precision).

(beneficence) and will do less harm (nonmaleficence), for example, by having identified subsets of a research sample that are harmed by an intervention that helps most participants. Such high-standards studies are also more likely to be fair when assigning participants to conditions at random, for example, to treatment versus wait-list/control (justice), and when results are reported with honesty (integrity). Finally, research conducted with higher standards of data analysis and reporting treats participants more respectfully by not wasting their time by conducting inferior data analyses and reporting (respect).

An example of an opportunity to increase utilities in the primary data analysis involves what Rosenthal (1994) described as "snooping around in the data," an idea we alluded to earlier when we spoke of coping with outliers. For a time, researchers and data analysts were taught that it is technically improper, and maybe even immoral, to analyze and reanalyze the data in many ways (i.e., to snoop around in the data). Assess the prediction with one preplanned statistical test, the students were told, and if the result turns out not to be significant at the .05 level, do not look any further at the data. On the contrary, snooping around in the data (sometimes referred to as data mining) is technically advisable (in the sense of being transparent, informative, precise, accurate, and well grounded), and it can be a way of increasing the utility of data insofar as it is likely to turn up something new, interesting, and important (e.g., Hoaglin, Mosteller, & Tukey, 1983; Tukey, 1977). Data are costly in terms of time, effort, money, and other resources. The antisnooping dogma makes for bad ethics because it is a wasteful consumption that betrays the public trust that scientists will be prudent in their use of funding, material, people's time, energy, and other resources. If the data were worth collecting in the first place, then they are worth a thorough analysis, being held up to the light in many different ways so that everyone who contributed (research subjects, a funding agency, science, and society) will get their time and their money's worth (Rosenthal, 1994). It is true, of course, that snooping around in the data can affect the $p$ values obtained, but we can use well-grounded statistical adjustments to deal with this problem, in case getting a more accurate $p$ value is important to the investigator. Readers can find a more detailed treatment of such statistical adjustments in Hubert and Wainer (Chapter 4, this volume). If no adjustment is made for $p$ values computed post hoc, we can all agree that replication will be required. Of course, replications are important even beyond this requirement. Replications that are very similar to the original design will, if successful, increase our confidence in the stability of the original finding, whereas replications that systematically vary a fundamental aspect of the original design (e.g., operationalization of a dependent variable or varying the population sampled) will, if successful, extend the generalizability of the original findings (cf. Rosenthal, 1990).

Having data to wisely, openly, and honestly snoop around in secondary data analyses would also address the problem that data sets are not always accessible, although there is now the requirement of federal agencies that large, expensive grants over a certain amount per year must develop plans for data archiving, so that others can reanalyze the data and benefit. But what about studies that are funded by a drug manufacturer that feels under no obligation to publish negative results? In a carefully documented article (including pictures of internal documents and e-mail), Spielmans and Parry (2010) elaborated on a litany of problems in the pharmaceutical industry's marketing of medicines. As stated by Spielmans and Parry: "One could argue that rather than [evidence-based medicine], we are actually now entrenched in marketing-based medicine, in which science has largely been taken captive in the name of increasing profits for pharmaceutical firms" (p. 13). A cited example was the publication of supposedly positive results that turned out to be exaggerations, but the exaggerated results were blithely disseminated as part of the company's marketing strategy. Another example that was in the news recently was the revelation that numerous articles in leading medical journals had not been written by the bylined authors, but rather had been drafted by "hidden writers" who were employed by pharmaceutical and medical device companies to promote their company products (Singer & Wilson, 2009; Wilson & Singer, 2009). In an editorial in the online journal *PLoS Medicine*, the editors asked: "How did we get to the point that falsifying the medical literature is acceptable? How did an industry whose products have contributed to astounding advances in global health over the past several decades come to accept such practices as the norm?" The editorial exhorted journal editors to identify and retract any ghostwritten articles and banish their authors (*PLoS Medicine* Editors, 2009).

There are ways of increasing the collective utility of data retroactively, and a good example is by the use of meta-analysis (Rosenthal, 1994). The collective cost of time, attention, and effort of the human participants in the individual studies are all more justified when the data are entered into a meta-analysis. (For a discussion of what should and should not be reported in meta-analyses, see Cooper and Dent, Chapter 16, this volume.) Other costs of the individual studies (e.g., funding, supplies, space, investigator time and effort, and other resources) are also more justified because the utility of individual studies is so increased by the borrowed strength obtained when the data from more studies are summed up and explored in a sophisticated quantitative way. Conversely, not using meta-analytic procedures where they could be used has ethical implications because the opportunity to increase the benefits of past individual studies is relinquished. This means not simply providing overall estimates of the size of pooled effects, but also looking for moderator variables and plausible explanations of the inevitable variation in the size of effects obtained in different studies. Indeed, from

both the scientific and ethical perspectives, it no longer seems acceptable even to fund research studies that claim to contribute to the resolution of controversy (e.g., does Treatment A work?) unless the researcher has already conducted a meta-analysis showing that there is a meaningful disagreement and not an illusory controversy based on a straw man argument. In some situations, meta-analyses resolve illusory controversies by eliminating two common problems in the evaluation of replications. One problem is the myth that when one study obtains a statistically significant effect and a replication does not, this indicates "a failure to replicate." However, a failure to replicate is properly measured by the magnitude of difference between the effect sizes of the two studies, accompanied by a confidence interval estimate. The second problem that meta-analyses can often eliminate is the naive belief that if there is a real effect in a situation, each contributing study will show a statistically significant effect (Rosenthal, 1994).

## Conclusion

Once upon a time, it was thought that science was "morally neutral" by its very nature because the moment that science starts sorting facts into "good ones" and "bad ones" it is no longer science. How curious that illusion now seems. Nowadays, every aspect of human endeavor, including science, is viewed not as morally neutral, but as fed by a wellspring of values, biases, motives, and goals, which in turn are infused with illusions and self-delusions. We are reminded of what the philosopher Abraham Kaplan (1964) called the "law of the instrument": Give a small boy a hammer and he will find that everything he encounters needs pounding (pp. 28–29). Given the burgeoning growth of ethical mandates, it is hardly surprising that virtually every facet of science is seen in some quarters as in need of moral pounding. We believe there is another way of looking at it that is more hopeful and exciting. The boundaries imposed by moral considerations can also become new horizons that challenge us to see opportunities in quantitative methodology for new explorations and new scientific triumphs when ethical and technical standards intersect.

## References

American Psychological Association. (1973). *Ethical principles in the conduct of research with human participants*. Washington, DC: Author.

American Psychological Association. (1982). *Ethical principles in the conduct of research with human participants*. Washington, DC: Author.

American Psychological Association. (1998). *The ethics of research with human participants* (draft report). Washington, DC: Author.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author. Retrieved from http://www.apa.org/ethics.code/index.aspx

American Statistical Association. (1999). *The ASA code, Ethical Guidelines for Statistical Practice*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/committees/ethics/index.cfm

Asch, S. E. (1959). A perspective on social psychology. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3, pp. 363–383). New York: McGraw-Hill.

Beecher, H. K. (1966, July 2). Documenting the abuses. *Saturday Review,* pp. 45–46.

Behnke, S. (2009). *Ethics rounds: Reading the ethics code more deeply.* Retrieved from http://www.apa.org/monitor/2009/04/ethics.html

Blanck, P. D., Bellack, A. S., Rosnow, R. L., Rotheram-Borus, M. J., & Schooler, N. R. (1992). Scientific rewards and conflicts of ethical choices in human subjects research. *American Psychologist, 47,* 959–965.

Bok, S. (1978). *Lying: Moral choice in public and private life.* New York: Pantheon.

Bok, S. (1984). *Secrets: On the ethics of concealment and revelation*. New York: Vintage Books.

Bragger, J. D., & Freeman, M. A. (1999). Using a cost-benefit analysis to teach ethics and statistics. *Teaching of Psychology, 26,* 34–36.

Ceci, S. J., Peters, D., & Plotkin, J. (1985). Human subjects review, personal values, and the regulation of social science research. *American Psychologist, 40,* 994–1002.

Cook, S. W., Hicks, L. H., Kimble, G. A., McGuire, W. J., Schoggen, P. H., & Smith, M. B. (1972, May). Ethical standards for research with human subjects. *APA Monitor,* I–XIX.

Cook, S. W., Kimble, A., Hicks, L. H., McGuire, W. J., Schoggen, P. H., & Smith, M. B. (1971, July). Ethical standards for psychological research: Proposed ethical principles submitted to the APA membership for criticism and modification (by the ad hoc) Committee on Ethical Standards in Psychological Research. *APA Monitor,* 9–28.

Edmonds, D., & Eidinow, J. (2001). *Wittgenstein's poker: The story of a ten-minute argument between two great philosophers*. New York: HarperCollins.

Fairchild, A. L., & Bayer, R. (1999). Uses and abuse of Tuskegee. *Science, 284,* 918–921.

Franzén, T. (2005). *Gödel's theorem: An incomplete guide to its use and abuse.* Wellesley, MA: A. K. Peters.

Gödel, K. (1992). *On formally undecidable propositions of principia mathematica and related systems.* New York: Dover. (Originally written in German and published in an Austrian scientific journal in 1931.)

Harris, G. (2009, March 10). *Doctor admits pain studies were frauds, hospital says.* Retrieved from http://www.nytimes.com/2009/03/11/health/research/11pain.html?ref=us

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). *Understanding robust and exploratory data analysis.* New York: Wiley.

Holton, G. (1978). From the endless frontier to the ideology of limits. In G. Holton & R. S. Morison (Eds.), *Limits of scientific inquiry* (pp. 227–241). New York: Norton.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers.* Milwaukee, WI: ASQC Quality Press.

Jones, H. H. (1993). *Bad blood: The Tuskegee syphilis experiment* (Rev. ed.). New York: Free Press.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* Scranton, PA: Chandler.

Kimmel, A. J. (1991). Predictable biases in the ethical decision making of American psychologists. *American Psychologist, 46,* 786–788.

Kimmel, A. J. (1996). *Ethical issues in behavioral research: A survey*. Cambridge, MA: Blackwell.

Kimmel, A. J. (1998). In defense of deception. *American Psychologist, 53,* 803–805.

Koch, S. (Ed.). (1959). *Psychology: A study of a science* (Vol. 1, pp. v–vii). New York: McGraw-Hill.

Koocher, G. P., & Keith-Spiegel, P. C. (1998). *Ethics in psychology* (2nd ed.). Washington, DC: American Psychological Association.

Kosslyn, S. M. (1994). *Elements of graph design.* New York: W. H. Freeman.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* Retrieved from http://www.fda.gov/ohrms/dockets/ac/05/briefing/2005-4178b_09_02_Belmont%20Report.pdf

National Commission on Research. (1980). Accountability: Restoring the quality of the partnership. *Science, 207,* 1177–1182.

*PLoS Medicine* Editors (2009, September). Ghostwriting: The dirty little secret of medical publishing that just got bigger. *PLoS Medicine, 6,* e1000156. Retrieved from http://www.plosmedicine.org/static/ghostwriting.action

Polanyi, M. (1966). *The tacit dimension.* New York: Doubleday Anchor.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach.* Oxford, UK: Oxford University Press.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality, 5,* 1–30.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5,* 127–134.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Rosenthal, R., & Rosnow, R. L. (1984). Applying Hamlet's question to the ethical conduct of research. *American Psychologist, 45,* 775–777.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York: McGraw-Hill.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge, UK: Cambridge University Press.

Rosnow, R. L. (1981). *Paradigms in transition: The methodology of social inquiry*. New York: Oxford University Press.

Rosnow, R. L. (1990). Teaching research ethics through role-play and discussion. *Teaching of Psychology, 17,* 179–181.

Rosnow, R. L. (1997). Hedgehogs, foxes, and the evolving social contract in psychological science: Ethical challenges and methodological opportunities. *Psychological Methods, 2,* 345–356.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Rosnow, R. L., & Rosenthal, R. (1995). "Some things you learn aren't so": Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science, 6,* 3–9.

Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: W. H. Freeman.

Rosnow, R. L., & Rosenthal, R. (2008). Assessing the effect size of outcome research. In A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research* (pp. 379–401). New York: Oxford University Press.

Rosnow, R. L., Rotheram-Borus, M. J., Ceci, S. J., Blanck, P. D., & Koocher, G. P. (1993). The institutional review board as a mirror of scientific and ethical standards. *American Psychologist*, *48*, 821–826.

Russell, B. (1992). *Human knowledge: Its scope and limits.* London: Routledge. (Original work published in 1948.)

Sales, B. D., & Folkman, S. (Eds.). (2000). *Ethics in research with human participants.* Washington, DC: American Psychological Association.

Sambursky, S. (Ed.). (1974). *Physical thought from the presocratics to the quantum physicists: An anthology.* New York: Pica Press.

Saxe, L. (1991). Lying: Thoughts of an applied social psychologist. *American Psychologist, 46,* 409–415.

Schuler, H. (1981). Ethics in Europe. In A. J. Kimmel (Ed.), *Ethics of human subject research* (pp. 41–48). San Francisco: Jossey-Bass.

Scott-Jones, D., & Rosnow, R. L. (1998). Ethics and mental health research. In H. Friedman (Ed.), *Encyclopedia of mental health* (Vol. 2, pp. 149–160). San Diego: Academic Press.

Singer, N., & Wilson, D. (2009, September 18). Unmasking the ghosts: Medical editors take on hidden writers. *The New York Times,* pp. B1, B5.

Smith, M. B. (2000). Moral foundations in research with human participants. In B. D. Sales & S. Folkman (Eds.). *Ethics in research with human participants* (pp. 3–9). Washington, DC: American Psychological Association.

Spielmans, G. I., & Parry, P. I. (2010). From evidence-based medicine to marketing-based medicine: Evidence from internal industry documents. *Bioethical Inquiry, 7,* 13–29. doi:10.1007/s11673-010-9208-8

Starobin, P. (1997, January 28). Why those hidden cameras hurt journalism. *The New York Times,* p. A21.

Strohmetz, D. B., & Skleder, A. A. (1992). The use of role-play in teaching research ethics: A validation study. *Teaching of Psychology, 19,* 106–108.

*Trials of war criminals before the Nuernberg military tribunals under control council law no. 10,* October 1946–April 1949, Vol. II. Washington, DC: U.S. Government Printing Office.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1984). How to display data badly. *American Statistician, 38,* 137–147.

Wainer, H. (2000). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* Mahwah, NJ: Erlbaum.

Wainer, H. (2009). *Picturing the uncertain world*. Princeton, NJ: Princeton University Press.

Whitehead, A. N., & Russell, B. (1962). *Principia mathematica to *56.* Cambridge, UK: Cambridge University Press. (Original work published in three volumes in 1910, 1912, and 1913.)

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 595–604.

Wilson, D., & Singer, J. (2009, September 11). Study says ghostwriting rife in medical journals. *The New York Times,* p. B5.

Wittgenstein, L. (1978). *Tractatus logico-philosophicus*. London: Routledge & Kegan Paul. (Original work, titled *Logisch-Philosophische Abhandlung*, published 1921 in the German periodical *Annalen der Naturphilosophie*.)

# Section II

# Teaching Quantitative Ethics

# 4

# *A Statistical Guide for the Ethically Perplexed*

**Lawrence Hubert**
*University of Illinois at Urbana-Champaign*

**Howard Wainer**
*National Board of Medical Examiners*

The meaning of "ethical" adopted here is one of being in accordance with the accepted rules or standards for right conduct that govern the practice of some profession. The professions we have in mind are statistics and the behavioral sciences, and the standards for ethical practice are what we try to instill in our students through the methodology courses we offer, with particular emphasis on the graduate statistics sequence generally required for all the behavioral sciences. Our hope is that the principal general education payoff for competent statistics instruction is an increase in people's ability to be critical and ethical consumers and producers of the statistical reasoning and analyses faced in various applied contexts over the course of their careers.

Generations of graduate students in the behavioral and social sciences have completed mandatory year-long course sequences in statistics, sometimes with difficulty and possibly with less than positive regard for the content and how it was taught. Before the 1960s, such a sequence usually emphasized a cookbook approach where formulas were applied unthinkingly using mechanically operated calculators. The instructional method could be best characterized as "plug and chug," where there was no need to worry about the meaning of what one was doing, only that the numbers could be put in and an answer generated. It was hoped that this process would lead to numbers that could then be looked up in tables; in turn, $p$ values were sought that were less than the magical .05, giving some hope of getting an attendant paper published.

The situation began to change for the behavioral sciences in 1963 with the publication of *Statistics for Psychologists* by William Hays. For the first time, graduate students could be provided both the needed recipes and

some deeper understanding of and appreciation for the whole enterprise of inference in the face of uncertainty and fallibility. Currently the Hays text is in its fifth edition, with a shortened title of *Statistics* (1994); the name of Hays itself stands as the eponym for what kind of methodology instruction might be required for graduate students, that is, at the level of Hays, and cover to cover. Although now augmented by other sources for related computational work (e.g., by SAS, SPSS, or SYSTAT), the Hays text remains a standard of clarity and completeness. Many methodologists have based their teaching on this resource for more than four decades. Hays typifies books that, although containing lucid explanations of statistical procedures, are too often used by students only as a cookbook of statistical recipes. The widespread availability of statistical software has made it clear that we no longer have a need for cookbooks and instead require a *Guide to Gastronomy*.

In teaching graduate statistics, there are multiple goals:

1. to be capable of designing and analyzing one's own studies, including doing the computational "heavy lifting" by one's self, and the ability to verify what others attached to a project may be doing;
2. to understand and consume other research intelligently, both in one's own area, but more generally as a statistically and numerically literate citizen;
3. to argue for and justify analyses when questioned by journal and grant reviewers or others, and to understand the basic justification for what was done.

For example, an ability to reproduce a formal proof of the central limit theorem is unnecessary, but a general idea of how it is formulated and functions is relevant, as well as that it might help justify assertions of robustness being made for the methods used. These skills in understanding are not "theoretical" in a pejorative sense, although they do require more thought than just being content to run the SPSS machine blindly. They are absolutely crucial in developing both the type of reflective teaching and research careers we would hope to nurture in graduate students, and more generally for the quantitatively literate citizenry we would wish to make up our society.

Graduate instruction in statistics requires the presentation of general frameworks and how to reason from these. These frameworks can be conceptual: for example, (a) the Fisherian view that provided the evidence of success in the Salk polio vaccine trials where the physical act of randomization led to credible causal inferences; or (b) to the unification given by the notion of maximum likelihood estimation and likelihood ratio tests

both for our general statistical modeling as well as for more directed formal modeling in a behavioral science subdomain, such as image processing or cognitive neuroscience. These frameworks can also be based on more quantitatively formal structures: for example, (a) the general linear model and its special cases of analysis of variance (ANOVA), analysis of covariance, and so on, along with model comparisons through full and reduced models; (b) the general principles behind prediction/selection/correlation in simple two-variable systems, with extensions to multiple-variable contexts; and (c) the various dimensionality reduction techniques of principal component/factor analysis, multidimensional scaling, cluster analysis, and discriminant analysis.

The remainder of the sections in this chapter will attempt to sketch some basic structures typically introduced in the graduate statistics sequence in the behavioral sciences, along with some necessary cautionary comments on usage and interpretation. The purpose is to provide a small part of the formal scaffolding needed in reasoning ethically about what we see in the course of our careers, both in our own work and that of others, or what might be expected of a statistically literate populace generally. Armed with this deeper understanding, graduates can be expected to deal more effectively with whatever ethically charged situations they might face.

## Probability Theory

The formalism of thought offered by *probability theory* is one of the more useful portions of any beginning course in statistics in helping to promote ethical reasoning. As typically presented, we speak of an event represented by a capital letter, say $A$, and the probability of the event as some number in the range from 0 to 1, written as $P(A)$. The value of 0 is assigned to the "impossible" event that can never occur; 1 is assigned to the "sure" event that will always occur. The driving condition for the complete edifice of all probability theory is one postulate: for two mutually exclusive events, $A$ and $B$ (where mutually exclusivity implies that both events cannot occur at the same time), $P(A \text{ or } B) = P(A) + P(B)$. As one final beginning definition, we say that two events are independent whenever the probability of the joint event, $P(A \text{ and } B)$, factors as the product of the individual probabilities, $P(A)P(B)$.

The idea of *statistical independence* and the factoring of the joint event probability immediately provide a formal tool for understanding several historical miscarriages of justice. In particular, if two events are not independent, then the joint probability cannot be generated by a simple product

of the individual probabilities. A recent example is the case of Sally Clark; she was convicted in England of killing her two children, partially on the basis of an inappropriate assumption of statistical independence. The purveyor of statistical misinformation in this case was Sir Roy Meadow, famous for Meadow's Law: "one sudden infant death is a tragedy, two is suspicious, and three is murder." We quote part of a news release from the Royal Statistical Society (2001):

> The Royal Statistical Society today issued a statement, prompted by issues raised by the Sally Clark case, expressing its concern at the misuse of statistics in the courts.
>
> In the recent highly-publicised case of R v. Sally Clark, a medical expert witness drew on published studies to obtain a figure for the frequency of sudden infant death syndrome (SIDS, or 'cot death') in families having some of the characteristics of the defendant's family. He went on to square this figure to obtain a value of 1 in 73 million for the frequency of two cases of SIDS in such a family.
>
> This approach is, in general, statistically invalid. It would only be valid if SIDS cases arose independently within families, an assumption that would need to be justified empirically. Not only was no such empirical justification provided in the case, but there are very strong a priori reasons for supposing that the assumption will be false. There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely.
>
> The well-publicised figure of 1 in 73 million thus has no statistical basis. Its use cannot reasonably be justified as a 'ballpark' figure because the error involved is likely to be very large, and in one particular direction. The true frequency of families with two cases of SIDS may be very much less incriminating than the figure presented to the jury at trial.

Numerous other examples for a misuse of the idea of statistical independence exist in the legal literature, such as the notorious 1968 jury trial in California, People v. Collins. Here, the prosecutor suggested that the jury merely multiply several probabilities together, which he conveniently provided, to ascertain the guilt of the defendant. In overturning the conviction, the Supreme Court of California criticized both the statistical reasoning and the framing of the decision for the jury:

> We deal here with the novel question whether evidence of mathematical probability has been properly introduced and used by the prosecution in a criminal case. … Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search of truth, must not cast a spell over him. We conclude that on the record before us, defendant should not have had his guilt

> determined by the odds and that he is entitled to a new trial. We reverse the judgement.

We will return to both the Clark and Collins cases later when Bayes' rule is discussed in the context of *conditional probability* confusions and what is called the "Prosecutor's Fallacy."

Besides the concept of independence, the definition of conditional probability plays a central role in all our uses of probability theory; in fact, most misapplications of statistical/probabilistic reasoning involve confusions of some sort regarding conditional probabilities. Formally, the conditional probability of some event $A$ given that $B$ has already occurred, denoted $P(A|B)$, is defined generally as $P(A \text{ and } B)/P(B)$; when $A$ and $B$ are independent, $P(A|B) = P(A)P(B)/P(B) = P(A)$; or in words, knowing that $B$ has occurred does not alter the probability of $A$ occurring. If $P(A|B) > P(A)$, we will say that $B$ is "facilitative" of $A$; when $P(A|B) < P(A)$, $B$ is said to be "inhibitive" of $A$. As a small example, suppose $A$ is the event of receiving a basketball scholarship; $B$, the event of being 7 feet tall; and $C$, the event of being 5 feet tall. One obviously expects $B$ to be facilitative of $A$ [i.e., $P(A|B) > P(A)$] and of $C$ to be inhibitive of $A$ [i.e., $P(A|C) < P(A)$]. In any case, the size and sign of the difference between $P(A|B)$ and $P(A)$ is an obvious raw descriptive measure of how much the occurrence of $B$ is associated with an increased or decreased probability of $A$, with a value of zero corresponding to statistical independence.

One convenient device for interpreting probabilities and understanding how events can be "facilitative" or "inhibitive" is through the use of a simple $2 \times 2$ table that cross-classifies a set of objects according to the events $A$ and $\overline{A}$ and $B$ and $\overline{B}$. For example, suppose we have a collection of $N$ balls placed in a container; each ball is labeled with $A$ or $\overline{A}$, and also with $B$ or $\overline{B}$, according to the notationally self-evident table of frequencies shown in Table 4.1.

The process we consider is one of picking a ball blindly from the container (where the balls are assumed to be mixed thoroughly) and noting the occurrence of the events $A$ or $\overline{A}$ and $B$ or $\overline{B}$. Based on this physical idealization of such a selection process, it is intuitively reasonable to

**TABLE 4.1**

A Generic $2 \times 2$ Contingency Table

| | $A$ | $\overline{A}$ | Row Sums |
|---|---|---|---|
| $B$ | $N_{AB}$ | $N_{\overline{A}B}$ | $N_B$ |
| $\overline{B}$ | $N_{A\overline{B}}$ | $N_{\overline{A}\overline{B}}$ | $N_{\overline{B}}$ |
| Column Sums | $N_A$ | $N_{\overline{A}}$ | $N$ |

assign probabilities according to the proportion of balls in the container satisfying the attendant conditions:

$$P(A) = N_A/N;\ P(\bar{A}) = N_{\bar{A}}/N;\ P(B) = N_B/N;\ P(\bar{B}) = N_{\bar{B}}/N;$$

$$P(A \mid B) = N_{AB}/N_B;\ P(B \mid A) = N_{AB}/N_A;$$

$$P(\bar{A} \mid B) = N_{\bar{A}B}/N_B;\ P(B \mid \bar{A}) = N_{\bar{A}B}/N_{\bar{A}};$$

$$P(\bar{B} \mid A) = N_{A\bar{B}}/N_A;\ P(A \mid \bar{B}) = N_{A\bar{B}}/N_{\bar{B}};$$

$$P(\bar{A} \mid \bar{B}) = N_{\bar{A}\bar{B}}/N_{\bar{B}};\ P(\bar{B} \mid \bar{A}) = N_{\bar{A}\bar{B}}/N_{\bar{A}}.$$

By noting the relationships: $N_B = N_{AB} + N_{\bar{A}B}$; $N_{\bar{B}} = N_{A\bar{B}} + N_{\bar{A}\bar{B}}$; $N_A = N_{AB} + N_{A\bar{B}}$; $N_{\bar{A}} = N_{\bar{A}B} + N_{\bar{A}\bar{B}}$; $N_B + N_{\bar{B}} = N_A + N_{\bar{A}} = N$, a variety of interesting connections can be derived and understood that can assist immensely in our probabilistic reasoning. We present a short numerical example below on how these ideas might be used in a realistic context; several such uses are then expanded on in the subsections to follow.

As a numerical example of using a $2 \times 2$ contingency table to help explicate probabilistic reasoning, suppose we have an assumed population of 10,000, cross-classified according to the presence or absence of colorectal cancer (CC) [$A$: +CC; $\bar{A}$: –CC], and the status of a fecal occult blood test (FOBT) [$B$: +FOBT; $\bar{B}$: –FOBT]. Using the data from Gerd Gigerenzer, *Calculated Risks* (2002), we have the $2 \times 2$ Table 4.2.

The probability, $P(+CC|+FOBT)$, is simply 15/314 = .048, using the frequency value of 15 for the cell (+FOBT, +CC) and the +FOBT row sum of 314. The marginal probability, P(+CC), is 30/10,000 = .003, and thus, a positive FOBT is "facilitative" of a positive CC because .048 is greater than .003. The size of the difference, $P(+CC|+FBOT) - P(+CC) = +.045$, may not be large in any absolute sense, but the change does represent a 15-fold increase over the marginal probability of .003. (But note that if you have a positive FOBT, more than 95% of the time you do not have cancer, i.e., there are 95% $\left(\frac{299}{314}\right)$ *false positive*s.)

There are many day-to-day contexts faced where our decisions might best be made from conditional probabilities (if we knew them) instead

**TABLE 4.2**

A $2 \times 2$ Contingency Table Between Colorectal Cancer and the Fecal Occult Blood Test

| | +CC | –CC | Row Sums |
|---|---|---|---|
| +FOBT | 15 | 299 | 314 |
| –FOBT | 15 | 9,671 | 9,686 |
| Column Sums | 30 | 9,970 | 10,000 |

of from marginal information. When deciding on a particular medical course of action, for example, it is important to condition on our own circumstances of age, risk factors, family medical history, our own psychological needs and makeup, and so on. A recent and controversial instance of this, where the conditioning information is "age," is reported in *The New York Times* article by Gina Kolata, *In Reversal, Panel Urges Mammograms at 50, Not 40* (November 17, 2009).

There are a variety of probability results that prove useful throughout our attempt to reason probabilistically and follow the field of statistical inference. We list some of these below, with uses given throughout this chapter.

1. For the complementary event, $\bar{A}$, which occurs when $A$ does not,

$$P(\bar{A}) = 1 - P(A).$$

2. For events $A$ and $B$ that are not necessarily mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

3. The *rule of total probability*: given a collection of mutually exclusive and exhaustive events, $B_1, \ldots, B_K$ (i.e., all are pairwise mutually exclusive and their union gives the sure event),

$$P(A) = \sum_{k=1}^{K} P(A \mid B_k) P(B_k).$$

4. Bayes' theorem (or rule) for two events, $A$ and $B$:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})}.$$

5. Bonferroni inequality: for a collection of events, $A_1, \ldots, A_K$,

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k).$$

6. $P(A \text{ and } B) \leq P(A \text{ or } B) \leq P(A) + P(B)$. In words, the first inequality results from the event "$A$ and $B$" being wholly contained within the event "$A$ or $B$"; the second obtains from the Bonferroni inequality restricted to two events.
7. $P(A \text{ and } B) \leq \text{minimum } (P(A), P(B)) \leq P(A) \text{ or } \leq P(B)$. In words, the first inequality results from the event "$A$ and $B$" being wholly contained both within $A$ and within $B$; the second inequalities are more generally appropriate—the minimum of any two numbers is always less than or equal to either of the two numbers.

### The (Mis-)Assignment of Probabilities

Although the assignment of probabilities to events consistent with the disjoint rule may lead to an internally valid system mathematically, there is still no assurance that this assignment is "meaningful" or bears any empirical validity for observable long-run expected frequencies. There seems to be a never-ending string of misunderstandings in the way probabilities can be generated that are either blatantly wrong or more subtly incorrect, irrespective of the internally consistent system they might lead to. Some of these problems are briefly sketched below, but we can only hope to be representative of a few possibilities, not exhaustive.

One inappropriate way of generating probabilities is to compute the likelihood of some joint occurrence after some of the outcomes are already known. There is the story about the statistician who takes a bomb aboard a plane, reasoning that if the probability of one bomb on board is small, the probability of two is infinitesimal. Or, during World War I, soldiers were actively encouraged to use fresh shell holes as shelter because it was very unlikely for two shells to hit the same spot during the same day. And the (Minnesota Twins) baseball manager who bats for a player, who earlier in the game hit a home run, because it would be very unlikely for him to hit two home runs in the same game. Although these (slightly) amusing stories may provide obvious misassignments of probabilities, other related situations are more subtle. For example, whenever coincidences are culled or "hot spots" identified from some search of available information, the probabilities that are then regenerated for these situations may not be valid. There are several ways of saying this: When some set of observations is the source of an initial suspicion, those same observations should not be used in a calculation that then tests the validity of the suspicion. In Bayesian terms, you do not get the Posterior from the same information that gave you the Prior.

Alternatively said, it makes no sense to do formal hypothesis assessment (by finding estimated probabilities) when the data themselves have suggested the hypothesis in the first place. Some cross-validation strategy is necessary; for example, collecting independent data. Generally, when some process of search or optimization has been used to identify an unusual situation (e.g., when a "good" regression equation is found through a step-wise procedure [see Freedman, 1983, for a devastating critique]; when data are "mined" and unusual patterns identified; when DNA databases are searched for "cold-hits" against evidence left at a crime scene; when geographic "hot spots" are identified for, say, some particularly unusual cancer, and so on), the same methods for assigning probabilities before the particular situation was identified are generally no longer appropriate post hoc.

A second general area of inappropriate probability assessment concerns the model postulated to aggregate probabilities over several events.

Campbell (1974) cites an article in the *New York Herald Tribune* (May, 1954) stating that if the probability of knocking down an attacking airplane was .15 at each of five defense positions before reaching the target, then the probability of knocking down the plane before it passed all five barriers would be .75 ($5 \times .15$), this last value being the simple sum of the probabilities, and an inappropriate model. If we could correctly assume independence between the Bernoulli trials at each of the five positions, a more justifiable value would be one minus the probability of passing all barriers successfully: $1.0 - (.85)^5 \approx .56$. The use of similar binomial modeling possibilities, however, may be specious—for example, when dichotomous events occur simultaneously in groups (e.g., the World Trade Center disaster on September 11, 2001); when the success proportions are not valid; when the success proportions change in value over the course of the trials; when time dependencies are present in the trials (e.g., tracking observations above and below a median over time); and so on. In general, when wrong models are used to generate probabilities, the resulting values may have little to do with empirical reality. For example, in throwing dice and counting the sum of spots that result, it is not true that each of the integers from 2 through 12 is equally likely. The model of what is equally likely may be reasonable at a different level (e.g., pairs of integers appearing on the two dice) but not at all aggregated levels. There are some stories, probably apocryphal, of methodologists meeting their demise by making these mistakes for their gambling patrons.

Flawed calculations of probability can have dire consequences within our legal systems, as the case of Sally Clark and related others make clear. One broad and current area of possible misunderstanding of probabilities is in the context of DNA evidence (which is exacerbated in the older and much more fallible system of identification through fingerprints). In the use of DNA evidence (and with fingerprints), one must be concerned with the random match probability (RMP): the likelihood that a randomly selected unrelated person from the population would match a given DNA profile. Again, the use of independence in RMP estimation is questionable; also, how does the RMP relate to, and is it relevant for, "cold-hit" searches in DNA databases? In a confirmatory identification case, a suspect is first identified by non-DNA evidence; DNA evidence is then used to corroborate traditional police investigation. In a "cold-hit" framework, the suspect is first identified by a search of DNA databases; the DNA evidence is thus used to identify the suspect as perpetrator, to the exclusion of others, directly from the outset (this is somewhat akin to shooting an arrow into a tree and then drawing a target around it). Here, traditional police work is no longer the focus. For a thorough discussion of the probabilistic context surrounding DNA evidence (which extends with even greater force to fingerprints), the article by Jonathan Koehler (1993) is recommended.

In 1989, and based on urging from the FBI, the National Research Council (NRC) formed the Committee on DNA Technology in Forensic Science, which issued its report in 1992 (*DNA Technology in Forensic Science*; or more briefly, NRC I). The NRC I recommendation about the cold-hit process was as follows:

> The distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important. The chance of finding a match in the second case is considerably higher. … The initial match should be used as probable cause to obtain a blood sample from the suspect, but only the statistical frequency associated with the additional loci should be presented at trial (to prevent the selection bias that is inherent in searching a databank).

A follow-up report by a second NRC panel was published in 1996 (*The Evaluation of Forensic DNA Evidence*; or more briefly, NRC II), having the following main recommendation about cold-hit probabilities and using what has been called the *database match probability* (DMP):

> When the suspect is found by a search of DNA databases, the random-match probability should be multiplied by *N*, the number of persons in the database.

The term *database match probability* is somewhat unfortunate; this is not a real probability but more of an expected number of matches given the RMP. A more legitimate value for the probability that another person matches the defendant's DNA profile would be $1 - \left(1 - \frac{1}{RMP}\right)^N$, for a database of size *N*, that is, one minus the probability of no matches over *N* trials. For example, for an RMP of 1/1,000,000 and an *N* of 1,000,000, the above probability of another match is .632; the DMP (not a probability) number is 1.00, being the product of *N* and RMP. In any case, NRC II made the recommendation of using the DMP to give a measure of the accuracy of a cold-hit match (and did not support the more legitimate "probability of another match" using the formula given above [possibly because it was considered too difficult?]):

> A special circumstance arises when the suspect is identified not by an eyewitness or by circumstantial evidence but rather by a search through a large DNA database. If the only reason that the person becomes a suspect is that his DNA profile turned up in a database, the calculations must be modified. There are several approaches, of

> which we discuss two. The first, advocated by the 1992 NRC report, is to base probability calculations solely on loci not used in the search. That is a sound procedure, but it wastes information, and if too many loci are used for identification of the suspect, not enough might be left for an adequate subsequent analysis. … A second procedure is to apply a simple correction: Multiply the match probability by the size of the database searched. This is the procedure we recommend.

## The Probabilistic Generalizations of Logical Fallacies Are No Longer Fallacies

In our roles as instructors of beginning statistics, we commonly introduce some simple logical considerations early on that revolve around the usual "if *p*, then *q*" statements, where *p* and *q* are two propositions. As an example, we might let *p* be "the animal is a yellow Labrador retriever," and *q*, "the animal is in the order *Carnivora*." Continuing, we note that if the statement "if *p*, then *q*" is true (which it is), then logically so must be the contrapositive of "if not *q*, then not *p*," that is, if "the animal is not in the order *Carnivora*," then "the animal is not a yellow Labrador retriever." However, two fallacies await the unsuspecting:

Denying the antecedent: if not *p*, then not *q* (if "the animal is not a yellow Labrador retriever," then "the animal is not in the order *Carnivora*").

Affirming the consequent: if *q*, then *p* (if "the animal is in the order *Carnivora*," then "the animal is a yellow Labrador retriever").

Also, when we consider definitions given in the form of "*p* if and only if *q*," (e.g., "the animal is a domesticated dog" if and only if "the animal is a member of the subspecies *Canis lupus familiaris*"), or equivalently, "*p* is necessary and sufficient for *q*," these separate into two parts:

"If *p*, then *q*" (i.e., *p* is a sufficient condition for *q*).

"If *q*, then *p*" (i.e., *p* is a necessary condition for *q*).

So, for definitions, the two fallacies are not present.

In a probabilistic context, we reinterpret the phrase "if *p*, then *q*" as *B* being facilitative of *A*; that is, $P(A|B) > P(A)$, where *p* is identified with *B* and *q* with *A*. With such a probabilistic reinterpretation, we no longer have the fallacies of denying the antecedent [i.e., $P(\bar{A} \mid \bar{B}) > P(\bar{A})$], or of affirming the consequent [i.e., $P(B|A) > P(B)$]. Both of the latter two probability statements can be algebraically shown true using the simple

$2 \times 2$ cross-classification frequency table and the equivalences among frequency sums given earlier:

$$\text{(original statement) } P(A|B) > P(A) \Leftrightarrow N_{AB}/N_B > N_A/N \Leftrightarrow$$

$$\text{(denying the antecedent) } P(\bar{A}|\bar{B}) > P(\bar{A}) \Leftrightarrow N_{\bar{A}\bar{B}}/N_{\bar{B}} > N_{\bar{A}}/N \Leftrightarrow$$

$$\text{(affirming the consequent) } P(B|A) > P(B) \Leftrightarrow N_{AB}/N_A > N_B/N \Leftrightarrow$$

$$\text{(contrapositive) } P(\bar{B}|\bar{A}) > P(\bar{B}) \Leftrightarrow N_{\bar{A}\bar{B}}/N_{\bar{A}} > N_{\bar{B}}/N$$

Another way of understanding these results is to note that the original statement of $P(A|B) > P(A)$ is equivalent to $N_{AB} > N_A N_B / N$, or in the usual terminology of a $2 \times 2$ contingency table, the frequency in the cell labeled $(A,B)$ is greater than the typical expected value constructed under independence of the attributes based on the row total, $N_B$, times the column total, $N_A$, divided by the grand total, $N$. The other probability results follow from the observation that with fixed marginal frequencies, a $2 \times 2$ contingency table has only one degree of freedom. These results derived from the original of $B$ being facilitative for $A$, $P(A|B) > P(A)$ could have been restated as $\bar{B}$ being inhibitive of $A$, or as $\bar{A}$ being inhibitive of $B$.

In reasoning logically about some situation, it would be rare to have a context that would be so cut-and-dried as to lend itself to the simple logic of "if $p$, then $q$," and where we could look for the attendant fallacies to refute some causal claim. More likely, we are given problems characterized by fallible data and subject to other types of probabilistic processes. For example, even though someone may have some genetic marker that has a greater presence in individuals who have developed some disease (e.g., breast cancer and the *BRAC1* gene), it is not typically an unadulterated causal necessity; in other words, it is not true that "if you have the marker, then you must get the disease." In fact, many of these situations might be best reasoned through using our simple $2 \times 2$ tables—$A$ and $\bar{A}$ denote the presence/absence of the marker; $B$ and $\bar{B}$ denote the presence/absence of the disease. Assuming $A$ is facilitative of $B$, we could go on to ask about the strength of the facilitation by looking at, say, the difference, $P(B|A) - P(B)$.

The idea of arguing probabilistic causation is, in effect, our notion of one event being facilitative or inhibitive of another. If we observe a collection of "$q$" conditions that would be the consequence of a single "$p$," we may be more prone to conjecture the presence of "$p$." Although this process may seem like merely affirming the consequent, in a probabilistic context this could be referred to as "inference to the best explanation," or as a variant of the Charles Pierce notion of abductive reasoning. In any case, with a probabilistic reinterpretation, the assumed fallacies of logic may not be such; moreover, most uses of information in contexts that are legal (forensic) or

medical (through screening), or that might, for example, involve academic or workplace selection, need to be assessed probabilistically.

### Using Bayes' Rule to Assess the Consequences of Screening for Rare Events

*Bayes' theorem,* or rule, was given in a form appropriate for two events, $A$ and $B$; it allows the computation of one conditional probability, $P(A|B)$, from two other conditional probabilities, $P(B|A)$ and $P(B|\bar{A})$, and the prior probability for the event $A$, $P(A)$. A general example might help show the importance of Bayes' rule in assessing the value of screening for the occurrence of rare events.

Suppose we have a test that assesses some relatively rare quantity (e.g., disease, ability, talent, terrorism propensity, drug/steroid usage, antibody presence, being a liar [where the test is a polygraph], and so forth). Let $B$ be the event that the test says the person has "it," whatever that may be; $A$ is the event that the person really does have "it." Two "reliabilities" are needed:

1. The probability, $P(B|A)$, that the test is positive if the person has "it"; this is called the *sensitivity* of the test.
2. The probability, $P(\bar{B}|\bar{A})$, that the test is negative if the person does not have "it"; this is the *specificity* of the test. The conditional probability used in the denominator of Bayes' rule, $P(B|\bar{A})$, is merely $1 - P(\bar{B}|\bar{A})$, and is the probability of a "false positive."

The quantity of prime interest, called the *positive predictive value* (PPV), is the probability that a person has "it" given that the test says so, $P(A|B)$, and is obtainable from Bayes' rule using the specificity, sensitivity, and prior probability, $P(A)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(\bar{B}|\bar{A}))(1 - P(A))}.$$

To understand how well the test does, the facilitative effect of $B$ on $A$ needs interpretation, that is, a comparison of $P(A|B)$ with $P(A)$, plus an absolute assessment of the size of $P(A|B)$ by itself. Here, the situation is usually dismal whenever $P(A)$ is small (i.e., screening for a relatively rare quantity) and the sensitivity and specificity are not perfect. Although $P(A|B)$ will generally be greater than $P(A)$, and thus, $B$ facilitative of $A$, the absolute size of $P(A|B)$ is typically so small that the value of the screening may be questionable.

As an example, consider the efficacy of mammograms in detecting breast cancer. In the United States, about 180,000 women are found to have breast

cancer each year from among the 33.5 million women who annually have a mammogram. Thus, the probability of a tumor is 180,000/33,500,000 = .0054. Mammograms are no more than 90% accurate, implying that $P(\textit{positive mammogram} \mid \textit{tumor}) = .90$; $P(\textit{negative mammogram} \mid \textit{no tumor}) = .90$. Because we do not know whether a tumor is present—all we know is whether the test is positive—Bayes' theorem must be used to calculate the probability we really care about, the PPV: $P(\textit{tumor} \mid \textit{positive mammogram})$. All the pieces are available to use Bayes' theorem to calculate this probability, and we will do so below. But first, as an exercise for the reader, try to estimate the order of magnitude of that probability, keeping in mind that cancer is rare and the test for it is 90% accurate. Do you guess that if you test positive, you have a 90% chance of cancer? Or perhaps 50%, or 30%? How low must this probability drop before we believe that mammograms may be an unjustifiable drain on resources? Using Bayes' rule, the PPV of the test is .047:

$$P(\textit{tumor} \mid \textit{positive mammogram}) = \frac{.90(.0054)}{.09(.0054) + .10(.9946)} = .047,$$

which is obviously greater than the prior probability of .0054 but still very small in magnitude, that is, more than 95% of the positive tests that arise turn out to be incorrect.

Whether using a test that is wrong 95% of the time is worth doing is, at least partially, an ethical question, for if we decide that it is not worth doing, what is the fate of the 5% or so of women who are correctly diagnosed? We will not attempt a full analysis, but some factors considered might be economic, for 33.5 million mammograms cost about $3.5 billion, and the 3.5 million women incorrectly diagnosed can be, first, dysfunctionally frightened, and second, they must use up another day for a biopsy, in turn costing at least $1,000 and adding another $3.5 billion to the overall diagnostic bill. Is it worth spending $7 billion to detect 180,000 tumors? That is about $39,000/tumor detected. And, not to put too fine a point on it, biopsies have their own risks: 1% yield staphylococcal infections, and they too have false-positive results, implying that some women end up being treated for nonexistent cancers. Also, the majority of the cancers detected in the 5% alluded to above are generally not life-threatening and just lead to the ills caused by overdiagnosis and invasive overtreatment. The statistics calculated do not make the decision about whether it is ethical to do mammograms, but such a decision to be ethical should be based on accurate information. Two recent articles discuss how the American Cancer Society may itself be shifting its stance on screening; the "page one, above the fold" pieces are by Gina Kolata *(In Shift, Cancer Society Has Concerns on Screenings, The New York Times,* October, 21 2009; *In Reversal, Panel Urges Mammograms at 50, Not 40, The New York Times,*

November 17, 2009). A third recent article discusses the odds and economics of screening (with calculations similar to those given here): *Gauging the Odds (and the Costs) in Health Screening* (Richard H. Thaler, *The New York Times,* December 20, 2009).

As we have seen in subsequent reactions to these "new" recommendations regarding screening for breast cancer, it is doubtful whether individual women will comply, or even that their doctors will advise them to. Health recommendations, such as these, pertain to an aggregate populace, possibly subdivided according to various demographic categories. But an individual who seeks some kind of control over (breast) cancer is not going to give up the only means she has to do so; all women know (at least indirectly) various individuals for whom breast cancer was detected early (and "cured," even though the given cancer may not have been harmful); similarly, all women know about individuals who died after a cancer had metastasized before screening located it. What might be justifiable public health policy in the aggregate may not be so when applied at the level of individuals; also, the issue that trumps all in the mammogram discussion is what women want (or think they want, which amounts to the same thing). It is doubtful whether a reasoned argument for diminished screening could ever be made politically palatable. To many, a statistical argument for a decrease of screening practice would merely be another mechanism by which insurance companies can deny coverage and make yet more money. To paraphrase a quote about General Motors, it is not true that "what is good for the Insurance Industry is good for the country" (or for that matter, for any single individual living in it). Two very cogent articles on these issues of screening both for individuals and the aggregate appeared on the same day (November 20, 2009) in *The New York Times: A Medical Culture Clash* by Kevin Sack and *Addicted to Mammograms* by Robert Aronowitz.

It might be an obvious statement to make, but in our individual dealings with doctors and the medical establishment generally, it is important for all to understand the PPVs for whatever screening tests we now seem to be constantly subjected to, and thus, the number, (1 – PPV), referring to the false positives, that is, if a patient tests positive, what is the probability that "it" is not actually present. It is a simple task to plot PPV against $P(A)$ from 0 to 1 for any given pair of sensitivity and specificity values. Such a plot can show dramatically the need for highly reliable tests in the presence of low values of $P(A)$ to attain even mediocre PPV values.

Besides a better understanding of how PPVs are determined, there is a need to recognize that even when a true positive exists, not every disease needs to be treated. In the case of another personal favorite of ours, prostate cancer screening (in that its low accuracy makes mammograms look good), where the worst danger is one of overdiagnosis and overtreatment, leading to more harm than good (see, e.g., Gina Kolata, *Studies Show*

*Prostate Test Save Few Lives, The New York Times,* March 19, 2009). Armed with this information, we no longer have to hear the snap of a latex glove behind our backs at our yearly physical, nor do we give blood for a prostate-specific antigen screening test. When we so informed our doctors as to our wishes, they agreed completely; the only reason such tests were done routinely was to practice "defensive medicine" on behalf of their clinics, and to prevent possible lawsuits arising from such screening tests not being administered routinely. In other words, clinics get sued for underdiagnosis but not for overdiagnosis and overtreatment.

## Bayes' Rule and the Confusion of Conditional Probabilities

One way of rewriting Bayes' rule is to use a ratio of probabilities, $P(A)/P(B)$, to relate the two conditional probabilities of interest, $P(B|A)$ (test sensitivity) and $P(A|B)$ (PPV):

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}.$$

With this rewriting, it is obvious that $P(A|B)$ and $P(B|A)$ will be equal only when the prior probabilities, $P(A)$ and $P(B)$, are the same. Yet, this confusion error is so common in the forensic literature that it is given the special name of the *Prosecutor's Fallacy*. In the behavioral sciences research literature, this Prosecutor's Fallacy is sometimes called the "Fallacy of the Transposed Conditional" or the "Inversion Fallacy." In the context of statistical inference, it appears when the probability of seeing a particular data result conditional on the null hypothesis being true, $P(\text{data}|H_o)$, is confused with $P(H_o|\text{data})$; that is, the probability that the null hypothesis is true given that a particular data result has occurred.

As a case in point, we return to the Sally Clark conviction where the invalidly constructed probability of 1 in 73 million was used to successfully argue for Sally Clark's guilt. Let $A$ be the event of innocence and $B$ the event of two "cot deaths" within the same family. The invalid probability of 1 in 73 million was considered to be for $P(B|A)$; a simple equating with $P(A|B)$, the probability of innocence given the two cot deaths, led directly to Sally Clark's conviction. We continue with the Royal Statistical Society Press Release:

> Aside from its invalidity, figures such as the 1 in 73 million are very easily misinterpreted. Some press reports at the time stated that this was the chance that the deaths of Sally Clark's two children were accidental. This (mis-)interpretation is a serious error of logic known as the Prosecutor's Fallacy.
>
> The Court of Appeal has recognised these dangers (R v. Deen 1993, R v. Doheny/Adams 1996) in connection with probabilities used for

> DNA profile evidence, and has put in place clear guidelines for the presentation of such evidence. The dangers extend more widely, and there is a real possibility that without proper guidance, and well-informed presentation, frequency estimates presented in court could be misinterpreted by the jury in ways that are very prejudicial to defendants.
>
> Society does not tolerate doctors making serious clinical errors because it is widely understood that such errors could mean the difference between life and death. The case of R v. Sally Clark is one example of a medical expert witness making a serious statistical error, one which may have had a profound effect on the outcome of the case.
>
> Although many scientists have some familiarity with statistical methods, statistics remains a specialised area. The Society urges the Courts to ensure that statistical evidence is presented only by appropriately qualified statistical experts, as would be the case for any other form of expert evidence.

The situation with Sally Clark and the Collins case in California (which both involved the Prosecutor's Fallacy) is not isolated. There was the recent miscarriage of justice in The Netherlands involving a nurse, Lucia de Berk, accused of multiple deaths at the hospitals where she worked. This case aroused the international community of statisticians to redress the apparent ills visited on Lucia de Berk. One source for background (although now somewhat dated) is Mark Buchanan at *The New York Times* Blogs (*The Prosecutor's Fallacy,* May 16, 2007). The Wikipedia article on "Lucia de Berk" provides the details of the case and the attendant probabilistic arguments, up to her complete exoneration in April of 2010.

A much earlier and historically important fin-de-siècle case is that of Alfred Dreyfus, the much maligned French Jew and captain in the military, who was falsely imprisoned for espionage. In this example, the nefarious statistician was the rabid anti-Semite Alphonse Bertillon, who through a convoluted argument, reported a very small probability that Dreyfus was "innocent"; this meretricious probability had no justifiable mathematical basis and was generated from culling coincidences involving a document, the handwritten *bordereau* (without signature) announcing the transmission of French military information. Dreyfus was accused and convicted of penning this document and passing it to the (German) enemy. The Prosecutor's Fallacy was more or less invoked to ensure a conviction based on the fallacious small probability given by Bertillon. In addition to Emile Zola's famous 1898 article, *J'Accuse*, in the newspaper *L'Aurore* on January 13, 1898, it is interesting to note that well-known turn-of-the-century statisticians and probabilists from the French Academy of Sciences (among them Henri Poincairé) demolished Bertillon's probabilistic arguments

and insisted that any use of such evidence needs to proceed in a fully Bayesian manner, much like our present understanding of evidence in current forensic science and the proper place of probabilistic argumentation. A detailed presentation of all the probabilistic and statistical issues and misuses present in the Dreyfus case is given by Champod, Taroni, and Margot (1999). (Also, see the comprehensive text by Aitken and Taroni, 2004, *Statistics and the Evaluation of Evidence for Forensic Scientists*.)

We observe the same general pattern in all of the miscarriages of justice involving the Prosecutor's Fallacy. There is some very small reported probability of "innocence," typically obtained incorrectly either by culling, misapplying the notion of statistical independence, or using an inappropriate statistical model. Such a probability is calculated by a supposed expert with some credibility in court: a community college mathematics instructor for Collins, Roy Meadow for Clark, Henk Elffers for de Berk, and Alphonse Bertillon for Dreyfus. The Prosecutor's Fallacy then takes place, leading to a conviction for the crime. Various outrages ensue from the statistically literate community, with the eventual emergence of some "statistical good guys" hoping to redress the wrongs done: an unnamed court-appointed statistician for the California Supreme Court for Collins, Richard Gill for de Berk, Henri Poincairé (among others) for Dreyfus, and the Royal Statistical Society for Clark. After long periods, convictions are eventually overturned, typically after extensive prison sentences have already been served. We can only hope to avoid similar miscarriages of justice in cases yet to come by recognizing the tell-tale pattern of occurrence for the Prosecutor's Fallacy.

There seem to be any number of conditional probability confusions that can arise in important contexts (and possibly when least expected). A famous instance of this is in the O.J. Simpson case, where one conditional probability, say, $P(A|B)$, was confused with another, $P(A|B \text{ and } D)$. We quote the clear explanation of this obfuscation by Krämer and Gigerenzer (2005):

> Here is a more recent example from the U.S., where likewise $P(A|B)$ is confused with $P(A|B \text{ and } D)$. This time the confusion is spread by Alan Dershowitz, a renowned Harvard Law professor who advised the O.J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflected a motive to kill, advancing the premise that "a slap is a prelude to homicide." Dershowitz, however, called this argument "a show of weakness" and said: "We knew that we could prove, if we had to, that an infinitesimal percentage—certainly fewer than 1 of 2,500—of men who slap or beat their domestic partners go on to murder them." Thus, he argued that the probability of the event $K$ that a husband killed his wife if he battered her was small, $P(K|\text{battered}) = 1/2{,}500$. The relevant probability, however, is not this one, as Dershowitz would have us believe. Instead, the relevant probability is that of a man murdering his partner given

> that he battered her and that she was murdered, $P(K|\text{battered and murdered})$. This probability is about 8/9. It must of course not be confused with the probability that O.J. Simpson is guilty; a jury must take into account much more evidence than battering. But it shows that battering is a fairly good predictor of guilt for murder, contrary to Dershowitz's assertions. (p. 228)

## The Basic Sampling Model and Related Issues

From *The New York Times* article by David Stout (April 3, 2009), *Obama's Census Choice Unsettles Republicans:*

> Robert M. Groves, a former census official and now a sociology professor at the University of Michigan, was nominated Thursday by President Obama to run the Census Bureau, a choice that instantly made Republicans nervous.
>
> Republicans expressed alarm because of one of Mr. Groves's specialties, statistical sampling—roughly speaking, the process of extrapolating from the numbers of people actually counted to arrive at estimates of those uncounted and, presumably, arriving at a realistic total.
>
> If minorities, immigrants, the poor and the homeless are those most likely to be missed in an actual head count, and if political stereotypes hold true, then statistical sampling would presumably benefit the Democrats.
>
> Republicans have generally argued that statistical sampling is not as reliable as its devotees insist. "Conducting the census is a vital constitutional obligation," Representative John A. Boehner of Ohio, the House minority leader, said Thursday. "It should be as solid, reliable and accurate as possible in every respect. That is why I am concerned about the White House decision to select Robert Groves as director of the Census Bureau."
>
> Mr. Boehner, recalling that controversy (from the early 1990s when Mr. Groves pushed for statistically adjusting the 1990 census to make up for an undercount), said Thursday that "we will have to watch closely to ensure the 2010 census is conducted without attempting similar statistical sleight of hand."

We begin by refreshing our memories about the distinctions between *population* and *sample*, *parameters* and *statistics*, and *population distributions* and *sampling distributions*. Someone who has successfully completed a year-long graduate sequence in statistics should know these distinctions very well. Here, only a simple univariate framework is considered explicitly, but obvious and straightforward generalizations exist for the multivariate context.

A *population* of interest is posited, operationalized by some random variable, say $X$. In this *Theory World* framework, $X$ is characterized by *parameters*, such as the expectation of $X$, $\mu = E(X)$, or its variance, $\sigma^2 = V(X)$. The random variable $X$ has a (*population*) *distribution*, which is often assumed normal. A *sample* is generated by taking observations on $X$, say, $X_1, \ldots, X_n$, considered independent and identically distributed as $X$, that is, they are exact copies of $X$. In this *Data World* context, statistics are functions of the sample, and therefore, characterize the sample: the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$; the sample variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$, with some possible variation in dividing by $n - 1$ to generate an unbiased estimator for $\sigma^2$. The statistics, $\hat{\mu}$ and $\hat{\sigma}^2$, are *point estimators* of $\mu$ and $\sigma^2$. They are random variables by themselves, so they have distributions called *sampling distributions*. The general problem of statistical inference is to ask what sample statistics, such as $\hat{\mu}$ and $\hat{\sigma}^2$, tell us about their population counterparts, $\mu$ and $\sigma^2$. In other words, can we obtain a measure of accuracy for estimation from the sampling distributions through, for example, *confidence intervals*?

Assuming that the population distribution is normally distributed, the sampling distribution of $\hat{\mu}$ is itself normal with expectation $\mu$ and variance $\sigma^2/n$. Based on this result, an approximate 95% confidence interval for the unknown parameter $\mu$ can be given by

$$\hat{\mu} \pm 2.0\frac{\hat{\sigma}}{\sqrt{n}}.$$

Note that it is the square root of the sample size that determines the length of the interval (and not the sample size per se). This is both good news and bad. Bad, because if you want to double precision, you need a fourfold increase in sample size; good, because sample size can be cut by four with only a halving of precision.

Even when the population distribution is not originally normally distributed, the central limit theorem (CLT) says that $\hat{\mu}$ is approximately normal in form and becomes exactly so as $n$ goes to infinity. Thus, the approximate confidence interval statement remains valid even when the underlying distribution is not normal; such a result underlies many claims of robustness; that is, when a procedure remains valid even if the assumption under which it was derived may not be true, as long as some particular condition is satisfied—here, that condition is for the sample size to be reasonably large. Although how large is big enough for a normal approximation to be adequate depends generally on the form of the underlying population distribution, a glance at a "$t$ table" will show that when the degrees-of-freedom are larger than 30, the values given are indistinguishable from that for the normal. Thus, we surmise that sample sizes above 30 should generally be large enough to invoke the benefits that the CLT provides.

Besides the robustness of the confidence interval calculations for $\mu$, the CLT also encompasses what is called the *law of large numbers* (LLN). As the sample size increases, the estimator, $\hat{\mu}$, gets closer and closer to $\mu$, and converges to $\mu$ at the limit of $n$ going to infinity. This is seen most directly in the sampling variance for $\hat{\mu}$, which gets smaller as the sample size gets larger.

The basic results obtainable from the CLT and LLN that averages are both less variable and more normal in distribution than individual observations, and that averages based on larger sample sizes will show less variability than those based on smaller sample sizes, have far- ranging and sometimes very subtle influences on our reasoning skills. For example, suppose we would like to study organizations, such as schools, health care units, or governmental agencies, and have some measure of performance on the individuals in the units and the average for each unit. To identify those units exhibiting best performance (or, in the current jargon, "best practice"), the top 10%, say, of units in terms of performance are identified; a determination is then made of what common factors might characterize these top-performing units. We are pleased when able to isolate one very salient feature—most units in this top tier are small; we proceed on this observation to advise in the breakup of larger units. Is such a policy really justified based on these data? Probably not, if one also observes that the bottom 10% are also small units. Given that smaller entities just tend to be inherently more variable than the larger would seem to vitiate a recommendation of breaking up the larger units for performance improvement. Evidence that the now defunct "small schools movement," funded heavily by the Gates Foundation, was a victim of the "square root of $n$ law" was presented by Wainer (2009, Chapter 1).

Another implication of the basic sampling model is that when the size of the population is effectively infinite, this does not affect the accuracy of our estimate, which is driven by sample size. Thus, if we want a more precise estimate, we need only draw a larger sample. For some reason, this confusion resurfaces and is reiterated every 10 years when the U.S. Census is planned, where the issues of complete enumeration, as demanded by the Constitution, and the problems of undercount are revisited. The beginning quotations from John Boehner in relation to the 2010 census are a good case in point. And the ethical implications of his statistical reasoning skills should be fairly clear.

An area of almost mythic proportions in which a misunderstanding, or at least, a misappreciation for randomness exists, is in sports. A reasonable model for sports performance is one of "observed performance" being the sum of "intrinsic ability" (or true performance) and "error," leading to natural variability in outcome either at the individual or the team level. Somehow it appears necessary for sports writers, announcers, and other pundits to continually give reasons for what is most likely just random variability. We hear of team "chemistry," good or bad, being present or

not; individuals having a "hot hand" (or a "cold hand," for that matter); someone needing to "pull out of a slump"; why there might be many more .400 hitters early in the season but not later; a player being "due" for a hit; free-throw failure because of "pressure"; and so on. Making decisions based on natural variation being somehow "predictive" or "descriptive" of the truth is not very smart, to say the least. But it is done all the time—sports managers are fired and CEOs replaced for what may be just the traces of natural variability.

In asking people to generate random sequences, they tend to underestimate the amount of variation present in such a stochastic process—not enough (longer) runs are present; there is a tendency to produce too many short alternations; and so on. In a similar way, we do not see the naturalness in what will be called in a later section, *regression toward the mean*—where extremes are followed by less extreme observations just because of fallibility in observed performance. And again, causes are sought. We hear about multiround golf tournaments where a good performance on the first day is followed by a less adequate score the second (probably the result of "pressure"); or a bad performance on the first day is followed by an improved performance the next (he or she must have been able to "play loose"). Or in baseball, at the start of a season, an underperforming Derek Jeter might be under "pressure" or too much "media scrutiny," or the difficulties of performing in a "New York Market." When an individual starts off well but then appears to fade, it must be people trying to stop him or her (i.e., "gunning" for someone). One should always remember that in estimating intrinsic ability, an individual is unlikely to be as good (or as bad) as the pace he or she is on. It is always a better bet to vote against someone eventually breaking some record, even when he or she is "on a pace" to so do early in the season. This may be one origin for the phrase "sucker bet"—a gambling wager where your expected return is significantly lower than your wager.

Another area where one expects to see a lot of anomalous results is when the data set is split into ever finer categorizations that end up having very few observations in them, and thus subject to much greater variability. For example, should we be overly surprised if Albert Pujols does not seem to bat well in domed stadiums at night when batting second against left-handed pitching? The pundits look for "causes" for these kinds of extremes when they should just be marveling at the beauty of natural variation and the effects of sample size. A similar and probably more important misleading effect occurs when our data are on the effectiveness of some medical treatment, and we try to attribute positive or negative results to ever finer-grained classifications of our clinical subjects.

Random processes are a fundamental part of nature and are ubiquitous in our day-to-day lives. Most people do not understand them or, worse, fall under an "illusion of control," where one believes they have influence

over how events progress. Thus, we have almost a mystical belief in the ability of a new coach, CEO, or President to "turn things around." Part of these strong beliefs may result from the operation of regression toward the mean, or the natural unfolding of any random process. We continue to get our erroneous beliefs reconfirmed when we attribute cause when none may be present. As humans we all wish to believe we can affect our future, but when events have dominating stochastic components, we are obviously not in complete control. There appears to be a fundamental clash between our ability to recognize the operation of randomness and the need for control in our lives.

## Correlation

The association between two variables measured on the same set of objects is commonly referred to as their correlation and often measured by Pearson's product moment correlation coefficient. Specifically, suppose $Z_{X_1}, \ldots, Z_{X_N}$ and $Z_{Y_1}, \ldots, Z_{Y_N}$ refer to $Z$ scores (i.e., having mean zero and variance one) calculated for our original observational pairs, $(X_i, Y_i)$, $i = 1, \ldots, N$; then the correlation between the original variables, $r_{XY}$, is defined as:

$$r_{XY} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} Z_{X_i} Z_{Y_i},$$

or the average product of the $Z$ scores. As usually pointed out early in any statistics sequence, $r_{XY}$ measures the linearity of any relation that might be present; thus, if some other (nonlinear) form of association exists, different means of assessing it are needed.

In any reasoning based on the presence or absence of a correlation between two variables, it is imperative that graphical mechanisms be used in the form of scatterplots. One might go so far to say that if only the value of $r_{XY}$ is provided and nothing else, we have a *prima facie* case of statistical malpractice. Scatterplots are of major assistance in a number of ways: (a) to ascertain the degree to which linearity might be the type of association present between the variables; this assessment could take the form of directly imposing various scatterplot smoothers and using these to help characterize the association present, if any; (b) to identify outliers or data points that for whatever reason are not reflective of the general pattern exhibited in the scatterplot, and to hopefully figure out why; and (c) to provide a graphical context for assessing the influence of a data point on a correlation, possibly by the size and/or color of a plotting symbol, or contour lines indicating the change in value for the correlation that would result if it were to be removed.

One of the most shopworn adages we hear in any methodology course is that "correlation does not imply causation." It is usually noted that other "lurking" or third variables might affect both $X$ and $Y$, producing a spurious association; also, because $r_{XY}$ is a symmetric measure of association, there is no clue in its value as to the directionality of any causal relationship. For example, we have had some recent revisions in our popular views on the positive effects of moderate drinking; it may be that individuals who otherwise lead healthy lifestyles also drink moderately. Or in a football sports context, "running the ball" does not cause winning; it is more likely that winning causes "running the ball." Teams that get an early lead try to run the ball frequently because it keeps the clock running and decreases the time for an opponent to catch up.

In any multiple variable context, it is possible to derive the algebraic restrictions present among some subset of the variables based on the given correlations for another subset. The simplest case involves three variables, say $X$, $Y$, and $W$. From the basic formula for the partial correlation between $X$ and $Y$ "holding" $W$ constant, an *algebraic* restriction is present on $r_{XY}$ given the values of $r_{XW}$ and $r_{YW}$:

$$r_{XW}r_{YW} - \sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)} \leq r_{XY} \leq r_{XW}r_{YW} + \sqrt{(1 - r_{XW}^2)(1 - r_{XW}^2)}.$$

Note that this is not a probabilistic statement (i.e., it is not a confidence interval); it says that no data set exists where the correlation $r_{XY}$ lies outside of the upper and lower bounds provided by $r_{XW}r_{YW} \pm \sqrt{(1 - r_{XW}^2)(1 - r_{XW}^2)}$. As a numerical example, suppose $X$ and $Y$ refer to height and weight, respectively, and $W$ is some measure of age. If, say, the correlations, $r_{XW}$ and $r_{YW}$ are both .8, then $.28 \leq r_{XY} \leq 1.00$. In fact, if a high correlation value of .64 were observed for $r_{XY}$, should we be impressed about the magnitude of the association between $X$ and $Y$? Probably not; if the partial correlation between $X$ and $Y$ "holding" $W$ constant were computed with $r_{XY} = .64$, a value of zero would be obtained. All the observed high association between $X$ and $Y$ can be attributed to their association with the developmentally driven variable. These very general restrictions on correlations have been known for a very long time and appear, for example, in Yule's first edition (1911) of *An Introduction to the Theory of Statistics* under the title: *Conditions of Consistence Among Correlation Coefficients*. Also, in this early volume, see Yule's chapter on *Fallacies in the Interpretation of Correlation Coefficients*.

A related type of algebraic restriction for a correlation is present when the distribution of the values taken on by the variables includes ties. In the extreme, consider a $2 \times 2$ contingency table, and the fourfold point correlation; this is constructed by using a 0/1 coding of the category information on the two attributes and calculating the usual Pearson correlation. Because of the "lumpy" marginal frequencies present in the $2 \times 2$ table,

the fourfold correlation cannot extend over the complete ±1 range. The achievable bounds possible can be computed (see Carroll, 1961); it may be of some interest descriptively to see how far an observed fourfold correlation is away from its achievable bounds, and possibly, to even normalize the observed value by such a bound.

The bounds of ±1 on a Pearson correlation can be achieved only by data sets demonstrating a perfect linear relationship between the two variables. Another measure that achieves the bounds of ±1 whenever the data sets merely have consistent rank orderings is Guttman's (weak) monotonicity coefficient, $\mu_2$:

$$\mu_2 = \frac{\sum_{i=1}^{n}\sum_{h=1}^{n}(x_h - x_i)(y_h - y_i)}{\sum_{i=1}^{n}\sum_{h=1}^{n}|x_h - x_i|\,|y_h - y_i|},$$

where $(x_h, y_h)$ denote the pairs of values being "correlated" by $\mu_2$. The coefficient, $\mu_2$, expresses the extent to which values on one variable increase in a particular direction as the values on another variable increases, without assuming that the increase is exactly according to a straight line. It varies between –1 and +1, with +1[–1] reflecting a perfect monotonic trend in a positive [negative] direction. The adjective "weak" refers to the untying of one variable without penalty. In contrast to the Pearson correlation, $\mu_2$ can equal +1 or –1, even though the marginal distributions of the two variables differ from one another. When the Pearson correlation is +1.00 or –1.00, $\mu_2$ will have the same value; in all other cases, the absolute value of $\mu_2$ will be higher than that of the Pearson correlation, including the case of a fourfold point correlation. Here, $\mu_2$ reduces to what is called Yule's *Q* (which is a special case of the Goodman–Kruskal gamma statistic for a $2 \times 2$ contingency table [a measure of rank order consistency]).

There are several other correlational pitfalls that seem to occur in various forms whenever we try to reason through data sets involving multiple variables. We briefly mention four of these areas in the sections to follow.

## Illusory Correlation

An *illusory correlation* is present whenever a relationship is seen in data where none exists. Common examples would be between membership in some minority group and rare and typically negative behavior, or in the endurance of stereotypes and an overestimation of the link between group membership and certain traits. Illusory correlations seem to depend on the novelty or uniqueness of the variables considered. Some four decades ago, Chapman and Chapman (1967, 1969) studied such false associations in relation to psychodiagnostic signs seen in projective tests. For example, in the Draw-a-Person test, a client draws a person on a blank piece of paper. Some psychologists believe that drawing a person with big eyes is a sign

of paranoia. Such a correlation is illusionary but very persistent. When data that are deliberately uncorrelated are presented to college students, the same diagnostic signs are found that some psychologists still believe in. It is of some historical interest to know that this very same notion of illusory correlation has been around since the early 1900s—see, for example, Yule's first edition (1911) of *An Introduction to the Theory of Statistics* and the chapter entitled: *Illusory Associations*.

There are several faulty reasoning relatives for the notion of an illusory correlation. One is *confirmation bias* where there are tendencies to search for, interpret, and remember information only in a way that confirms one's preconceptions or working hypotheses. No one will soon forget the country's collective confirmation bias in identifying "weapons of mass destruction" in the run-up to the Iraq war; this is related to the "I'm not stupid" fallacy that rests on the belief that if one is mistaken, one must therefore be stupid, and we generally believe that we are not stupid—witness the prosecutor who refuses to drop charges against an obviously innocent suspect because otherwise, he or she would need to admit error and wasted effort. At an extreme, we have (the trap of) *apophenia*, or seeing patterns or connections in random or meaningless data. A subnotion is *pareidolia*, where vague and random stimuli (often images or sounds) are perceived as significant, for example, the Virgin Mary is seen on a grilled cheese sandwich. One particular problematic realization of apophenia is in epidemiology when residential cancer clusters are identified that rarely, if ever, result in identifiable causes. What seems to be occurring is sometimes labeled the Texas Sharpshooter Fallacy—like a Texas sharpshooter who shoots at the side of a barn and then draws a bull's-eye around the largest cluster of bullet holes. In residential cancer clusters, we tend to notice cases first, for example, multiple cancer patients on the same street, and then define the population base around them. A particularly well-presented piece on these illusory associations is by Atul Gawande in the February 8, 1998, *New Yorker: The Cancer-Cluster Myth*.

## Ecological Correlation

An *ecological correlation* is one calculated between variables that are group means, in contrast to obtaining a correlation between variables measured at an individual level. There are several issues with the use of ecological correlations: They tend to be a lot higher than individual-level correlations, and assuming what is seen at the group level also holds at the level of the individual is so pernicious, it has been labeled the "ecological fallacy" by Selvin (1958). The term *ecological correlation* was popularized from a 1950 article by William Robinson (Robinson, 1950), but the idea has been around for some time (e.g., see the 1939 article by E. L. Thorndike, *On the Fallacy of Imputing Correlations Found for Groups to the*

*Individuals or Smaller Groups Composing Them*). Robinson computed a correlation of .53 between literacy rate and the proportion of the population born outside the United States for the 48 states of the 1930 census. At the individual level, however, the correlation was –.11, so immigrants were on average less literate than their native counterparts. The high ecological correlation of .53 was due to immigrants settling in states with a more literate citizenry. A recent discussion of ecological correlation issues in our present political climate is the entertaining (at least for statisticians) piece in the *Quarterly Journal of Political Science* by Gelman, Shor, Bafumi, and Park (2007): *Rich State, Poor State, Red State, Blue State: What's the Matter With Connecticut*. An expansion of this article in book form is Gelman, Park, Shor, Bafumi, and Cortina (2010; *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*).

A problem related to ecological correlation is the modifiable areal unit problem (MAUP), where differences in spatial units used in the aggregation can cause wide variation in the resulting correlations (e.g., anywhere from minus to plus 1.0). Generally, the manifest association between variables depends on the size of areal units used, with increases as areal unit size gets larger. A related "zone" effect concerns the variation in correlation caused by reaggregating data into different configurations at the same scale. Obviously, the MAUP has serious implications for our abilities to reason with data: When strong relationships exist between variables at an individual level, these can be obscured through aggregation; conversely, aggregation can lead to apparently strong association when none is present. A thorough discussion of the modifiable unit problem appears in Yule and Kendall (1968).

### Restriction of Range for Correlations

The famous psychologist Clark Hull noted in 1928 that psychological tests did not predict job performance very well, with correlations rarely above .30. The implication taken was that tests could never be of much use in personnel selection because job performance could not be predicted very well. In one of the most famous articles in all of industrial and organizational psychology, Taylor and Russell (1939) responded to Hull, noting the existence of the restriction of range problem: In a group selected on the basis of some test, the correlation between test and performance must be lower than it would be in an unselected group. Taylor and Russell provided tables and charts for estimating an unselected from the selected correlation based on how the selection was done (the famous Taylor–Russell charts).

An issue related to the restriction of range in its effect on correlations is the need to deal continually with fallible measurement. Generally, the more unreliable our measures, the lower (or more attenuated) the correlations. The field of psychometrics has for some many decades provided

a mechanism for assessing the effects of fallible measurement through its "correction for attenuation": the correlation between "true scores" for our measures is the observed correlation divided by the square roots of their reliabilities. Various ways are available for estimating reliability, so implementing attenuation corrections is an eminently feasible enterprise. Another way of stating this correction is to note that any observed correlation must be bounded above by the square root of the product of the reliabilities. Obviously, if reliabilities are not very good, observed correlations can never be very high.

Another type of range restriction problem (Figure 4.1) is observed in the empirical fact of a negative correlation between Law School Admission Test (LSAT) scores and undergraduate grade point average (UGPA) within almost all law schools. Does this mean that the worse you perform in college courses, the better you will do on the LSAT? Well, no; it is because if you did very well on both, you went to Harvard, and if you did poorly on both, you did not get into law school. So at all other law schools, there were admittees who did relatively better on one than on the other. A graph of the LSAT scores versus UGPA shows thin bands running from the upper left to the lower right representing each law school, with the better schools higher up on both; the overall picture, however, is a very positive data swirl with the lower triangle not admitted.

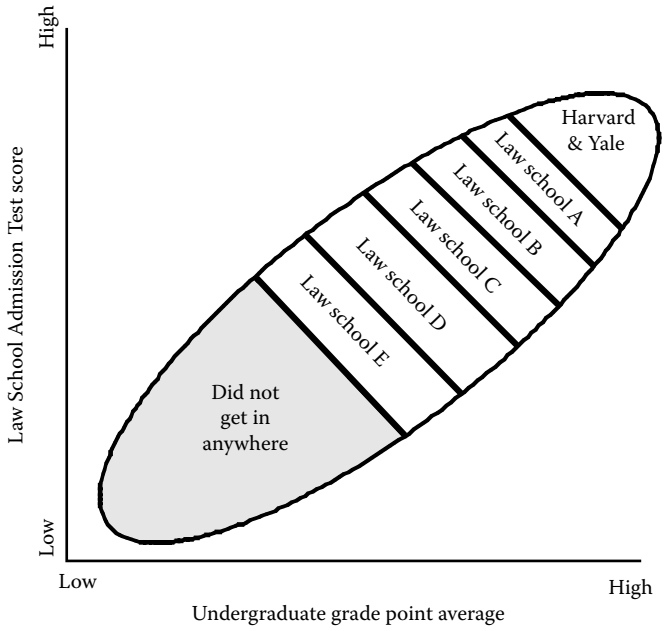



**FIGURE 4.1**
A restriction of range issue between undergraduate grade point averages and the Law School Admission Test.

### Odd Correlations

A recent article (Vul, Harris, Winkielman, & Pashler, 2009) in a journal from the Association for Psychological Science, *Perspectives on Psychological Science*, has the intriguing title of *Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition* (renamed from the earlier and more controversial, *Voodoo Correlations in Social Neuroscience*). These authors comment on the extremely high (e.g., > .8) correlations reported in the literature between brain activation and personality measures and point out the fallaciousness of how they were obtained. Typically, huge numbers of separate correlations were calculated, and only the mean of those correlations exceeding some threshold (based on a very small significance level) is reported. It is tautological that these correlations selected for size must be large in their average value. With no cross-validation attempted to see the shrinkage expected in these measures on new samples, we have sophistry at best. Any of the usual understanding of yardsticks provided by the correlation or its square, the proportion of shared variance, is inappropriate. In fact, as noted by Vul et al. (2009), these inflated mean correlations typically exceed the upper bounds provided by the correction for attenuation based on what the reliabilities should be for the measures being correlated.

When a correlation reported in the literature seems odd, it is incumbent on a literate consumer of such information to understand why. Sometimes it is as simple as noting the bias created by the selection process as in the fMRI correlations, and that such selection is not being mitigated by any cross-validation. Or, possibly, inflated or deflated association measures may occur because of the use of ecological correlations or modifiable areal units, restriction of range, the fallibility of the behavioral measures, the presence of a nonlinear relationship, and so on. The reason behind apparent correlational artifacts can be subtle and require a careful explication of the processes leading to the measures being correlated and on what objects. For example, if correlations are being monitored over time, and the group on which the correlations are based changes composition, the effects could be dramatic. Such composition changes might be one of different sex ratios, immigrant influxes, economic effects on the available workforce, age, and so on. One particularly unusual example is discussed by Dawes (1975) on the relation between graduate admission variables and future success. Because admission criteria tend to be compensatory (where good values on certain variables can make up for not so good values on others), the covariance structure among admissions variables in the selected group is unusual in that it involves negative correlations. As argued nicely by Dawes, it must be the case that the variables used to admit graduate students have low correlation with future measures of success.

A related odd correlational effect (Figure 4.2) occurs in graduate admissions for departments that specialize in technical subjects—there is a

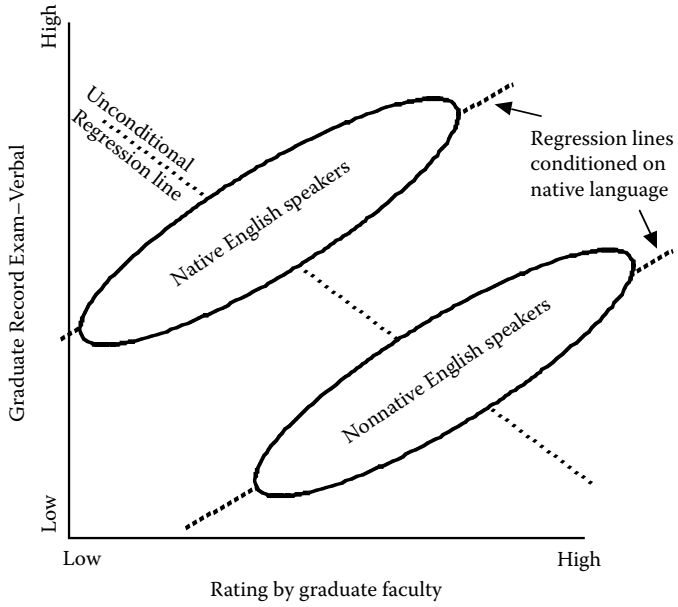



**FIGURE 4.2**
The relation between rating by graduate faculty and the Graduate Record Examination–Verbal.

negative correlation of performance in graduate school (as judged by faculty ratings) and Graduate Record Examination–Verbal (GRE-V) scores. Does this imply that faculty judge badly? Or that the poorer your English proficiency, the better you will do in graduate school? The answer is more subtle and is generated by the large number of students with foreign (often Chinese) backgrounds, whose performance on the GRE-V may be relatively poor but who do well in graduate school. This interpretation is confirmed when we condition on the binary variable "Native English Speaker" or "Not" and find that the correlation is strongly positive within either of the two classes. Again, this becomes clear with a graph that shows two tight ovals at different heights corresponding to the two language groups, but the overall regression line runs across the two ovals and in the opposite direction.

## Prediction

The attempt to predict the values on some (dependent) variable by a function of (independent) variables is typically approached by simple or multiple regression for one and more than one predictor, respectively. The

most common combination rule is a linear function of the independent variables obtained by least squares; that is, the linear combination minimizes the sum of the squared residuals between the actual values on the dependent variable and those predicted from the linear combination. In the case of simple regression, scatterplots again play a major role in assessing linearity of the relationship, the possible effects of outliers on the slope of the least squares line, and the influence of individual objects in its calculation. The regression slope, in contrast to the correlation, is neither scale invariant nor symmetric in the dependent and independent variables. One usually interprets the least squares line as one of expecting, for each unit change in the independent variable, a regression slope change in the dependent variable.

There are several topics in prediction that arise continually when we attempt to reason ethically with fallible multivariable data. We discuss briefly four such areas in the subsections to follow: regression toward the mean, the distinction between actuarial (statistical) and clinical prediction, methods involved in using regression for prediction that incorporate corrections for unreliability, and differential prediction effects in selection based on tests.

### Regression Toward the Mean

Regression toward the mean is a phenomenon that will occur whenever dealing with (fallible) measures with a less than perfect correlation. The word *regression* was first used by Sir Francis Galton in his 1886 article, *Regression Toward Mediocrity in Hereditary Stature*, where he showed that heights of children from very tall or short parents would regress toward mediocrity (i.e., toward the mean)—exceptional scores on one variable (parental height) would not be matched with such exceptionality on the second (child height). This observation is purely due to the fallibility for the various measures (i.e., the lack of a perfect correlation between the heights of parents and their children).

Regression toward the mean is a ubiquitous phenomenon and is given the name *regressive fallacy* whenever cause is ascribed where none exists. Generally, interventions are undertaken if processes are at an extreme, for example, a crackdown on speeding or drunk driving as fatalities spike; treatment groups formed from individuals who are seriously depressed; individuals selected because of extreme behaviors, both good or bad; and so on. In all such instances, whatever remediation is carried out will be followed by some more moderate value on a response variable. Whether the remediation was itself causative is problematic to assess given the universality of regression toward the mean.

There are many common instances where regression may lead to invalid reasoning: I went to my doctor and my pain has now lessened;

I instituted corporal punishment and behavior has improved; he was jinxed by a *Sports Illustrated* cover because subsequent performance was poorer (i.e., the "sophomore jinx"); although he had not had a hit in some time, he was "due," and the coach played him; and on and on. More generally, any time one optimizes with respect to a given sample of data by constructing prediction functions of some kind, there is an implicit use and reliance on data extremities. In other words, the various measures of goodness of fit or prediction we might calculate need to be cross-validated either on new data or a clever sample reuse strategy such as the well-known jackknife or bootstrap procedures. The degree of "shrinkage" we see in our measures based on this cross-validation is an indication of the fallibility of our measures and the adequacy of the given sample sizes.

The misleading interpretive effects engendered by regression toward the mean are legion, particularly when we wish to interpret observational studies for some indication of causality. There is a continual violation of the old adage that "the rich get richer and the poor get poorer," in favor of "when you are at the top, the only way is down." Extreme scores are never as extreme as they first appear. Many of these regression artifacts are explicated in the cautionary source, *A Primer on Regression Artifacts* (Campbell & Kenny, 2002), including the various difficulties encountered in trying to equate intact groups by matching or analysis of covariance. Statistical equating creates the illusion but not the reality of equivalence. As summarized by Campbell and Kenny, "the failure to understand the likely direction of bias when statistical equating is used, is one of the most serious difficulties in contemporary data analysis."

There are a variety of phrases that seem to get attached whenever regression toward the mean is probably operative. We have the "winner's curse," where someone is chosen from a large pool (e.g., of job candidates), who then does not live up to expectation; or when we attribute some observed change to the operation of "spontaneous remission." As Campbell and Kenny note, "Many a quack has made a good living from regression toward the mean." Or, when a change of diagnostic classification results on repeat testing for an individual given subsequent one-on-one tutoring (e.g., after being placed in a remedial context), or, more personally, there is "editorial burn out" when someone is chosen to manage a prestigious journal at the apex of one's career, and things go quickly downhill from that point forward.

## Actuarial Versus Clinical Prediction

Paul Meehl in his classic 1954 monograph, *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of

data combination, such as multiple regression, outperform (expert) clinical prediction. The enormous amount of literature produced since the appearance of this seminal contribution has uniformly supported this general observation; similarly, so have the extensions suggested for combining data in ways other than by multiple regression; for example, by much simpler unit weighting schemes (Wainer, 1976) or those using other prior weights. It appears that individuals who are conversant in a field are better at selecting and coding information than they are at integrating it. Combining such selected information in some more mechanical manner will generally do better than the person choosing such information in the first place. This conclusion can be pushed further: If we formally model the predictions of experts using the same chosen information, we can generally do better than the experts themselves. Such formal representations of what a judge does are called "paramorphic."

In an influential review article, Dawes (1979) discussed what he called proper and improper linear models and argued for the "robust beauty of improper linear models." A proper linear model is one obtained by some optimization process, usually least squares; improper linear models are not "optimal" in this latter sense, and typically have their weighting structures chosen by a simple mechanism; for example, random or unit weighting. Again, improper linear models generally outperform clinical prediction, but even more surprisingly, improper models typically outperform proper models in cross-validation. What seems to be the reason is the notorious instability of regression weights with correlated predictor variables, even if sample sizes are very large. Generally, we know that simple averages are more reliable than individual observations, so it may not be so surprising that simple unit weights are likely to do better on cross-validation than those found by squeezing "optimality" out of a sample. Given that the *sine qua non* of any prediction system is its ability to cross-validate, the lesson may be obvious—statistical optimality with respect to a given sample may not be the best answer when we wish to predict well.

The idea that statistical optimality may not lead to the best predictions seems counterintuitive, but as argued well by Roberts and Pashler (2000), just the achievement of a good fit to observations does not necessarily mean we have found a good model. In fact, because of the overfitting of observations, choosing the model with the absolute best fit is apt to result in poorer predictions. The more flexible the model, the more likely it is to capture not only the underlying pattern but also unsystematic patterns such as noise. A single general purpose tool with many adjustable parameters is prone to instability and greater prediction error as a result of high error variance. An observation by John von Neumann is particularly germane: "With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk." More generally, this notion that "less is

more" is difficult to get one's head around, but as Gigerenzer and others have argued (e.g., see Gigerenzer & Brighton, 2009), it is clear that simple heuristics, such as "take the best," can at times be more accurate than complex procedures. All the work emanating from the idea of the "robust beauty of improper linear models" and sequelae may force some reassessment of what the normative ideals of rationality might be; most reduce to simple cautions about overfitting one's observations, and then hoping for better predictions because an emphasis has been placed on immediate optimality instead of the longer-run goal of cross-validation.

## Incorporating Reliability Corrections in Prediction

There are two aspects of variable unreliability in the context of prediction that might have consequences for ethical reasoning. One is in estimating a person's true score on a variable; the second is in how regression might be handled when there is measurement error in the independent and/or dependent variables. In both of these instances, there is an implicit underlying model for how any observed score, $X$, might be constructed additively from a true score, $T_X$, and an error score, $E_X$, where $E_X$ is typically assumed uncorrelated with $T_X$: $X = T_X + E_X$. When we consider the distribution of an observed variable over, say, a population of individuals, there are two sources of variability present in the true and the error scores. If we are interested primarily in structural models among true scores, then some correction must be made because the common regression models implicitly assume that variables are measured without error.

The estimation, $\hat{T}_X$, of a true score from an observed score, $X$, was derived using the regression model by Kelley in the 1920s (see Kelley, 1947), with a reliance on the observation that the squared correlation between observed and true score is the reliability. If we let $\hat{\rho}$ be estimated reliability, Kelley's equation can be written as $\hat{T}_X = \hat{\rho}X + \left(1 - \hat{\rho}\right)\bar{X}$, where $\bar{X}$ is the mean of the group to which the individual belongs. In other words, depending on the size of $\hat{\rho}$, a person's estimate is compensated for by where they are in relation to the group—upward if below the mean; downward if above. The application of this statistical tautology in the examination of group differences provides such a surprising result to the statistically naive, that this equation has been called *Kelley's Paradox* (see Wainer, 2005, Chapter 10). We might note that this notion of being somewhat punitive of performances better than the group to which one supposedly belongs was not original with Kelley but was known at least 400 years earlier; in the words of Miguel de Cervantes (1547–1616): "Tell me what company you keep and I'll tell you what you are."

In the topic of errors in variables regression, we try to compensate for the tacit assumption in regression that all variables are measured without error. Measurement error in a response variable does not bias the

regression coefficients per se, but it does increase standard errors, and thereby reducing power. This is generally a common effect: unreliability attenuates correlations and reduces power even in standard ANOVA paradigms. Measurement error in the predictor variables biases the regression coefficients. For example, for a single predictor, the observed regression coefficient is the "true" value multiplied by the reliability coefficient. Thus, without taking account of measurement error in the predictors, regression coefficients will generally be underestimated, producing a biasing of the structural relationship among the true variables. Such biasing may be particularly troubling when discussing econometric models where unit changes in observed variables are supposedly related to predicted changes in the dependent measure; possibly the unit changes are more desired at the level of the true scores.

### Differential Prediction Effects in Selection

One area in which prediction is socially relevant is in selection based on test scores, whether for accreditation, certification, job placement, licensure, educational admission, or other high-stakes endeavors. We note that most of these discussions about fairness of selection need to be phrased in terms of regression models relating a performance measure to a selection test, and whether the regressions are the same over all the identified groups of relevance, for example, ethnic, gender, age, and so on. Specifically, are slopes and intercepts the same; if so or if not, how does this affect the selection mechanism being implemented, and whether it can be considered fair? It is safe to say that depending on the pattern of data within groups, all sorts of things can happen; generally, an understanding of how a regression/selection model works with this kind of variation is necessary for a literate discussion of its intended or unintended consequences. To obtain a greater sense of the complications that can arise, the reader is referred to Allen and Yen (2001; Chapter 4.4, *Bias in Selection*).

## Data Presentation and Interpretation

The goal of statistics is to gain understanding from data; the methods of presentation and analyses used should not only allow us to "tell the story" in the clearest and fairest way possible, but more primarily, to help learn what the story is in the first place. When results are presented, there is a need to be sensitive to the common and maybe not so common missteps that result from a superficial understanding and application of the methods in statistics. It is insufficient to just "copy and paste" without

providing context for how good or bad the methods are that are being used and without understanding what is behind the procedures producing the numbers. We will present in this introductory section some of the smaller pitfalls to be avoided; a number of larger areas of concern will be treated in separate subsections:

1. Even very trivial differences will be "significant" when sample sizes are large enough. Also, significance should never be confused with importance; the current emphasis on the use of confidence intervals and the reporting of effect sizes reflects this point. (For a further discussion of this topic, see Cumming & Fidler, Chapter 11, this volume.)
2. As some current textbooks still report inappropriately, a significance test does not evaluate whether a null hypothesis is true. A *p* value measures the "surprise value" of a particular observed result conditional on the null hypothesis being true.
3. *Degrees-of-freedom* do not refer to the number of independent observations within a data set; the term indicates how restricted the quantities are that are being averaged in computing various statistics; for example, sums of squares between or within groups.
4. Although the CLT comes to the assistance of robustness issues when dealing with means, the same is not true for variances. The common tests on variances are notoriously nonrobust and should never be used; robust alternatives are available in the form of sample-reuse methods such as the jackknife and bootstrap.
5. Do not carry out a test for equality of variances before performing a two-independent samples *t* test. A quote, usually attributed to George Box, comments on the good robustness properties of the *t* test in relation to the nonrobustness of the usual tests for variances: "to test for equality of variances before carrying out an independent samples *t* test is like putting a row boat out on the ocean to see if it is calm enough for the Queen Mary."
6. Measures of central tendency and dispersion, such as the mean and variance, are not *resistant* in that they are influenced greatly by extreme observations; the median and interquartile range, on the other hand, are resistant, and each observation counts the same in the calculation of the measure.
7. Do not ignore the repeated measures nature of your data and just use methods appropriate for independent samples. For example, do not perform an independent samples *t* test on "before" and "after" data in a time series intervention study. Generally,

the standard error of a mean difference must include a correction for correlated observations, as routinely done in a paired (matched samples) *t* test. (For more development of these issues, see Goldstein, Chapter 13, this volume.)

8. The level of the measurement model used for your observations limits meaningful inferences. For example, interpreting the relative sizes of differences makes little sense on data measured with a model yielding only nominal or ordinal level characteristics.
9. Do not issue blanket statements as to the impossibility of carrying out reasonable testing, confidence interval construction, or cross-validation. It is almost always now possible to use resampling methods that do not rely on parametric models or restrictive assumptions, and which are computer implemented for immediate application. The appropriate statement is not that "This can't be done," but rather, "I don't know how to do this as yet."
10. Keep in mind the distinctions between fixed and random effects models and the differing test statistics they may necessitate. The output from some statistical packages may use a default understanding of how the factors are to be interpreted. If your context is different, then appropriate calculations must be made, sometimes "by hand." To parody the Capital One Credit Card commercial: "What's in your denominator?"
11. Do not report all of the eight or so decimal places given in typical computer output. Such false precision (or spurious accuracy) is a dead giveaway that you really do not know what you are doing. Two decimal places are needed at most, and often, only one is really justified. As an example, consider how large a sample is required to support the reporting of a correlation to more than one decimal place (answer: given the approximate standard error of $\frac{1}{\sqrt{n}}$, a sample size greater than 400 would be needed to give a 95% confidence interval of ± 0.1).
12. It is wise generally to avoid issuing statements that might appear to be right, but with some deeper understanding, are just misguided:
    a. "Given the huge size of a population, it is impossible to achieve accuracy with a sample"; this reappears regularly with the discussion of undercount and the census.
    b. "It is incumbent on us to always divide by $n - 1$ when calculating a variance to give the 'best' estimator"; well, if you divide by $n + 1$, the estimator has a smaller expected error of estimation, which to many is more important than just being

"unbiased." Also, why is it that no one ever really worries that the usual correlation coefficient is a "biased" estimate of its population counterpart?

c. "ANOVA is so robust that all of its assumptions can be violated at will"; although it is true that normality is not that crucial if sample sizes are reasonable in size (and the CLT is of assistance), and homogeneity of variances does not really matter as long as cell sizes are close, the independence of errors assumption is critical and one can be led very far astray when it does not hold—for intact groups, spatial contexts, and repeated measures. (Again, for further discussion, see Goldstein, Chapter 13, this volume.)

d. Do not lament the dearth of one type of individual from the very upper scores on some test without first noting possible differences in variability. Even though mean scores may be the same for groups, those with even slightly larger variances will tend to have more representatives in both the upper and lower echelons.

13. Avoid using one-tailed tests. Even the carriers of traditional one-tailed hypotheses, the chi-square and *F* distributions, have two tails, and both ought to be considered. The logic of hypothesis testing is that if an event is sufficiently unlikely, we must reconsider the truth of the null hypothesis. Thus, for example, if an event falls in the lower tail of the chi-square distribution, it implies that the model fits too well. If investigators had used two-tailed tests, the data manipulations of Cyril Burt may have been uncovered much earlier.

In concluding these introductory comments about the smaller missteps to be avoided, we note the observations of Edward Tufte on the ubiquity of PowerPoint (PP) for presenting quantitative data and the degradation it produces in our ability to communicate (Tufte, 2006, his italics):

> *The PP slide format has the worst signal/noise ratio of any known method of communication on paper or computer screen*. Extending PowerPoint to embrace paper and internet screens pollutes those display methods. (p. 26)

Generally PP is poor at presenting statistical evidence and is no replacement for more detailed technical reports, data handouts, and the like. It is now part of our "pitch culture," where, for example, we are sold on what drugs to take but are not provided with the type of detailed numerical evidence we should have for an informed decision about benefits and

risks. In commenting on the obscuring of important data that surrounded the use of PP-type presentations to give the crucial briefings in the first shuttle accident of Challenger in 1986, Richard Feyman noted (reported in Tufte, 2006):

> Then we learned about "bullets"—little black circles in front of phrases that were supposed to summarize things. There was one after another of these little goddamn bullets in our briefing books and on slides. (p. 17)

## Multivariable Systems

Whenever results are presented within a multivariate context, it is important to remember there is a system present among the variables, and this has a number of implications for how we proceed: Automated systems that cull through collections of independent variables to locate the "best" regression equations (e.g., by forward selection, backward elimination, or the hybrid of stepwise regression) are among the most misused statistical methods available in all the common software packages. They offer a false promise of blind theory building without user intervention, but the incongruities present in their use are just too great for this to be a reasonable strategy of data analysis: (a) one does not necessarily end up with the "best" prediction equations for a given number of variables; (b) different implementations of the process do not necessarily end up with the same equations; (c) given that a system of interrelated variables is present, the variables not selected cannot be said to be unimportant; (d) the order in which variables enter or leave in the process of building the equation does not necessarily reflect their importance; and (e) all the attendant significance testing and confidence interval construction methods become completely inappropriate (see Freedman, 1983).

Several methods, such as the use of Mallow's $C_p$ statistic for "all possible subsets (of the independent variables) regression," have some possible mitigating effects on the heuristic nature of the blind methods of stepwise regression. They offer a process of screening all possible equations to find the better ones, with compensation for the differing numbers of parameters that need to be fit. Although these search strategies offer a justifiable mechanism for finding the "best" according to ability to predict a dependent measure, they are somewhat at cross-purposes for how multiple regression is typically used in the behavioral sciences. What is important is in the structure among the variables as reflected by the regression, and not so much in squeezing the very last bit of variance accounted for out of our methods. More pointedly, if we find a "best" equation with fewer than the maximum number of available independent

variables present, and we cannot say that those not chosen are less important than those that are, then what is the point?

A more pertinent analysis was demonstrated by Efron and Gong (1983) in which they bootstrapped the entire model-building process. They showed that by viewing the frequency with which each independent variable finds its way into the model, we can assess the stability of the choice of variables. Examining the structure of the independent variables through, say, a principal component analysis will alert us to irreducible uncertainty as a result of high covariance among predictors. This is always a wise step, done in conjunction with bootstrapping, but not instead of it.

The implicit conclusion of the last argument extends more generally to the newer methods of statistical analysis that seem to continually demand our attention, for example, in hierarchical linear modeling, nonlinear methods of classification, procedures that involve optimal scaling, and so on. When the emphasis is solely on getting better "fit" or increased prediction capability, thereby modeling "better," the methods may not be of much use in "telling the story" any more convincingly. And that should be the ultimate purpose of any analysis procedure we choose. Also, as Roberts and Pashler (2000) note rather counterintuitively, "goodness of fit" does not necessarily imply "goodness of model."

Even without the difficulties presented by a multivariate system when searching through the set of independent variables, there are several admonitions to keep in mind when dealing with a single equation. The most important may be to remember that regression coefficients cannot be interpreted in isolation for their importance using their sizes, even when based on standardized variables (i.e., those that have been *Z*-scored). Just because one coefficient is bigger than another does not imply it is therefore more important. For example, consider the task of comparing the relative usefulness of the Scholastic Aptitude Test (SAT) scores and high school grade point averages (HSGPAs) in predicting freshmen college grades. Both independent variables are highly correlated; so when grades are predicted with SAT scores, a correlation of about .7 is found. Correlating the residuals from this prediction with HSGPA gives a small value. It would be a mistake to conclude from this that SAT is a better predictor of college success than HSGPA. If the order of analysis is reversed, we would find that HSGPA correlates about .7 with freshmen grades, and the residuals from this analysis have only a small correlation with SAT score.

If we must choose between these two variables, or try to evaluate a claim that one variable is more important than another, it must be from some other basis. For example, SAT scores are like the product of an experiment; they can be manipulated and improved. Flawed test items can be discovered and elided. But HSGPAs are like the result of an observational study; they are just found, lying on the ground. We are never sure exactly what they mean. If one teacher harbors a secret bias and gives students of

a particular ilk grades that do not represent their true accomplishments, how are we to know? There are some formal methods that can at times help reduce our ignorance. We will discuss them next, but first remember that no formal procedure guarantees success in the face of an unthinking analysis.

The notion of importance may be explored by comparing models with and without certain variables present, and comparing the changes in variance-accounted-for that ensue. Similarly, the various significance tests for the regression coefficients are not really interpretable independently, for example, a small number of common factors may underlie all the independent variables and thus, generate significance for all the regression coefficients. In its starkest form, we have the one, two, and three asterisks scattered around in a correlation matrix, suggesting an ability to evaluate each correlation by itself without consideration of the multivariable system that the correlation matrix reflects in its totality. Finally, for a single equation, the size of the squared multiple correlation ($R^2$) gets inflated by the process of optimization and needs to be adjusted, particularly when sample sizes are small. One beginning option is to use the commonly generated Wherry "adjusted $R^2$," which makes the expected value of $R^2$ zero when the true squared multiple correlation is itself zero. Note that the name of "Wherry's shrinkage formula" is a misnomer because it is not a measure based on any process of cross-validation. A cross-validation strategy is now routine in software packages, such as SYSTAT, using the "hold out one-at-a-time" mechanism. Given the current ease of implementation, such cross-validation processes should be routinely carried out.

## Graphical Presentation

The importance of scatterplots in evaluating the association between variables was reiterated several times in our earlier discussions of correlation and prediction. Generally, graphical and other visual methods of data analysis are central to an ability to tell what data may be reflecting and what conclusions are warranted. In a time when graphical presentation may have been more expensive than it is now, it was common to only use summary statistics, even when various reporting rules were followed; for example, "never present just a measure of central tendency without a corresponding measure of dispersion." Or, in providing the results of a poll, always give the margin of error (usually, the 95% confidence interval) to reflect the accuracy of the estimate based on the sample size being used. If data are not nicely unimodal, however, more is needed than just means and variances. Both "stem and leaf" and "box and whisker" plots are helpful in this regard and should be routinely used for data presentation.

Several egregious uses of graphs for misleading presentations were documented many years ago in the very popular book by Darrell Huff,

*How to Lie with Statistics* (1954), and updated in Wainer's oft-cited 1984 classic from *The American Statistician, How to Display Data Badly* (also, see Chapter 1 in Wainer, 1997/2000). Both of these deal with visual representation and how graphs can be used to distort; for example, by truncating bottoms of line or bar charts, so differences are artificially magnified, or using two- and three-dimensional objects to compare values on a one-dimensional variable where images do not scale the same way as do univariate quantities. Tufte (e.g., see Tufte, 1983) has lamented on the poor use of graphics that use "chart junk" for questionable visual effect, or gratuitous color or three dimensions in bar graphs that do not represent anything real. In extending some of these methods of misrepresentation to the use of maps, it is particularly easy to deceive given the effects of scale-level usage, ecological correlation, and the modifiable areal unit problem. What should be represented generally in our graphs and maps must be as faithful as possible to the data represented, without the distracting application of unnecessary frills that do not communicate any information of value.

There is one particularly insidious use of a graphical format that almost always misleads: the double *y*-axis plot. In this format there are two vertical axes, one on the left and one on the right, depicting two completely different variables—say, death rates over time for smokers shown on the left axis (time is on the horizontal axis) and death rates for nonsmokers shown on the right axis. Because the scales on the two vertical axes are independent, they can be chosen to show anything the graph maker wants. Compare the first version in Figure 4.3 (after the Surgeon General's report on the dangers of smoking) with the second in Figure 4.4, prepared by someone attentive to the needs of big tobacco that uses the double *y*-axis format. Few other graphic formats lend themselves so easily to the misrepresentation of quantitative phenomena.

In providing data in the form of matrices, such as subject by variable, we should consider the use of "heat maps," where numerical values, assumed commensurable over variables, are mapped into color spectra reflecting magnitude. The further imposing of nondestructively obtained orderings on rows and columns to group similar patches of color together can lead to useful data displays. A survey of the history of heat maps, particularly as developed in psychology, has been given by Wilkinson and Friendly (2009); this article should be mandatory reading in any part of a statistics course concerned with accurate and informative graphical data presentation. Also, see Bertin (1973/1983), Tufte (1983, 1990, 1996), Tukey (1977), and Wainer (1997, 2005, 2009).

## Problems With Multiple Testing

A difficulty encountered with the use of automated software analyses is that of multiple testing, where the many significance values provided are

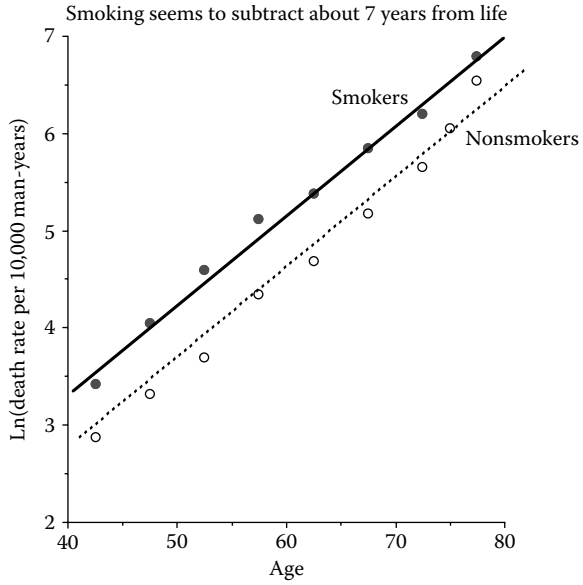



**FIGURE 4.3**
A graph showing that smokers die sooner than nonsmokers.

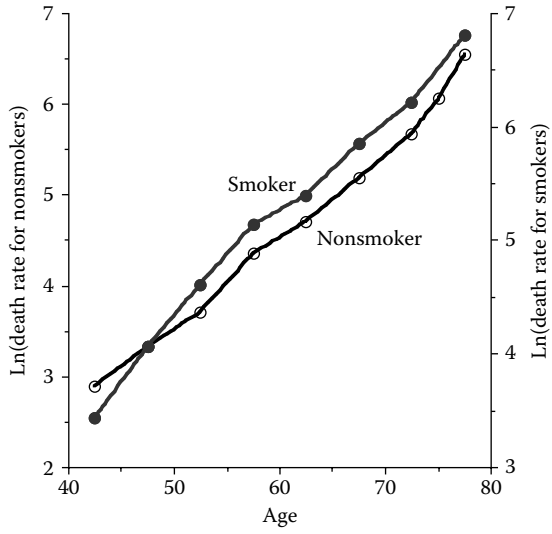



**FIGURE 4.4**
A graph showing that aging is the primary cause of death.

all given as if each were obtained individually without regard for how many tests were performed. This situation gets exacerbated when the "significant" results are then culled, and only these are used in further analysis. A good case in point was reported earlier in the section on odd correlations where highly inflated correlations get reported in fMRI studies because an average is taken only over those correlations selected to have reached significance according to a stringent threshold. Such a context is a clear violation of a dictum given in any beginning statistics class: You cannot legitimately test a hypothesis on the same data that first suggested it.

Exactly the same issue manifests itself, although in a more subtle, implicit form, in the modern procedure known as data mining. Data mining consists of using powerful graphical methods to view high-dimensional data sets of moderate to large size, looking for interesting features. When such a feature is uncovered, it is isolated and saved—a finding! Implicit in the search, however, are many, many comparisons that the viewer makes and decides are not interesting. Because the searching and comparing is done in real time, it is difficult to keep track of how many "insignificant" comparisons were discarded before alighting on a significant one. Without knowing how many, we cannot judge the significance of the interesting features found without an independent confirmatory sample. Such independent confirmation is too rarely done.

To be more formal about the problem of multiple testing, suppose there are $K$ hypotheses to test, $H_1, \ldots, H_K$, and, for each, we set the criterion for rejection at the fixed Type I error value of $\alpha_k$, $k = 1, \ldots, K$. If the events, $A_1, \ldots, A_K$, are defined as: $A_k$ is the incorrect rejection of $H_k$ (i.e., rejection when it is true), the Bonferroni inequality gives:

$$P(A_1 \text{ or } \ldots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) = \sum_{k=1}^{K} \alpha_k.$$

Noting that the event ($A_1$ or … or $A_K$) can be verbally restated as one of "rejecting incorrectly *one or more* of the hypotheses," the experimentwise (or overall) error rate is bounded by the sum of the $K$ alpha values set for each hypothesis. Typically, we set $\alpha_1 = \ldots = \alpha_K = \alpha$, and the bound is then $K\alpha$. Thus, the usual rule for controlling the overall error rate through the Bonferroni correction sets the individual $\alpha$ at some small value, for example, $.05/K$; the overall error rate is then guaranteed to be no larger than .05.

The problems of multiple testing and the failure to practice "safe statistics" appear in both blatant and more subtle forms. For example, companies may suppress unfavorable studies until those to their liking occur. There is a possibly apocryphal story that toothpaste companies promoting fluoride in their products in the 1950s did repeated studies until large effects

could be reported for their "look Ma, no cavities" television campaigns. This may be somewhat innocent advertising hype for toothpaste, but when drug or tobacco companies engage in the practice, it is not so innocent and can have a serious impact on our health. It is important to know how many things were tested to assess the importance of those reported. For example, when given only those items from some inventory or survey that produced significant differences between groups, be very wary!

In the framework of multiple testing, there are a number of odd behaviors that people sometimes engage in. We list a few of these below in summary form:

1. It is not legitimate to do a Bonferroni correction post hoc; that is, find a set of tests that lead to significance, and then evaluate just this subset with the correction.
2. Scheffé's method (and relatives) is the only true post hoc procedure to control the overall error rate. An unlimited number of comparisons can be made (no matter whether identified from the given data or not), and the overall error rate remains constant.
3. You cannot look at your data to decide which planned comparisons to do.
4. Tukey's method is not post hoc because you plan to do all possible pairwise comparisons.
5. Even though the comparisons you might wish to test are independent (e.g., they are defined by orthogonal comparisons), the problem of inflating the overall error rate remains; similarly, in performing a multifactor ANOVA or testing multiple regression coefficients, all the tests carried out should have some type of overall error control imposed.
6. It makes no sense to perform a multivariate analysis of variance before you then go on to evaluate each of the component variables one by one. Typically, a multivariate analysis of variance (MANOVA) is completely noninformative as to what is really occurring, but people proceed in any case to evaluate the individual univariate ANOVAs irrespective of what occurs at the MANOVA level—we may not reject the null hypothesis at the overall MANOVA level but then illogically ask where the differences are at the level of the individual variables. Plan to do the individual comparisons beforehand, and avoid the typically noninterpretable overall MANOVA test completely.

We cannot, in good conscience, leave the important topic of multiple comparisons without at least a mention of what is now considered the most useful method—the false discovery rate (Benjamini & Hochberg,

1995). But even this strategy is not up to the most vexing of problems of multiplicity. We have already mentioned data mining as one of these; a second arises in the search for genetic markers. A typical paradigm in this crucial area is to isolate a homogeneous group of individuals, some of whom have a genetic disorder and others do not, and then to see whether one can determine which genes are likely to be responsible. One such study is currently being carried out with a group of 200 Mennonites in Pennsylvania. Macular degeneration is common among the Mennonites, and this sample was chosen so that 100 of them had macular degeneration and a matched sample of 100 did not. The genetic structure of the two groups was very similar, and so the search was on to see which genes were found much more often in the group that had macular degeneration than in the control group. This could be determined with a *t* test. Unfortunately, the power of the *t* test was diminished considerably when it had to be repeated for more than 100,000 separate genes. The Bonferroni inequality was no help, and the false discovery rate, although better, was still not up to the task. The search continues to find a better solution to the vexing problem of multiplicity.

## (Mis-)Reporting of Data

The Association for Psychological Science publishes a series of timely monographs on *Psychological Science in the Public Interest*. One recent issue was from Gerd Gigerenzer and colleagues, entitled *Helping Doctors and Patients Make Sense of Health Statistics* (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2008); it details some issues of statistical literacy as it concerns health, both our own individually as well as societal health policy more generally. Some parts of being statistically literate may be fairly obvious—we know that just making up data, or suppressing information even of supposed outliers without comment, is unethical. The topics touched on by Gigerenzer et al., however, are more subtle; if an overall admonition is needed, it is that "context is always important," and the way data and information are presented is absolutely crucial to an ability to reason appropriately and act accordingly. We touch on several of the major issues raised by Gigerenzer et al. in the discussion to follow.

We begin with a quote from Rudy Giuliani from a New Hampshire radio advertisement that aired on October 29, 2007, during his run for the Republican Presidential nomination (this example was also used by Gigerenzer et al., 2008):

> I had prostate cancer, five, six years ago. My chances of surviving prostate cancer and thank God I was cured of it, in the United States, 82 percent. My chances of surviving prostate cancer in England, only 44 percent under socialized medicine.

Not only did Giuliani not receive the Republican Presidential nomination, he was just plain wrong on survival chances for prostate cancer. The problem is confusion between survival and mortality rates. Basically, higher survival rates with cancer screening do not imply longer life.

To give a more detailed explanation, we define a 5-year survival rate and an annual mortality rate:

Five-year survival rate = (number of diagnosed patients alive after 5 years)/(number of diagnosed patients)

Annual mortality rate = (number of people who die from a disease over 1 year)/(number in the group)

The inflation of a 5-year survival rate is caused by a *lead-time bias*, where the time of diagnosis is advanced (through screening) even if the time of death is not changed. Moreover, such screening, particularly for cancers such as prostate, leads to an *overdiagnosis bias*—the detection of a pseudo-disease that will never progress to cause symptoms in a patient's lifetime. Besides inflating 5-year survival statistics over mortality rates, overdiagnosis leads more sinisterly to overtreatment that does more harm than good (e.g., incontinence, impotence, and other health-related problems).

It is important to keep in mind that screening does not "prevent cancer," and early detection does not diminish the risk of getting cancer. One can only hope that cancer is caught, either by screening or other symptoms, at an early enough stage to help. It is also relevant to remember that more invasive treatments are not automatically more effective. A recent and informative summary of the dismal state and circumstances surrounding cancer screening generally, appeared in *The New York Times*, page one and "above the fold," article by Natasha Singer (Friday, July 17, 2009), *In Push for Cancer Screening, Limited Benefits*.

A major area of concern in the clarity of reporting health statistics is in how the data are framed as relative risk reduction or as absolute risk reduction, with the former usually seeming much more important than the latter. As examples that present the same information:

*Relative risk reduction*—If you have this test every 2 years, it will reduce your chance of dying from the disease by about one third over the next 10 years.

*Absolute risk reduction*—If you have this test every 2 years, it will reduce your chance of dying from the disease from 3 in 1,000 to 2 in 1,000 over the next 10 years.

We also have a useful variant on absolute risk reduction given by its reciprocal, the *number needed to treat*—if 1,000 people have this test every 2 years, one person will be saved from dying from the disease every 10 years.

Because bigger numbers garner better headlines and more media attention, it is expected that relative rather than absolute risks are the norm. It is especially disconcerting, however, to have potential benefits (of drugs, screening, treatments, and the like) given in relative terms, but harm in absolute terms that is typically much smaller numerically. The latter has been called "mismatched framing" by Gigerenzer and colleagues (2008).

An ethical presentation of information avoids nontransparent framing of information, whether unintentional or intentional. Intentional efforts to manipulate or persuade people are particularly destructive, and unethical, by definition. As Tversky and Kahneman (e.g., 1981) have noted many times in their published contributions, framing effects and context have major influences on a person's decision processes. Whenever possible, give measures that have operational meanings with respect to the sample at hand (e.g., the Goodman–Kruskal $\gamma$), and avoid measures that do not, such as odds ratios. This advice is not always followed [see, e.g., Agency for Healthcare Research and Quality's 2008 *National Healthcare Disparities Report*, in which the efficacy of medical care is compared across various groups in plots with the odds ratio as the dependent variable. As might be expected, this section's impact on the public consciousness was severely limited].

In a framework of misreporting data, we have the all too common occurrence of inflated (and sensational) statistics intended to have some type of dramatic effect. As noted succinctly by Joel Best in his article, *Lies, Calculations and Constructions* (2005): "Ridiculous statistics live on, long after they've been thoroughly debunked; they are harder to kill than vampires." We typically see a three-stage process in the use of inflated statistics: first, there is some tale of atrocity (think Roman Polanski's *Rosemary's Baby*); the problem is then given a name (e.g., the presence of satanic cults in our midst); and finally, some inflated and, most likely, incorrect statistic is given that is intended to alarm (e.g., there are well over 150,000 active satanic cults throughout the United States and Canada).

Another issue in the reporting of data is when the context for some statement is important but is just not given (or is suppressed), resulting in a misinterpretation (or at least, an overinterpretation). These examples are legion and follow the types illustrated below:

1. The chances of a married man becoming an alcoholic are double those of a bachelor because 66% of souses are spouses. (This may not be so dramatic when we also note that 75% of all men over 20 are married.)
2. Among 95% of couples seeking divorce, either one or both do not attend church regularly. (This example needs some base rate information to effect a comparison, e.g., what is the proportion of couples generally where one or both do not attend church regularly.)

3. More than 65% of all accidents occur within 25 miles of home and at a speed of 40 miles per hour or less. (An obvious question to ask is where most of one's driving is done.)
4. Hector Luna, who went 2 for 5 and raised his average to .432, had his fourth straight multihit game for the Cardinals, who have won six of seven overall (Associated Press; St. Louis Cardinals vs. Pittsburgh Pirates, April 26, 2004). (Reporting of data should provide a context that is internally consistent; here, the word "raised" is odd.)

### Pitfalls of Software Implementations

Most of our statistical analyses are now done through the use of packages such as SYSTAT, SPSS, or SAS. Because these systems are basically blind to what data you may be analyzing and what questions you may want to ask, it is up to the user to know some of the pitfalls to avoid. For example, just because an analysis of covariance is extremely easy to do, does not mean it should be done or that it is possible to legitimately equate intact groups statistically. Also, just because output may be provided does not automatically mean it should be used. Cases in point are the inappropriate reporting of indeterminate factor scores, the gratuitous number of decimal places typically given, Durbin–Watson tests when the data are not over time, uninformative overall MANOVAs, nonrobust tests for variances, and so on. We mention two more general traps we have seen repeatedly, and which need to be recognized to avoid embarrassment:

1. In the construction of items or variables, the numbers assigned may at times be open to arbitrary keying. For example, instead of using a 1 to 10 scale, where "1" means "best" and "10" "worst," the keying could be reversed so "1" means "worst" and "10" best. When an intercorrelation matrix is obtained among a collection of variables subject to this kind of scoring arbitrariness, it is possible to obtain some pretty impressive (two-group) structures in methods of multidimensional scaling and cluster analysis that are merely artifacts of the keying and not of any inherent meaning in the items themselves. In these situations, it is common to "reverse score" a subset of the items, so, it is hoped, an approximate "positive manifold" is obtained for the correlation matrix; that is, there are few if any negative correlations that cannot be attributed to just sampling error. (The topic of reverse scoring for the ubiquitous Likert scales is noted, at least in passing, in a variety of measurement sources; one recent and readable account is given by Dunn-Rankin, Knezek, Wallace, & Zhang, 2004.)

2. There are certain methods of analysis (e.g., most forms of multidimensional scaling, *K*-means and mixture model cluster analyses, and some strategies involving optimal scaling) that are prone to local optima; that is, a result is presented but one that is not the best possible according to the goodness of fit measure being optimized. The strategies used in the optimization are not able to guarantee global optimality because of the structure of the functions being optimized (e.g., those that are highly nonconvex). One standard method of local optimality exploration is to repeatedly start (randomly) some specific analysis method, observe how bad the local optima problem is for a given data set, and choose the best analysis found for reporting a final result. Unfortunately, none of the current packages (SPSS, SAS, SYSTAT) offer these random start options for all the methods that may be prone to local optima (for a good case in point involving *K*-means clustering, see Steinley, 2003). These local optimality difficulties are one of the reasons for allowing more than the closed analysis systems in graduate statistics instruction, and the general move (or maybe, we should say rush) toward using environments such as MATLAB® and R (or at least to choose packages that allow an exploration of local optima, e.g., MPlus includes a facility for supplying sets of random starting values for model-based mixture analyses).

The ease to which analyses can be done with closed statistical systems requiring little or no understanding of what the "point and clicks" are really giving, may at times be more of an impediment to clear reasoning than it is of assistance. The user does not need to know much before being swamped with copious amounts of output, and with little or no help on how to wade through the results or, when necessary, to engage in further exploration (e.g., in investigating local minima or alternative analyses). One of the main reasons for now using some of the newer statistical environments (such a R and MATLAB®) is that they do not rely on pull-down menus to do one's thinking; instead, they are built up from functions that take various inputs and provide outputs—but you need to know what to ask for and the syntax of the function being used. Also, the source code for the routines is available and can be modified if some variant of an analysis is desired—again, this assumes more than a superficial understanding of how the methods work; these are valuable skills to have when attempting to reason from data. The R environment has become the *lingua franca* for framing cutting-edge statistical development and analysis, and is becoming the major computational tool we need to develop in the graduate-level statistics sequence. It is also open source and free, so there are no additional instructional costs incurred with the adoption of R.

## Simpson's Paradox

In the presentation of multiway contingency tables, an unusual phenomenon occurs so frequently it has been given the label of *Simpson's Paradox* (Simpson, 1951; Yule, 1903). Basically, various relationships that appear to be present when data are conditioned on the levels of one variable either disappear or change "direction" when aggregation occurs over the levels of the conditioning variable. A well-known real-life example is the Berkeley sex bias case involving women applying to graduate school (see Bickel, Hammel, & O'Connell, 1975). Table 4.3 shows the aggregate admission figures for fall 1973: There appears to be a *prima facie* case for bias given the lower rate of admission for women compared with men.

Although there appears to be bias at the aggregate level, the situation becomes less clear once the data are broken down by major (Table 4.4) (these data are for only the top six majors in number of applicants; therefore, the numbers do not add to those in Table 4.3). Here, no department is significantly biased against women, and, in fact, most have a small bias against men; Simpson's paradox has occurred! Apparently, based on Table 4.4, women tend to apply to competitive departments with lower rates of admission among qualified applicants (e.g., English); men tend to apply to departments with generally higher rates of admission (e.g., engineering).

A different example showing a similar point can be given using data on the differential imposition of a death sentence depending on the race of the defendant and the victim (see Table 4.5). These data are from 20 Florida counties during 1976–1977; our source is Radelet (1981), but they are repeated in many categorical data analysis texts (e.g., see Agresti, 2007). Because 12% of white defendants receive the death penalty and only 10% of blacks, at this aggregate level there appears to be no bias against blacks. But when the data are disaggregated, the situation appears to change (Table 4.6), for when we condition on the race of the victim, in both cases the black defendant has the higher probability of receiving the death sentence compared with the white defendant (17% to 13% for white victims; 6% to 0% for black victims). The conclusion one can reach is disconcerting: The value of a victim is worth more if white than if black, and

**TABLE 4.3**

Berkeley Graduate School Admissions Data (1973)—Aggregate

| | Number of Applicants | Percentage Admitted |
|---|---|---|
| Men | 8,442 | 44 |
| Women | 4,321 | 35 |

**TABLE 4.4**

Berkeley Graduate School Admissions Data (1973)—Six Largest Majors

| | Men | | Women | |
|---|---|---|---|---|
| **Major** | **Applicants** | **Percentage Admitted** | **Applicants** | **Percentage Admitted** |
| A | 825 | 62 | 108 | 82 |
| B | 560 | 63 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 272 | 6 | 341 | 7 |

because more whites kill whites, at the aggregate level, there appears to be a slight bias against whites. But for both types of victims, blacks are more likely to receive the death penalty.

Although not explicitly a Simpson's Paradox context, there are similar situations that appear in various forms of multifactor analysis of variance that raise cautions about aggregation phenomena. The simplest dictum is that "you cannot interpret main effects in the presence of interaction." Some softening of this admonition is usually given when the interaction is not disordinal and where the graphs of means do not cross. In these instances, it may be possible to eliminate the interaction by some relatively simple transformation of the data and produce an "additive" model. Because of this, noncrossing interactions might be considered "unimportant." Similarly, the absence of parallel profiles (i.e., when interaction is present) may hinder the other tests for the main effects of coincident and horizontal profiles. Possibly, if the profiles again show only an "unimportant" interaction, such evaluations could proceed.

Although Simpson's Paradox has been known by this name only rather recently (as coined by Colin Blyth in 1972), the phenomenon has been recognized and discussed for well over 100 years; in fact, it has a complete textbook development in Yule's *An Introduction to the Theory of Statistics*, first published in 1911. In honor of Yule's early contribution (Yule, 1903), we sometimes see the title of the Yule–Simpson effect.

**TABLE 4.5**

Death Sentence Imposition for 20 Florida Counties (1976–1977)—Aggregate

| **Defendant** | **Death: Yes** | **Death: No** |
|---|---|---|
| White | 19 (12%) | 141 |
| Black | 17 (10%) | 149 |

**TABLE 4.6**

Death Sentence Imposition for 20 Florida Counties (1976–1977)—Disaggreaged by Victim Race

| Victim | Defendant | Death: Yes | Death: No |
|---|---|---|---|
| White | White | 19 (13%) | 132 |
| White | Black | 11 (17%) | 52 |
| Black | White | 0 (0%) | 9 |
| Black | Black | 6 (6%) | 97 |

## Conclusion

It is hoped a graduate course in statistics prepares students in a number of areas that have immediate implications for the practice of ethical reasoning. We review six broad areas in this concluding section that should be part of any competently taught sequence in the behavioral sciences: (a) formal tools to help think through ethical situations; (b) a basic understanding of the psychology of reasoning and how it may differ from that based on a normative theory of probability; (c) how to be (dis)honest in the presentation of information and how to avoid obfuscation; (d) some ability to ferret out specious argumentation when it has a supposed statistical basis; (e) the deleterious effects of culling in all its various forms (e.g., the identification of "false positives"), and the subsequent failures to either replicate or cross-validate; and (f) identifying plausible but misguided reasoning from data or from other information presented graphically.

One of the trite quantitative sayings that may at times drive individuals "up a wall" is when someone says condescendingly, "just do the math." Possibly, this saying can become a little less obnoxious when reinterpreted to mean working through a situation formally rather than just giving a quick answer based on first impressions that may be wrong. An example of this may help. In 1990, Craig Whitaker wrote a letter to Marilyn vos Savant and her column in *Parade* magazine (September 9, 1990, p. 16) stating what has been called the Monte Hall problem:

> Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? (p. 16)

The answer almost universally given to this problem is that switching does not matter, presumably with the reasoning that there is no way for

the player to know which of the two unopened doors is the winner, and each of these must then have an equal probability of being the winner. By "doing the math," however, possibly writing down three doors hiding one car and two goats and working through the options in a short simulation, it becomes clear quickly that the opening of a goat door changes the information one has about the original situation, and that always changing doors doubles the probability of winning from 1/3 to 2/3. (As an interesting historical note, the "Monte Hall" problem has been a fixture of probability theory from at least the 1890s; it is called the problem of the "three caskets" by Henri Poincairé, and is more generally known as [Joseph] Bertrand's Box Paradox.)

Any beginning statistics class should always include a number of formal tools to help "do the math." Several of these have been mentioned in early sections: Bayes' theorem and implications for screening using sensitivities, specificities, and prior probabilities; conditional probabilities more generally and how probabilistic reasoning might work for facilitative and inhibitive events; sample sizes and variability in, say, a sample mean, and how a confidence interval might be constructed that could be made as accurate as necessary by just increasing the sample size, and without any need to consider the exact size (assumed to be large) of the original population of interest; how statistical independence operates or does not; the pervasiveness of natural variability and the use of simple probability models (as the binomial) to generate stochastic processes; the computations involved in corrections for attenuation; and usage of Taylor–Russell charts.

A second area of interest in developing statistical literacy and learning to reason ethically is the large body of work produced by psychologists regarding the normative theory of choice and decisions derivable from probability theory, and how it may not be the best guide to the actual reasoning processes that individuals engage in. The Nobel Prize–level contributions of Tversky and Kahneman (e.g., 1971, 1974, 1981) are particularly germane, and the view that people rely on various simplifying heuristic principles to assess probabilities and engage in judgments under uncertainty; also, that the psychology of choice is dictated to a great extent by the framing of a decision problem. We give two classic Tversky and Kahneman (1983, pp. 297, 299) examples to illustrate how reasoning heuristics and framing might operate:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.
>
> Which is more probable? (a) Linda is a bank teller. (b) Linda is a bank teller and is active in the feminist movement.

For one group of subjects, 85% choose option (b), even though the conjunction of two events must be less likely than either of the constituent events. Tversky and Kahneman (1983) argue that this "conjunction fallacy" occurs because the "representativeness heuristic" is being used to make the judgment—the second option seems more representative of Linda based on the description given for her.

The representativeness heuristic operates where probabilities are evaluated by the degree to which *A* is representative of *B*; if highly representative, the probability that *A* originates from *B* is assessed to be higher. When representativeness heuristics are in operation, a number of related characteristics of the attendant reasoning processes become apparent: prior probabilities (base rates) are ignored; insensitivity develops to the operation of sample size on variability; and an expectation that a sequence of events generated by some random process, even when the sequence is short, will still possess all the essential characteristics of the process itself. This leads to the "gambler's fallacy" (or, "the doctrine of the maturity of chances"), where certain events must be "due" to bring the string more in line with representativeness—as one should know, corrections are not made in a chance process but only diluted as the process unfolds. When a belief is present in the "law of small numbers," even small samples must be highly representative of the parent population—thus, researchers put too much faith in what is seen in small samples and overestimate replicability and fail to recognize regression toward the mean because predicted outcomes should be maximally representative of the input, and therefore, be exactly as extreme.

A second powerful reasoning heuristic is availability. We quote from Tversky and Kahneman (1974):

> Lifelong experience has taught us that, in general, instances of large classes are recalled better and faster than instances of less frequent classes; that likely occurrences are easier to imagine than unlikely ones; and that the associative connections between events are strengthened when the events frequently co-occur. As a result, man has at his disposal a procedure (the availability heuristic) for estimating the numerosity of a class, the likelihood of an event, or the frequency of co-occurrences, by the ease with which the relevant mental operations of retrieval, construction, or association can be performed. (p. 1128)

Because retrievability can be influenced by differential familiarity and saliences, the probability of an event may not be best estimated by the ease to which occurrences come to mind. A third reasoning heuristic is one of adjustment and anchoring, which may also be prone to various biasing effects. Here, estimates are made based on some initial value that is then adjusted.

The power of framing in how decision situations are assessed can be illustrated well through an example and the associated discussion provided by Tversky and Kahneman (1981):

> Problem 1 [*N* = 152]: Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved. [72 percent]
>
> If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. [28 percent]
>
> Which of the two programs would you favor?
>
> The majority choice in this problem is risk averse: The prospect of certainly saving 200 lives is more attractive than a risky prospect of equal expected value, that is, a one in three chance of saving 600 lives.
>
> A second group of respondents was given the cover story of problem 1 with a different formulation of the alternative programs, as follows:
>
> Problem 2 [*N* = 155]:
>
> If Program C is adopted, 400 people will die. [22 percent]
>
> If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78 percent]
>
> Which of the two programs would you favor?
>
> The majority choice in problem 2 is risk taking: The certain death of 400 people is less acceptable than the two in three chance that 600 will die. The preferences in problems 1 and 2 illustrate a common pattern: Choices involving gains are often risk averse, and choices involving losses are often risk taking. However, it is easy to see that the two problems are effectively identical. The only difference between them is that the outcomes are described in problem 1 by the number of lives saved and in problem 2 by the number of lives lost. The change is accompanied by a pronounced shift from risk aversion to risk taking.

The effects of framing can be very subtle when certain (coded) words are used to provide salient contexts that influence decision processes either consciously or unconsciously. A recent demonstration of this in the framework of our ongoing climate change debate is given by Hardisty, Johnson, and Weber (2010) in the journal *Psychological Science*. The article has an interesting title: *A Dirty Word or a Dirty World? Attribute Framing,*

*Political Affiliation, and Query Theory*; an abstract first posted online follows:

> Paying more for carbon-producing activities is one way to compensate for carbon dioxide emissions, but new research suggests that policymakers should be mindful of how they describe such initiatives. Volunteers were asked to choose between two identical products, one option including a surcharge for emitted carbon dioxide. When the surcharge was labeled as an "offset," the majority of volunteers chose the more expensive, environmentally friendly product. However, when the surcharge was labeled as a "tax," Republican and Independent volunteers were more likely to choose the less expensive option; Democratic volunteers' preferences did not change.

When required to reason about an individual's motives in some ethical context, it may be best to remember the operation of the *fundamental attribution error*, where people presume that actions of others are indicative of the true ilk of a person, and not just that the situation compels the behavior.

The presentation of data is an obvious area of concern when developing the basics of statistical literacy. Some aspects may be obvious, such as not making up data or suppressing analyses or information that does not conform to prior expectations. At times, however, it is possible to contextualize (or to "frame") the same information in different ways that might lead to differing interpretations. As noted in Gigerenzer et al. (2008), distinctions should be made between survival and mortality rates, absolute versus relative risks, and natural frequencies versus probabilities. Generally, the presentation of information should be as honest and clear as possible. An example given by Gigerenzer et al. suggests the use of frequency statements instead of single-event probabilities, which removes the ambiguity of the reference class being referred to: instead of saying "There is a 30–50% probability of developing sexual problems with Prozac," use "Out of every 10 patients who take Prozac, 3–5 experience a sexual problem."

In presenting data to persuade, and because of the so-called "lead time bias" that medical screening produces, it is unethical to promote any kind of screening based on improved 5-year survival rates, or to compare such survival rates across countries where screening practices vary. As a somewhat jaded view of our current health situation, we have physicians practicing defensive medicine because there are no legal consequences for overdiagnosis and overtreatment—only for underdiagnosis. Or, as the editor of *Lancet* commented (quoted in Gigerenzer et al., 2008): "Journals have devolved into information laundering operations for the pharmaceutical industry." The ethical issues involved in medical screening and its associated consequences are socially important; for example, months

after false positives for HIV, mammograms, prostate cancer, and the like, considerable and possibly dysfunctional anxiety may still exist.

A fourth statistical literacy concern is to have enough of the formal skills and context to separate legitimate claims from those that might represent more specious arguments. As examples, one should recognize when a case for cause is made in a situation where regression toward the mean is as likely an explanation, or when test unfairness is argued for based on differential performance (i.e., impact) and not on actual test bias (i.e., same ability levels performing differently). A more recent example of the questionable promotion of a methodological approach, called optimal data analysis (ODA), is given in Yarnold and Soltysik (2004). We quote from the preface:

> To determine whether ODA is the appropriate method of analysis for any particular data set, it is sufficient to consider the following question: When you make a prediction, would you rather be correct or incorrect? If your answer is "correct," then ODA is the appropriate analytic methodology—by definition. That is because, for any given data set, ODA explicitly obtains a statistical model that yields the theoretical maximum possible level of predictive accuracy (e.g., number of correct predictions) when it is applied to those data. That is the motivation for ODA; that is its purpose. Of course, it is a matter of personal preference whether one desires to make accurate predictions. In contrast, alternative non-ODA statistical models do not explicitly yield theoretical maximum predictive accuracy. Although they sometimes may, it is not guaranteed as it is for ODA models. It is for this reason that we refer to non-ODA models as being *suboptimal*.

Sophistic arguments such as these have no place in the legitimate methodological literature. It is not ethical to call one's method "optimal" and refer pejoratively to others as therefore "suboptimal." The simplistic approach to classification underlying "optimal data analysis" is known to not cross-validate well (see, e.g., Stam, 1997); it is a huge area of operations research where the engineering effort is always to squeeze a little more out of an observed sample. What is most relevant in the behavioral sciences is stability and cross-validation (of the type reviewed in Dawes, 1979, on proper and improper linear models) and to know what variables discriminate and how, and to thereby "tell the story" more convincingly and honestly.

The penultimate area of review in this concluding section is a reminder of the ubiquitous effects of searching/selecting/optimization and the identification of "false-positives." We have mentioned some blatant examples in earlier sections—the weird neuroscience correlations, the small probabilities (mis)reported in various legal cases (such as the Dreyfus small probability for the forgery coincidences or that for the de Berk hospital fatalities pattern), and repeated clinical experimentation until positive

results are reached in a drug trial—but there are many more situations that would fail to replicate; we need to be ever-vigilant of results obtained by "culling" and then presented to us as evidence.

A general version of the difficulties encountered when results are culled is labeled the *file drawer problem*. This refers to the practice of researchers putting away studies with negative outcomes; that is, those not reaching reasonable statistical significance or when something is found contrary to what the researchers want or expect, or those rejected by journals who will only consider publishing articles demonstrating positive and significant effects. The file drawer problem can seriously bias the results of a meta-analysis (i.e., methods for synthesizing collections of studies in a particular domain), particularly if only published sources are used (and not, for example, unpublished dissertations or all the rejected manuscripts lying on a pile in someone's office). We quote from the abstract of a fairly recent review, *The Scientific Status of Projective Techniques* (Lilienfeld, Wood, & Garb, 2000):

> Although some projective instruments were better than chance at detecting child sexual abuse, there were virtually no replicated findings across independent investigative teams. This meta-analysis also provides the first clear evidence of substantial file drawer effects in the projectives literature, as the effect sizes from published studies markedly exceeded those from unpublished studies.

The subtle effects of culling with subsequent failures to replicate can have serious consequences for the advancement of our understanding of human behavior. A recent important case in point involves a gene–environment interaction studied by a team led by Avshalom Caspi. A polymorphism related to the neurotransmitter serotonin was identified that apparently could be triggered to confer susceptibility to life stresses and resulting depression. Needless to say, this behavioral genetic link caused quite a stir in the community devoted to mental health research. Unfortunately, the result could not be replicated in a subsequent meta-analysis (could this possibly be due to the implicit culling over the numerous genes affecting the amount of serotonin in the brain?). Because of the importance of this cautionary tale for all behavioral genetics research, we refer the reader to a *News of the Week* item from *Science*, written by Constance Holden (June 26, 2009): *Back to the Drawing Board for Psychiatric Genetics*.

Our final concluding statistical literacy issue is the importance of developing abilities to spot and avoid falling prey to the trap of specious reasoning known as an "argument from ignorance," or *argumentum ad ignorantiam*, where a premise is claimed to be true only because it has not been proven false, or that it is false because it has not been proven true. Sometimes this is also referred to as "arguing from a vacuum" (paraphrasing from Dawes,

1994)—what is purported to be true is supported not by direct evidence but by attacking an alternative possibility. Thus, a clinician might say: "Because the research results indicate a great deal of uncertainty about what to do, my expert judgment can do better in prescribing treatment than these results." Or to argue that people "need" drugs just because they have not solved their problems before taking them.

A related fallacy is "argument from personal incredulity," where because one personally finds a premise unlikely or unbelievable, the premise can be assumed false, or that another preferred but unproven premise is true instead. In both of these instances, a person regards the lack of evidence for one view as constituting proof that another is true. Related fallacies are (a) the *false dilemma* where only two alternatives are considered when there are, in fact, other options. The famous Eldridge Cleaver quote from his 1968 Presidential campaign is a case in point: "You're either part of the solution or part of the problem." Or, (b) the Latin phrase *falsum in uno, falsum in omnibus* (false in one thing, false in everything) implying that someone found to be wrong on one issue must be wrong on all others as well. In a more homey form, "When a clock strikes 13, it raises doubt not only to that chime, but to the 12 that came before." Unfortunately, we may have a current example of this in the ongoing climate change debate; the one false statistic proffered by a report from the Intergovernmental Panel on Climate Change on Himalayan glacier melt may serve to derail the whole science-based argument that climate change is real.

Fallacies with a strong statistical tinge related to *argumentum ad ignorantiam* would be the "margin of error folly," usually attributed to David Rogosa (the name, not the folly itself): If it could be, it is. Or, in a hypothesis testing context, if a difference is not significant, it is zero. We now can refer to all these reasoning anomalies under the umbrella term "truthiness," coined by Stephen Colbert from Comedy Central's *The Colbert Report*. Here, truth comes from the gut, not books, and refers to the preferring of concepts or facts one wishes to be true rather than concepts or facts known to be true. Thus, in 2010 we have the "birthers," who claim that President Obama was not born in the United States, so constitutionally he cannot be President; or that the Health Care Bill includes "death squads" ready to "pull the plug on granny"; or that there were weapons of mass destruction that justified the Iraq war; and on and on.

## References

Agency for Healthcare Research and Quality. (2008). *National healthcare disparities report*. Rockville, MD: Author.

Agresti, A. (2007). *An introduction to categorial data analysis* (2nd ed.). New York: Wiley-Interscience.

Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*. Chichester, UK: Wiley.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

Aronowitz R. (2009, November 20). Addicted to mammograms. *The New York Times*.

Associated Press. (2004, April 26) Recap of St. Louis Cardinals vs. Pittsburgh Pirates.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289–300.

Bertin, J. (1973). *Semiologie graphique* (2nd ed.). The Hague: Mouton-Gautier. (English translation by W. Berg & H. Wainer published as *Semiology of graphics*, Madison, WI: University of Wisconsin Press, 1983.)

Best, J. (2005). Lies, calculations and constructions: Beyond "How to Lie with Statistics." *Statistical Science, 20,* 210–214.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187,* 398–404.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67,* 364–366.

Buchanan, M. (2007, May 16). The prosecutor's fallacy. *The New York Times*.

Campbell, D. T., & Kenny, D. A. (2002). *A primer on regression artifacts*. New York: Guilford Press.

Campbell, S. K. (1974). *Flaws and fallacies in statistical thinking*. Englewood Cliffs, NJ: Prentice-Hall.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372.

Champod, C., Taroni, F., & Margot, P.-A. (1999). The Dreyfus case—an early debate on expert's conclusions. *International Journal of Forensic Document Examiners, 5,* 446–459.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

Committee on DNA Forensic Science, National Research Council. (1996). *The evaluation of forensic DNA evidence*. Washington, DC: National Academies Press.

Committee on DNA Technology in Forensic Science, National Research Council. (1992). *DNA technology in forensic science*. Washington, DC: National Academies Press.

Dawes, R. M. (1975). Graduate admissions criteria and future success. *Science, 187,* 721–723.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: The Free Press.

Dunn-Rankin, P., Knezek, G. A., Wallace, S. R., & Zhang, S. (2004). *Scaling methods* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37,* 36–48.

Freedman, D. A. (1983). A note on screening regression coefficients. *The American Statistician, 37,* 152–155.

Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute, 15*, 246–263.

Gawande, A. (1999, February 8). The cancer-cluster myth. *The New Yorker*, pp. 34–37.

Gelman, A., Shor, B., Bafumi, J., & Park, D. (2007). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science, 2,* 345–367.

Gelman, A., Park, D., Shor, B., Bafumi, J., & Cortina, J. (2010). *Rich state, blue state, rich state, poor state: Why Americans vote the way they do* (expanded ed.). Princeton, NJ: Princeton University Press.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G., & Brighton, H. (2009). *Homo heuristics*: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107–143.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8,* 53–96.

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? Attribute framing, political affiliation, and query-theory. *Psychological Science, 21,* 86–92.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Holden, C. (2009). Back to the drawing board for psychiatric genetics. *Science*, *324*, 1628.

Huff, D. (1954). *How to lie with statistics*. New York: Norton.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics Journal, 34,* 21–39.

Kolata, G. (2009a, March 19). Prostate test found to save few lives. *The New York Times*.

Kolata, G. (2009b, October 21). Cancer society, in shift, has concerns on screenings. *The New York Times*.

Kolata, G. (2009c, November 17). Panel urges mammograms at 50, not 40. *The New York Times*.

Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science, 20,* 223–230.

Lilienfeld, S. A., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Radelet, M. L. (1981). Racial characteristics and the imposition of the death penalty. *American Sociological Review, 46,* 918–927.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367.
Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.
Royal Statistical Society. (2001, October 23). *News release: Royal Statistical Society concerned by issues rasied in Sally Clark case*. London: Author.
Sack, K. (2009, November 20). Screening debate reveals culture clash in medicine. *The New York Times*.
Selvin, H. C. (1958). Durkheim's suicide and problems of empirical research. *American Journal of Sociology, 63,* 607–619.
Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13,* 238–241.
Singer, N. (2009, July 17). In push for cancer screening, limited benefits. *The New York Times*.
Stam, A. (1997). MP approaches to classification: Issues and trends. *Annals of Operations Research, 74,* 1–36.
Steinley, D. (2003). Local optima in *K*-means clustering: What you don't know may hurt you. *Psychological Methods, 8,* 294–304.
Stout, D. (2009, April 3). Obama's census choice unsettles Republicans. *The New York Times*.
Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 33,* 565–578.
Thaler, R. H. (2009, December 20). Gauging the odds (and the costs) in health screening. *The New York Times*.
Thorndike, E. L. (1939). On the fallacy of imputing correlations found for groups to the individuals or smaller groups composing them. *The American Journal of Psychology, 52,* 122–124.
Tufte, E. (2006). *The cognitive style of PowerPoint: Pitching out corrupts within* (2nd ed.). Cheshire, CT: Graphics Press.
Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.
Tufte, E. R. (1996). *Visual explanations*. Cheshire, CT: Graphics Press.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.
Tversky, A., & Kahnman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review, 90,* 293–315.
Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4,* 274–290.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no never-mind. *Psychological Bulletin, 83,* 213–217.

Wainer, H. (1984). How to display data badly. *The American Statistician, 38,* 137–147.
Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus Books (Reprinted in 2000, Hillsdale, NJ: Lawrence Erlbaum Associates).
Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.
Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate and control uncertainty through graphical display*. Princeton, NJ: Princeton University Press.
Whitaker, C. F. (1990, September 9). Formulation by Marilyn vos Savant of a question posed in a letter from Craig Whitaker: "Ask Marilyn" column. *Parade Magazine*, p. 16.
Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician, 63,* 179–184.
Yarnold, P. R., & Soltysik, R. C. (2004). *Optimal data analysis*. Washington, DC: American Psychological Association.
Yule, G. U. (1903). Notes on the theory of association of attributes of statistics. *Biometrika, 2,* 121–134.
Yule, G. U. (1911). *Introduction to the theory of statistics*. London: Griffin.
Yule, G. U., & Kendall, M. G. (1968). *An introduction to the theory of statistics* (14th ed., 5th impression). New York: Hafner Publishing Company.
Zola, E. (1898, January 13). J'Accuse [I accuse]. *L'Aurore*.

# Section III

# Ethics and Research Design Issues

# 5

# *Measurement Choices: Reliability, Validity, and Generalizability*

**Madeline M. Carrig**
*Duke University*

**Rick H. Hoyle**
*Duke University*

The choice of measurement instrument is a critical component of any research undertaking in the behavioral sciences and is a topic that has spawned theoretical development and debate virtually since the dawn of our field. Unlike the eminently observable subjects of many other fields of scientific inquiry—for example, the physical characteristics of rock cores in sedimentary stratigraphy or the velocity of blood flows in biomedical engineering—the subject of interest in behavioral research is often human thoughts, feelings, preferences, or cognitive abilities that are not readily apparent to the investigator, and which may even be out of the full awareness of the research participant. Over the years, many hundreds of tools, such as pencil-and-paper questionnaires, projective tests, neuropsychological batteries, and, more recently, electrophysiological and neuroimaging techniques, have been developed or tailored in an attempt to capture the essence of various behavioral phenomena. For the research (or indeed, applied) behavioral scientist, the question arises: When it is time to operationalize a behavioral construct of interest, how should I choose and implement an instrument in a way that is consistent with ethical practice?

Practitioners often look to their governing associations for guidance on matters of professional ethics, and fortunately, in its 2002 *Ethical Principles of Psychologists and Code of Conduct* (the ethics code), the American Psychological Association (APA) provides some beginning guidance in answer to this question. In the sections of the ethics code that are most relevant to the ethical selection and use of behavioral measurement instruments in research, the code states:

1. Psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques (Section 9.02.a, p. 13).
2. Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation (Section 9.02.b, p. 13).

Hence, we are reminded that it is ethical to select instruments that are *useful* and *properly applied*. We are particularly encouraged to administer measures whose *reliability* and *validity* have been established in the population of interest and to report supporting psychometric evidence. But what types of evidence are most germane? More fundamentally, how should the research behavioral scientist evaluate whether an instrument, as well as his or her application of that instrument, possesses the desired characteristics?

It is to this latter question that the present chapter is substantially devoted. Our overarching goal is to provide guidance to the research behavioral scientist on the ethical selection and implementation of behavioral measurement instruments. We begin with a discussion of reliability and validity—two properties of measurement instruments that promote usefulness and proper application—and provide an overview of the methods that are presently available to the research behavioral scientist for the assessment of these properties. With respect to the proper application of such instruments, we also address the importance of considering the level of measurement, especially when instruments are involved in quantitative data analysis. Next, we expand our discussion of the ethics of behavioral measurement in research to include a survey of current scientific and ethical reporting standards. We then provide a summary of recommendations for practice. Finally, we present a case example, with the aim of highlighting key features of ethical conduct.[1]

[1] Other issues pertinent to the issue of the ethics of behavioral measurement pertain more specifically to psychodiagnostic assessment as performed by the clinical or school psychologist, such as training and supervision issues, use of tests for diagnosis, and the security of test materials and results. Such issues are addressed by sections of the APA code not presented here and are also discussed in detail in Koocher and Keith-Spiegel's excellent 2008 text. See also Wright and Wright (2002) for an interesting discussion of the ethics of behavioral measurement that focuses on the participant as a research stakeholder.

## Reliability and Validity

As is reflected in the APA ethics code, it is generally agreed that the two most desirable properties of a behavioral measurement instrument are that instrument's reliability and validity. The original developer of a behavioral measurement instrument bears a responsibility for furnishing reliability and validity evidence that supports the use of the instrument for its stated purpose, and it is reasonable for the investigator to consider that evidence when making a selection among instruments. However, the investigator ultimately bears the responsibility of demonstrating the reliability and validity of the instrument in the particular setting in which it has been used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). Correspondingly, most recent conceptualizations of these desirable psychometric properties focus more strongly on the reliability and validity of a particular measure's implementation, rather than on assessment of the validity and reliability of the measure per se, as will be highlighted below.

### Reliability

Reliability may be defined as the consistency of measurement instrument scores across replications of the measurement procedure (Brennan, 2001). Fortunately, it is a property of measurement that lends itself directly to quantification and statistical evaluation. Perhaps less fortunately, a dizzying array of methods for quantifying reliability are available, most of which depend on adoption of particular statistical models of measurement and/or definitions of the set of replications across which reliability will be assessed. Most of these methods involve the computation of either a *standard error of measurement* or the estimation of a *reliability coefficient*. We provide a brief overview of the various approaches. Our discussion draws on the comprehensive chapter written by Haertel (2006), which itself draws on earlier works by Thorndike (1951), Stanley (1971), and Feldt and Brennan (1989).

#### *Classical Test Theory*

In classical test theory (CTT), the model $X = T + E$ is used to describe the relationship between an observed score $X$, a "true" (error-free) score $T$, and the total measurement error $E$, where $E$ may arise from any number of sources but is assumed to be uncorrelated with the true score $T$ (Lord & Novick, 1968). In CTT, the reliability coefficient may be defined as the proportion of the total variance in observed scores that can be attributed to

true-score variance, or equivalently, as the squared correlation between the observed and true scores. As such, the reliability coefficient will assume values between 0 and 1 inclusive, with larger values reflective of greater reliability.

*Estimation of the Reliability Coefficient*

Although the reliability coefficient of a particular measurement process is rarely—if ever—exactly known, it may be numerically estimated. Over the years, CTT has given rise to multiple methods for producing such estimates. These methods, reviewed by Haertel (2006) in detail, include (a) the *parallel forms* reliability estimate, which is the correlation of scores resulting from two interchangeable (parallel) forms of a single measurement instrument administered to a single sample of participants at two points in time; (b) the *test–retest* reliability estimate, which is the correlation of scores resulting from two identical forms of a single measurement instrument administered to a single sample of participants at two points in time; and (c) the *staggered equivalent split-half procedure* (Becker, 2000), which attempts to take advantage of parallel-forms reliability estimation under circumstances when only one form of the measurement instrument is available.

An especially large category of methods for estimating the reliability coefficient in CTT includes *internal consistency* estimates, which tend to be frequently used because they were developed for the assessment of reliability from a single administration of a measurement instrument. All forms of internal consistency estimation involve subdividing the items of a measurement instrument and then observing the consistency of scores across subdivisions. Types of internal consistency estimates include (a) estimates that rely on the subdivision of the instrument into two parts, such as the *Spearman–Brown*, *Flanagan* or *Guttman–Rulon split-half*, *Raju*, and *Angoff–Feldt* coefficients (with Feldt & Charter, 2003, providing some guidance on making the best selection between them); and (b) estimates that rely on the subdivision of the measurement instrument into more than two parts, including *coefficient alpha*, *Kuder–Richardson 20*, *Kuder–Richardson 21*, *standardized alpha*, and *Guttman's* $\lambda_2$. Although popular, internal consistency estimates are likely to overestimate a measurement instrument's reliability because they do not capture error associated with possible fluctuations over time in responses to the instrument. The reader is encouraged to consult Haertel (2006) for references and for technical and computational details. Haertel (2006) also addresses estimates of reliability that are appropriate for composite scores, including difference scores, and provides information on computation of the CTT *conditional standard error of measurement*, which provides the standard error of measurement for a particular true score and is therefore useful for computing true-score confidence intervals.

*Applications of Reliability Estimation in Statistical Analysis*

When we apply inferential statistical models, we are generally interested in investigating relationships among the true scores on the constructs we intended to measure. However, models are generally fit to observed scores, and because of the complexities of assessing intra- or extrapsychic human behavior, even the best-conceived behavioral measurement instrument is likely to fail to achieve perfect reliability. Unfortunately, use of observed scores that are not perfectly reliable in the context of inferential statistical models can produce seriously misleading results, with potentially dramatic repercussions on the development of theory, clinical practice, policy, and the direction of future research. Failure to account for the presence of measurement error in a covariate used within an analysis of covariance (ANCOVA) model, for example, can lead either to significant $F$ tests in the presence of no true adjusted effect or to nonsignificant $F$ tests in the presence of a true adjusted effect (Maxwell & Delaney, 2004). Cohen, Cohen, West, and Aiken (2003) point out the potential of instrument fallibility to distort partialled relationships (e.g., partial regression coefficients) and to increase Type I or Type II error rates in the more general multiple regression/correlation analysis framework. Likewise, via simulation results, Hoyle and Kenny (1999) have demonstrated that mediational analyses that fail to account for unreliability in the mediating variable can produce biased parameter estimates and increase Type I and Type II error rates for the associated statistical tests.

The real threat of unreliability to the correctness of statistical conclusions under many circumstances has led to the development of statistical frameworks within the CTT tradition that attempt to "correct" for observed scores' fallibility, providing measures of effect that more closely reflect the relationships among the true scores (constructs) under investigation. Huitema (1980), for example, addresses options for analysis that may correct the problem in the context of ANCOVA; Cohen et al. (2003) detail methods developed to correct for the attenuation of correlation coefficients associated with measurement error and provide an overview of the strengths and weaknesses of existing remedies for the distortion of partialled relationships. Furthermore, Hedges and Olkin (1985) address correction for unreliability-associated attenuation of effect size.

The methods just described apply after-the-fact adjustments to parameter estimates produced from fitting a statistical model to a set of fallible observed scores. In general, they rely on the assumption that a measurement instrument's reliability is known. However, the substitution of estimated reliabilities can lead to potentially problematic results (see, e.g., Dunivant, 1981). Fortunately, more sophisticated methods are available that account for measurement error during the estimation process itself. Structural equation modeling procedures (Jöreskög, 1970), for example,

allow for the specification of a measurement model in which an unobserved *latent variable*, which holds an individual's hypothetical true score, and a separate unobserved measurement error variable together predict the individual's observed score on each of a number of behavioral measures. Relationships among the unobserved latent variables—the true-score measures of the constructs of interest—may then be modeled as the investigator sees fit, with the resulting parameter estimates presumably being free from the deleterious effects of measurement error. *Instrumental variable* estimation may also be used to minimize, or perhaps even remove, the negative influence of measurement error on parameter estimates (e.g., Hägglund, 1982).

### *Generalizability Theory*

Because of the relative complexity of its associated models and data-analytic methods, *generalizability theory* (GT) will perhaps be less familiar to the research behavioral scientist than CTT. GT is largely (although not universally) viewed as an extension of CTT. Haertel (2006) provides a brief but readable introduction, and Brennan (2001) offers a more comprehensive treatment.

As noted above, the basic CTT measurement model includes one term ($E$) that captures the total of measurement error. Relative to CTT, GT offers many advantages in terms of the evaluation of a measurement process's reliability. Perhaps the two most important are (a) the inclusion in measurement models of terms that permit the specification of multiple and distinct types of error and (b) a more precise conceptualization of the set of replications across which reliability is to be evaluated.

Even a very basic application of GT to a reliability evaluation requires multiple definitions and decisions. For example, the investigator must identify the potential sources of error variance in the observed scores. These might include, for example, rater, test form, location of administration, and occasion of measurement. In GT, each source (e.g., rater) is named a *facet*, and each level within that source (e.g., Jane, Bill) is considered a *condition* of that facet. The investigator must also specify a so-called *universe of generalization*, defining the exact set of potential replications across which reliability will be defined for a particular measurement process. Accordingly, a single "measurement" within the universe of generalization might include a set of multiple observations (i.e., a collection of observed scores), each associated with a particular condition for each facet. Importantly, the investigator must also decide whether each of his or her facets is *random* or *fixed*. Random facets are those for which the particular conditions observed by the investigator in one measurement are viewed as a random sample from an infinitely large population of conditions to which the investigator seeks to generalize. Fixed facets, on the other hand,

are those involving a set of conditions that will not vary across the set of hypothetical measurements within the universe of generalization.

In a GT *decision study (D-study)*, a basic linear measurement model might explain an observed score as a function of a person (participant) effect, multiple facet effects, and perhaps effects that represent interactions among effects (together with a residual). Random-effects analysis of variance (ANOVA) is used to estimate the *variance components* (variances) of the various effects included in the measurement model. This procedure permits the estimation of a *universe score variance*, which captures the variability of the person effect across the hypothetical measurements in the universe of generalization. Estimated variance components can be used to compute *coefficients of generalizability*, which assess the reliability of a particular measurement instrument within the defined universe of generalization. Under some circumstances, certain generalizability coefficients (e.g., the $E\rho^2$ of Cronbach, Gleser, Nanda, & Rajaratnam, 1972) simplify to forms of the reliability coefficient defined in CTT. Extensions of GT for more complicated measurement models and data structures are available (cf. Brennan, 2001). Haertel (2006) provides a brief overview of the estimation of conditional standard errors of measurement from the perspective of GT.

### *Item Response Theory*

The *item response theory* (IRT) model (e.g., Lord, 1968; Lord & Novick, 1968)—sometimes also named the latent trait model (Lord, 1953), logistic test model (Birnbaum, 1968), or Rasch model (Rasch, 1960)—is a family of models that uses a function of a set of participant and item parameters to describe the probability that a participant will receive a particular score on an individual measurement instrument item. Specific models within the IRT family may be differentiated in terms of multiple characteristics, including (a) the type of score produced by the measurement instrument items (i.e., binary vs. a *polytomous*, or ordered-categorical, outcome); (b) the model's *dimensionality*, or in other words, the number of participant parameters (also known as *abilities*, *traits*, or *proficiencies*) included in the model; (c) the number and type of item parameters involved in the model (which may include, e.g., characteristics such as item difficulty or capacity to discriminate among participants of differing abilities); and (d) the particular mathematical function used to relate the participant and item parameters to the observed score (Yen & Fitzpatrick, 2006).

The IRT model may be distinguished from the CTT and GT models in multiple ways, including (a) the IRT model's greater focus on item versus test-level scores; (b) the IRT model's somewhat more restrictive definition of a replication, with all item parameters in the IRT framework typically viewed as being fixed across all possible replications; (c) differences across models in the exact meaning of "true score"; and (d) the lack of an error

term in IRT (cf. Brennan, 2006). Hambleton and Jones (1993) note that the assumptions made by the IRT model are relatively more difficult to satisfy than those of CTT but emphasize that if the model fits the observed data well, IRT offers the advantage of participant and item parameters that are sample independent. Brennan summarizes his view of the differences between IRT, CTT, and GT thusly: "IRT is essentially a scaling model, whereas classical test theory and generalizability theory are measurement models. The essential difference, as I see it, is that a measurement model has a built-in, explicit consideration of error" (p. 6).

In great part because of its model's lack of an error term, IRT does not provide the more traditional reliability coefficients offered by CTT and GT. However, the IRT test information function does yield its own version of the conditional standard error of measurement, with technical and computational details addressed by Yen and Fitzpatrick (2006). Although the IRT conditional standard error of measurement is often used in the same manner as its CTT and GT counterparts, the investigator should be aware that there exist subtle differences in their meanings and appropriate interpretations (cf. Brennan, 2006).

### *Recommendations*

In sum, reliability may be defined as the consistency of measurement instrument scores across replications of that measurement procedure. A number of statistics for estimating an instrument's (unknown) true reliability are available, including the many reliability coefficients offered by CTT, GT's generalizability coefficients, and the conditional standard errors of measurement yielded by CTT, GT, and IRT. For any particular implementation of a measurement instrument, each of these statistics will be associated with particular strengths and weaknesses associated with (a) the fit of the observed data to the proposed measurement model; (b) practical considerations, such as the sample size required to produce stable estimates (with the IRT statistics requiring somewhat larger samples); and (c) the relevance of the statistic to the applied setting in which it has been used.

With regard to the last consideration, several points are worthy of mention. First, it should be remembered that there exist sometimes-subtle differences between the various reliability and generalizability coefficients developed within the CTT and GT frameworks. In that connection, some coefficients will not be suitable for some intended purposes and populations. A test–retest reliability coefficient, for example, would not be the ideal estimate of the precision of a measure of mood, a construct that will itself vary over time, resulting in changes in observed scores that are unrelated to measurement error. Second, statistics developed within the CTT and GT frameworks are generally sample (e.g., population, D-study design, item) dependent. Third, the standard errors of measurement

yielded by the IRT test information function reflect the restricted configuration of measurement error that is addressed by internal consistency estimates of reliability, and should be interpreted accordingly (AERA, APA, & NCME, 1999). Finally, the investigator should be aware that the particular mathematical function used within the IRT framework to relate item parameters to the observed score can influence the estimated standard errors of measurement (AERA, APA, & NCME).

Does an assessment of the available methods' overall strengths and weaknesses allow for more specific recommendations for practice? The 1999 volume *Standards for Educational and Psychological Testing* (the *Standards*), which was jointly published by the AERA, APA, and NCME, provides some guidance on the most appropriate coefficient for a small set of specific testing purposes (e.g., it recommends that when a measurement instrument is designed to reflect rate of work, a test–retest or alternate-forms coefficient should be used), and its authors emphasize the increasing importance of precision as the potential consequences of measurement error grow in importance (e.g., as in a setting where a single score is used to make decisions about admission to graduate school). In general, however, the *Standards* provides no "cookbook" recommendations regarding the type of reliability evidence that should be sought, nor the level of precision that should be attained. We agree with the authors' assessment that:

> There is no single, preferred approach to quantification of reliability. No single index adequately conveys all of the relevant facts. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment. (AERA, APA, & NCME, 1999, p. 31)

Of course, for such judgment to be apt, the investigator must be conversant with the various approaches. Moreover, it is hoped that the investigator will possess sufficient technical resources such that the choice of reliability evidence will be made solely based on the methods' relative strengths and weaknesses and not based on his or her ability (or lack thereof) to enact the different techniques. We hope that our necessarily brief overview of the available methods will spur the reader to pursue any needed additional education on their derivation, computation, and interpretation.

## Validity

Brennan (2006) provides an excellent and informative overview of the evolution of measurement theory, and in particular, of historical developments in theoretical models of validity (see also Thompson & Daniel, 1996).

Brennan notes that earlier conceptualizations of validity involved multipartite models that focused on defining specific aspects of validity (e.g., the content, predictive, concurrent, and construct validities defined in the APA's 1954 *Technical Recommendations for Psychological Tests and Diagnostic Techniques*), but emphasizes that more recently theoretical developments have focused on more unified conceptualizations of validity (e.g., Messick, 1988b, 1989) that lend themselves to consideration of multiple means of accumulating evidence relevant to instrument validation.

In an influential 1989 work, Messick defines validity as follows:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.… [It] is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use. Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (Messick, 1989, p. 13)

Hence, in ascribing validity to a behavioral measurement process, Messick (1989) focuses on the importance of both (a) establishing the soundness of inferences drawn from the use of the measurement instrument and (b) considering the potential consequences of those inferences. In a separate work, Messick (1988a) offers a four-faceted question that he suggests as a guide for those interested in the process of evaluating validity in the context of behavioral measurement. He asks the potential user of the instrument to consider:

> What balance of evidence supports the interpretation or meaning of the scores; what evidence undergirds not only score meaning, but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting; what rationales make credible the value implications of the score interpretation and any associated implications for action; and what evidence and arguments signify the functional worth of the testing in terms of its intended and unintended consequences. (Messick, 1988a, p. 5)

The latter two facets of Messick's question in particular have engendered discussions of the value of assessing the social (vs. scientific) consequences of measurement (see, e.g., Lees-Haley, 1996; Messick, 1995); they are subjects rife with their own ethical complexities but which fall largely outside the scope of the present chapter. The first two facets of Messick's question, however, are very pertinent to the subject of ethical measurement choices in the research context, and will be addressed below in turn.

First: What balance of evidence supports the interpretation or meaning of the scores?

### *Score Interpretation*

In some applied settings—in recognition of the imperfect nature of behavioral measurement—scores resulting from measurement instruments are not viewed as being perfectly related to the trait or behavior being "measured," but rather, are used to generate hypotheses that are left open to rejection on further investigation. For example, the wise and ethical practicing clinical psychologist would not rely solely on the particular scores resulting from a Rorschach inkblot test to make a definitive diagnosis of psychotic disorder. Instead, the psychologist would consider such scores in the context of a wealth of additional information, such as unstructured interview data, behavioral observations, and records review. When behavioral measurement tools are used in this fashion—as a hypothesis-generating mechanism in the context of an in-depth, individualized assessment paradigm—the potentially negative consequences of imperfect measurement can be dramatically minimized.

In other applied settings, however, scores resulting from behavioral assessment tools are indeed expected to provide a relatively direct index of the trait or behavior being measured, and scores are used in a fashion that is consistent with that expectation. Many tests administered in educational settings, for example, generate scores that are presumed to provide a highly valid index of intellectual and educational aptitude and/or achievement (e.g., end-of-grade testing; the Scholastic Aptitude Test [SAT]). Such scores are used sometimes alone, or in concert with limited additional information, to make consequential decisions about student services, placement, and progression. Readers are encouraged to refer to the chapter written by Cizek and Rosenberg (Chapter 8, this volume) for a discussion of the ethical considerations relevant to these so-called "high-stakes" assessment situations.

In the research setting, and especially in studies that use quantitative methods, assessment tools are used to measure the behavioral constructs under investigation, and the particular scores resulting from these tools are construed as reflecting the level or degree of the trait or behavior being measured. Scores resulting from behavioral measurement tools administered to research participants are generally not of "high stakes" to the participating individuals, in the sense that scores are most often not shared either with the participant or with other decision makers in the participant's life, limiting their potential consequence to the individual. Nevertheless, when investigators apply inferential statistical methods to a full sample of such scores, the particular scores observed will obviously have a critical impact on conclusions drawn about the phenomena

of interest in the population sampled. Hence it is important to establish that the interpretation or assigned meaning of the scores is reflective of the level of the construct being measured.

Kane (2006) provides a highly useful framework for critically evaluating the argument that a particular measurement instrument produces scores that are appropriate for construct-relevant interpretations. Kane asserts that the first step in ensuring interpretable and meaningful scores is to obtain evidence relevant to the scoring of the instrument.

*Scoring*

Kane (2006) recommends that the user of a behavioral measurement instrument ensure that (a) the rubric used for scoring is appropriate, (b) the rules for scoring are implemented as specified during test construction, and (c) the scoring is unbiased. He notes that many forms of scoring-related evidence could serve to undermine the proposed score interpretations, including, for example, poor interrater agreement, evidence of inadequate training of scorers or raters, and the failure of scoring rules to include relevant criteria. Kane also notes that if a statistical model is used in scaling, it is important to empirically verify that the selected model is a good fit to the observed data.

In making this last point, Kane (2006) focuses primarily on the scaling of scores in the context of standardized testing programs (like the SAT). Under many circumstances, however, new (true or latent) scores are generated by the data analyst when an inferential statistical model is fit to the raw (or scaled) scores resulting from a measurement instrument. The appropriateness and fit of such models are key factors in assessing the validity of the resulting scores. We will return to a discussion of the ramifications of noncontinuous scoring, in particular, on the estimation of measurement models in a later section.

*Generalization*

The second step Kane (2006) recommends in developing an argument for the interpretability and meaningfulness of scores is provided in the language of GT. In particular, Kane advises that the investigator establish that the observed score to universe score generalization is appropriate for the present use of the instrument. Kane suggests that the investigator first evaluate whether the investigator's sample of observations is representative of the currently defined universe of generalization. Paraphrasing an argument made in an earlier article (Kane, 1996), he opines, "If a serious effort has been made to draw a representative sample from the universe of generalization, and there is no indication that this effort has failed, it would be reasonable to assume that the sample is representative" (p. 35). Concomitantly, Kane suggests that the investigator assess whether the sample size of the present measurement procedure is large

enough to compensate for sampling error. He notes that examination of D-study evidence can point to the presence of problematically large random sampling errors for one or more facets.

*Extrapolation*

Kane's (2006) extrapolation step involves assurance that the universe score established in the previous step is meaningfully related to the target construct. Such assurance can be obtained using multiple analytic and empirical results, including evaluation of (a) the extent to which the measurement instrument contains items or tasks that are as representative as possible of the construct being assessed (Kane notes that standardization can minimize error, but with the tradeoff that standardized instruments may be associated with a universe of generalization that does not always adequately sample the target domain); (b) *face validity*, or the extent to which the relevance of the measurement instrument to the proposed construct interpretation is apparent to the research participant; (c) *criterion validity*, or the extent to which observed scores on the measurement instrument correlate with scores on a clearly valid criterion measure; and (d) *convergent validity*, or the extent to which observed scores on the measurement instrument correlate with scores on other (perhaps established) measures that seek to tap the same (or a similar) construct.

*Implication*

The final step in Kane's (2006) framework for appraising whether scores are appropriate for construct-relevant interpretations involves evaluating whether the construct score on the measurement instrument is appropriately linked to the verbal description of that score, and to any implications created by that label. For example, evidence that an achievement-related construct score varies across racial/ethnic groups consisting of members who are otherwise very similar with regard to intellectual ability and educational background would raise serious doubts about the associated measurement procedure's validity.

*Threats to Validity*

Kane (2006) also urges the investigator to rule out two major threats to the appropriateness of scores to construct-relevant interpretations. The first threat is identified as *trait underrepresentation*, which occurs when a measurement process undersamples the processes and contexts germane to the construct of interest, possibly leading to an overly restrictive universe of generalization. In that connection, Cook and Campbell (1979) and Messick (1989), among others, have emphasized the value of using multiple methods of assessment. The second threat to validity considered by Kane is *irrelevant variance* (vs. random error; also known as *systematic error*), which

is present in the scores derived from a measurement instrument to the extent that those scores are systematically affected by processes that are unrelated to the construct of interest (e.g., rater bias). Multimodal assessment may also minimize irrelevant variance (Messick, 1989).

### *Score Relevance*

What evidence undergirds not only score meaning but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting? The second facet of Messick's (1988a) question may be viewed in part as addressing the question of *generalizability*. In particular, the investigator might ask of the measurement instrument under consideration: Was the instrument originally developed—and validated—for the population and applied purpose that will be the focus of the proposed research? If the answer is no, then the ethical investigator must shoulder the responsibility of seeking out evidence that the measure does operate as intended in the population, and for the purpose, of interest. If such evidence is not available, then he or she should be prepared, as is suggested in the 2002 APA ethics code, to emphasize in the research report the potential limitations of the measurement process and the associated inferences and interpretations.

The procedures outlined in the previous section may be used to validate a measurement instrument in a new population and/or for a novel applied purpose. Those who wish to provide statistical evidence of generalizability may also take advantage of methods for establishing *measurement invariance*. From this perspective, a measurement instrument is considered to be invariant, or generalizable, across populations if participants from different populations who possess the same level of the construct of interest have the same probability of attaining a given score on the instrument (Mellenbergh, 1989). In that connection, latent variable models within both the confirmatory factor analysis and item response theory traditions allow the investigator to evaluate whether fixing parameters relating observed scores to latent variables to be equal across populations results in a significant decrement in model fit (cf. Meade & Lautenschlager, 2004).

### *Recommendations*

In sum, the evaluation of the validity of a measurement process will involve the collection of evidence regarding the measurement instrument's proposed interpretation and use in the population and setting under investigation. Such evidence will likely include information relevant to scoring, generalization, extrapolation, and implication inferences, and will involve assessment of the extent to which the measurement instrument is invariant across populations of interest. It is important to

note that a measurement application that cannot be demonstrated to be adequately reliable will be unlikely to yield sufficient validity evidence (AERA, APA, & NCME, 1999).

## Noncontinuous Scores

Many of the methods described above, and especially those drawn from the CTT and GT traditions, are appropriate for the evaluation of measurement instruments that produce continuous scores. However, many forms of behavioral measurement involve classifications or scale types that may fail to produce continuous scores. The investigator who is considering the use of such instruments will encounter at least two decision points in his or her work.

The first question regards whether the noncontinuous nature of the resulting scores calls for different strategies for the evaluation of the instrument's reliability. The answer to this question is likely yes, and that new strategies should especially be considered when scores capture discrete group-membership information. Haertel (2006) provides a useful overview of specialized indices of reliability that are appropriate for either (a) continuous scores that are used to make categorical decisions (as when, e.g., a test is scored and then assigned a value of "pass" or "fail"), or (b) measurement procedures that directly generate classifications into a set of discrete categories. Methods appropriate for classifications involving the comparison of a continuous score with a cut score (or set of cut scores) include Livingston's $k^2$, Brennan and Kane's $\Phi$ and $\Phi(\lambda)$, and Cohen's $\kappa$, among many others. Blackman and Koval (1993) provide a discussion of multiple reliability indices that consider the extent of consistency across raters when a measurement procedure involves direct classification into categories.

The second issue confronted by the investigator who contemplates a measurement procedure with noncontinuous outcomes involves consideration of whether inferential measurement models that traditionally use continuous scores may be ethically applied to noncontinuous scores. Most behavioral scientists would acknowledge that few behavioral measurement instruments are possessive of an interval or ratio scale of measurement, and yet application of measurement models and estimation procedures that rely on continuous, interval-level measurement is very common, even when ordinal-level (ordered-categorical) scores have been observed. The appropriateness of such practice has been debated vigorously over the years, both on philosophical and applied grounds (see, e.g., Marcus-Roberts & Roberts, 1987; Michell, 1986; Townsend & Ashby, 1984). For the researcher interested in applying a particular measurement model

to ordinal-level data, both theoretical work and simulation studies, which investigate the behavior of statistical procedures when certain assumptions are violated, can be very informative.

For example, application of the common linear confirmatory factor analysis (CFA) model assumes that observed scores are continuous, or at least interval-level, in measurement. Multiple Monte Carlo simulation studies have addressed the impact within traditional CFA of ordinal-level measurement on the estimation of parameters linking latent variables to observed scores. Wirth and Edwards (2007) note that although results from some studies (e.g., DiStefano, 2002; Dolan, 1994) are suggestive that traditional maximum-likelihood estimation with adjustment might produce acceptable results when the number of ordered categories is five or greater, other findings (e.g., Cai, Maydeu-Olivares, Coffman, & Thissen, 2006) indicate that caution is warranted when applying traditional models to categorical data, even when single-moment adjustments are made.

Fortunately, as Wirth and Edwards (2007) report, multiple statistical frameworks have been developed to accommodate categorical data, among them item factor analysis (IFA). The authors acknowledge the lure of the application of traditional methods to ordinal-level data, especially in light of IFA models' greater complexity and the larger sample sizes typically required to produce stable results. Nonetheless, they recommend that investigators favor IFA when either (a) the number of response categories is fewer than five and/or (b) the present measurement procedure has not yet been well validated in the population of interest. Although they do not uniformly object to the use of traditional methods when measurement procedures produce more than five response categories, in such situations, they strongly encourage the researcher to verify the consistency of results of traditional techniques—obtained using a variety of estimation methods—with those obtained using IFA.

Thus, application of measurement models and estimation procedures that rely on continuous scores to ordinal-level data can potentially produce misleading results. Ideally, investigators will choose measurement models that are appropriate for the level of measurement attained by a particular measurement procedure. If the procedure's level of measurement does not fully satisfy the requirements of a statistical method, it is incumbent on the investigator to critically evaluate the relevant analytical and simulation study findings before applying the method. If the investigator proceeds, she or he should clearly delineate the potential limitations of the method, as applied to data with the observed characteristics, in the research report. Note that although the present discussion has been focused on measurement models, these points are equally relevant to the application of explanatory statistical models such as ANOVA and linear regression to ordinally scaled dependent variables.

We turn next to a review of current standards for the reporting of measurement strategies and associated analyses.

## Scientific and Ethical Reporting Standards

Practitioners increasingly are urged—even required—to use evidence-based decision making when choosing interventions and treatments. Such decision making requires access to relevant research that is reported in a manner that allows for an evaluation of its strengths and limitations. To ensure that research reports routinely include all the information necessary for weighing the evidence produced, a number of professional organizations have articulated reporting standards. These standards cover all facets of research—from conceptualization to design and analysis—and to varying degrees they address the integrity of the measures and measurement strategies used.

## Origins in Biomedical Research

Perhaps the most organized efforts at standardizing research reports and ensuring that they include all the information bearing on the strength and limitations of a study have been within the context of biomedical research. At least two statements and accompanying checklists are primarily aimed at guiding the reporting of biomedical research findings: the Consolidated Standards of Reporting Trials (CONSORT) statement, which applies to randomized trials, and the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) statement, which applies to nonrandomized evaluation studies. The standards prescribed by these statements largely focus on accounting for research participants and on description of the intervention or treatment, procedure for assigning participants to condition, and methods involved in statistical inference; however, both touch on measurement. Item 6a in the CONSORT statement, for example, prescribes "clearly defined primary and secondary outcome measures" (Altman et al., 2001, p. 669). The explanation associated with this item provides several guidelines for reporting. First, primary outcomes should be clearly identified and distinguished from secondary outcomes. Second, when scales or instruments are used, "authors should indicate [their] provenance and properties" (p. 669). Finally, the statement urges the use of "previously developed and validated scales" (p. 669; see also Marshall

et al., 2000). Item 6b in the CONSORT statement refers to steps taken to improve the reliability of measurement and indicates that the use of multiple measurements or assessor training should be described fully in the report. Although the CONSORT guidelines have been endorsed by more than 150 journals, general adherence by authors and enforcement by journal editors have not been uniform (Barbour, Moher, Sox, & Kahn, 2005).

The TREND statement was patterned after the CONSORT statement but adds items relevant to research in which participants are not randomized to condition (Des Jarlais, Lyles, & Crepaz, 2004). To the CONSORT measurement recommendations, the TREND statement adds only the prescription that "methods used to collect data" (e.g., self-report, interview, computer-assisted) should be described.

Before turning to standards for reporting behavioral research, we note one additional set of standards that is focused almost entirely on measurement: the Standards for the Reporting of Diagnostic Accuracy Studies (STARD; Bossuyt et al., 2003). The STARD checklist includes 25 items, of which approximately half concern the reporting of measurement methods and the analysis of the effectiveness of the "index test" to be used for diagnoses. Prescriptions to report information about the reference standard, reproducibility, and accuracy of classification—for the sample as a whole and for subgroups of interest—reflect the importance of a clearly articulated evidence base for tests that will be used by practitioners. Use of the STARD checklist is encouraged by more than 200 biomedical journals.

## Standards for Behavioral Science

In behavioral science, only recently have formal statements of reporting standards been published, and to date, these have not been formally endorsed by specific journals. As a result, the primary audience for such standards is manuscript authors. Adherence to the standards is voluntary and uneven; hence they currently appear to function more as recommendations than standards as strictly defined.

In 1999, the APA Task Force on Statistical Inference published a set of guidelines for the reporting of statistical methods (Wilkinson & the Task Force on Statistical Inference, 1999). Building on an earlier report by the International Committee of Medical Journal Editors (Bailar & Mosteller, 1988), the APA Task Force offered guidelines for reporting the investigator's selected methods and associated results and for drawing appropriate conclusions. Unlike the earlier report, the report of the Task Force included a lengthy section on measurement in which the authors offer guidance on describing measurement procedures. With regard to variables, the Task

Force asserts, "Naming a variable is almost as important as measuring it. We do well to select a name that reflects how a variable is measured" (p. 596). With regard to the use of a questionnaire measure, the Task Force urges authors to "summarize the psychometric properties of its scores with specific regard to the way the instrument is used in the population" (p. 596). *Psychometric properties* were defined as "measures of validity, reliability, and any other qualities affecting conclusions" (p. 596). Finally, the Task Force proposes that authors provide detail about how the measures were used, recommending that authors "clearly describe the conditions under which measurements are taken (e.g., format, time, place, personnel who collected the data)" (p. 596). Of particular concern to the Task Force were aspects of the measurement procedure that might introduce bias, and they instruct authors to describe measures taken to reduce or eliminate potential biases.

A similar report was produced by the American Education Research Association's (AERA, 2006) Task Force on Reporting of Research Methods in AERA Publications. To the information requested by the APA Task Force, the AERA Task Force adds with reference to the description of measures used in research that "information on access to these surveys, instruments, protocols, inventories, and guides should be specified" (p. 36). In addition, prescriptions for detailing steps taken to develop new measures or to classify research participants using scores are offered in a "Measurement and Classification" section.

The move from guidelines and suggestions to potential standards for reporting of behavioral science is more apparent with the efforts of the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (the Working Group, 2008). The Working Group began by consolidating the CONSORT, TREND, and AERA standards described earlier and then added new reporting recommendations. Like the CONSORT and TREND standards, the resultant recommendations refer to all aspects of a report of empirical research. In a section labeled "Measures and Covariates," the Working Group recommends that reports of new research include (a) definitions of all variables, including primary and secondary variables and covariates (to include mention of variables on which data were gathered but not analyzed for the report); (b) measurement methods, including reference to any training of individuals who administered measures and consistency between measures when administered more than once; and (c) "information on validated or ad hoc instruments created for individual studies, for example, psychometric and biometric properties" (p. 842). Unfortunately, perhaps because of the intended generality of the proposed standards, neither details nor examples are provided for the Working Group's recommendations.

The APA Working Group was one of seven working groups that contributed to the production of the sixth edition of the *Publication Manual of the*

*American Psychological Association* (APA, 2010), and its recommendations for research reports are reflected in that influential document. In fact, the content of the "Measures and Covariates" item from the Working Group report was carried forward into the *Manual* without elaboration. Notably, in a chapter on manuscript structure and content, the *Manual*'s recommendations are referred to as reporting "standards," although authors are encouraged to "balance the rules of the *Publication Manual* with good judgment" (p. 5).

To date, perhaps the most concrete, prescriptive, and exhaustive set of recommendations addressing the development, evaluation, and appropriate documentation of behavioral measurement instruments is provided by the 1999 *Standards* (AERA, APA, & NCME). Although the primary aim of the *Standards* was to provide guidance to those involved with educational, personnel, and program evaluation testing applications (see Cizek & Rosenberg, Chapter 8, this volume), the recommendations are sufficiently general that they are relevant for the broader applied measurement and behavioral science research community. In summarizing the overall purpose and intended audience of the *Standards*, the authors advocate that "within feasible limits, the relevant technical information be made available so that those involved in policy debate may be fully informed" (p. 2).

With regard to reporting associated with the precision of a measurement procedure, the *Standards* emphasizes that "general statements to the effect that a test is 'reliable' or that it is 'sufficiently reliable to permit interpretations of individual scores' are rarely, if ever, acceptable" (AERA, APA, & NCME, 1999, p. 31). For the selection, evaluation, and reporting of data germane to the assessment of a measurement instrument's reliability, the authors provide the following guidance:

- For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported (Standard 2.1, p. 31).
- The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original sale units and in units of each derived score recommended for use in test interpretation (Standard 2.2, p. 31).
- When test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences (Standard 2.3, p. 32).
- Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to

select examinees for reliability analyses and descriptive statistics on these samples should be reported (Standard 2.4, p. 32).

- A reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent (Standard 2.5, p. 32).
- Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score (Standard 2.14, p. 35).

Additional reliability standards address reporting for particular testing applications (e.g., tests designed to reflect rate of work, tests scored by raters, tests with both long and short forms). The reader is encouraged to consult the *Standards* for annotations and additional details.

The *Standards* (AERA, APA, & NCME, 1999) additionally addresses the reporting of validity evidence. Recommendations particularly relevant to the use of an existing measure in a research context include:

- If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary (Standard 1.4, p. 18).
- The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics (Standard 1.5, p. 18).
- When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided (from Standard 1.10, p. 19).
- If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided (Standard 1.11, p. 20).
- When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given (Standard 1.12, p. 20).
- When validity evidence includes empirical analyses of test responses together with data on other variables, the rationale

for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent (Standard 1.14, p. 20).

- When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported (Standard 1.16, p. 21).
- If test scores are used in conjunction with other quantifiable variables to predict some outcome or criterion, regression (or equivalent) analyses should include those additional relevant variables along with the test scores (Standard 1.17, p. 21).
- When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported (Standard 1.18, pp. 21–22).

Other standards not presented here address the reporting of validity evidence for applications such as treatment assignment, use of meta-analytic evidence for instrument validation, and consequences of testing. Again, the reader is encouraged to consult the 1999 volume for a comprehensive treatment of reporting recommendations.

## Conclusion

In sum, the 2002 APA ethics code urges the behavioral scientist to select instruments that are useful and properly applied and whose reliability and validity have been established in the population of interest. The code provides no recommendations for procedures for establishing reliability and validity; in this chapter, we have attempted to provide more specific guidance. The code further prescribes that the scientist report evidence supportive of his or her instrument selection. Multiple professional organizations have articulated standards that address the appropriate reporting of measurement strategies; earlier efforts originated in the field of biomedical research and are in general neither detailed nor comprehensive enough to address the varied measurement concerns associated with

behavioral sciences research. In recent years, formal statements of reporting standards for research have been published for the behavioral sciences in particular, but these statements, including the APA's 2010 *Publication Manual*, reflect barely expanded measurement sections. To date, perhaps the most usefully prescriptive and exhaustive set of reporting recommendations for the behavioral sciences is provided by the 1999 *Standards* (AERA, APA, & NCME). It is unfortunate that the APA's *Publication Manual*, which is so widely used and readily available to those in the field, does not contain a more comprehensive measurement section. Indeed, awareness of ethical measurement practice in the behavioral sciences might be heightened if that section were expanded in future editions. In our view, the accountability associated with the reporting of evidence supportive of an instrument's implementation can only serve to improve adherence to ethical conduct.

We believe that a fair share of the burden involved in assuring the ethical use of behavioral measurement instruments falls to the instrument developer. Ideally, the developer will provide evidence of the reliability and validity of the instrument in the clearly defined population and applied setting for which it was designed, and—anticipating future uses to the extent possible—both provide reliability data for alternative populations and warn potential users against unsupported interpretations (AERA, APA, & NCME, 1999). Unquestionably, however, the research user of a behavioral measurement instrument bears a significant responsibility for ensuring that his or her choice of instrument has led to valid inferences.

The investigator should be aware that the inferences derived from measurement choices occur at all stages of the research process. Accordingly, the investigator following ethical practice will evaluate validity evidence relating to scoring, generalization, extrapolation, and implication inferences (Kane, 2006) and the generalizability of the instrument to the present population. The researcher will also be conversant with the various statistical frameworks for estimating instrument precision and will evaluate the particular forms of reliability evidence that are most relevant to her or his proposed use of the instrument, taking care to consider the fit of any measurement models used to the observed data. In analyzing study data, the investigator will choose measurement models that are appropriate for the level of measurement attained by a particular measurement procedure, taking into account model assumptions and relevant analytical and simulation study findings; furthermore, if latent variable models are used, the names given to latent constructs will be selected and explained with great care. Finally, the investigator should adhere to the scientific and ethical reporting standards summarized above and refer especially to the *Standards* (AERA, APA, & NCME, 1999) for concrete recommendations relevant to particular measurement strategies. The researcher should discuss in the research report the potential limitations of the measurement

process and its associated inferences and interpretations. At all stages of the research and reporting process, informed professional judgment will be required.

## Case Example

Many of the issues we have raised are highlighted in the literature on the design and interpretation of the Implicit Association Test (IAT). We have stressed the importance of assessing the reliability and validity of a measurement instrument in a particular population and applied setting; in the present section, we do not contravene our earlier advice by offering discussion about the properties of the IAT per se, but we do attempt to pinpoint issues that have likely pertained to most, if not all, of the instrument's applied uses.

The IAT is billed primarily as a measure of implicit bias toward a specific group (e.g., an ethnic minority, the elderly). Here, an *implicit* bias refers to one that is beneath awareness and presumably outside conscious control (Greenwald & Banaji, 1995); it may be contrasted with an *explicit* bias, of which the individual presumably is aware and would be able to control if motivated to do so. An intriguing feature of the IAT is that those completing the measure reportedly feel unable to control their implicit biases, even when they realize their responses may be revealing them.

The initial description of the IAT was published in the June 1998 issue of the *Journal of Personality and Social Psychology* (Greenwald, McGhee, & Schwartz). In its most basic form, the IAT is administered by seating the respondent at a computer. Displayed on the monitor are words and/or images, to which the respondent reacts by pressing a key. Most frequently, the reaction involves classifying the word or image into one of two contrasting categories (e.g., *young* vs. *old*, *good* vs. *bad*). Response data reflect the computed latency between the time the word or image appears on the screen and the time at which the respondent presses the key to indicate the correct categorization. The response latencies for different types of categorization are combined to produce an overall score that reflects any bias favoring one group over the other. Imagine, for example, that the IAT was being used to assess bias against the elderly: If the respondent were faster to categorize young faces and positive using the same key and old faces and negative using the same key than they were to categorize young and negative and old and positive, then the respondent would be assumed to possess an unconscious bias in favor of young people and against older people. On the other hand, if the opposite pairing were faster, then the respondent's score would be assumed to reflect

an unconscious bias in favor of older people and against younger people. After its development, the IAT was quickly and widely endorsed by the larger research community, as evidenced by its use in at least 122 research studies published through January 2007 (summarized in the meta-analytic report of Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

The IAT has also captured interest outside academe. The IAT was published on a freely accessible website in October 1998; through Project Implicit, funded by the National Institute of Mental Health and the National Science Foundation, the IAT remains available for self-administration at https://implicit.harvard.edu/implicit, where it is completed approximately 15,000 times each week (with a total of approximately 4.5 million completions since it first appeared online; Project Implicit, n.d.). The measure is also regularly featured in the popular media (e.g., Chedd, 2007; Thompson, 2009; Tierney, 2008; Vedantam, 2005), where it is described using statements such as, "The tests get to the bottom of our true inclinations" (Thompson, para. 3).

The rare recognition and acceptance of the IAT by the larger public has led to scrutiny of the measure that is somewhat uncommon for instruments developed primarily for research purposes. The result of this scrutiny is a growing literature questioning the reliability and validity of implementations of the IAT. These questions focus primarily on three broad concerns.

The first concern is that the reliability evidence associated with the IAT may not be sufficiently consistent and strong to warrant the relatively unqualified acceptance the measure has enjoyed. To date, the small amount of information offered on the precision of the IAT has been in the form of test–retest reliability coefficients. Because the IAT is purported to tap an individual difference, its associated short-term test–retest coefficients should theoretically be high, perhaps in the neighborhood of .80 or greater. However, the range for 1-week to 1-month coefficients has typically ranged from .50 to .70 (e.g., Bosson, Swann, & Pennebaker, 2000). Such reliability estimates reflect reasonably high levels of measurement error and are furthermore not consistent with the idea that the IAT taps a stable characteristic.

The second concern is that the observed scores produced by the IAT may not be sufficiently valid measures of the construct of interest. The validity of the IAT as a measure of unconscious bias has been questioned since the measure was first introduced, and indeed, its developers offer access to more than 50 articles addressing validity concerns at http://faculty.washington.edu/agg/iat_validity.htm. Multiple forms of evidence appear to cast doubt on the appropriateness of construct-relevant interpretations. For example, evidence regarding the strength of association between implicit and explicit measures of the same bias is inconsistent, yet theoretical explanations for observed associations appear to adapt, to some

degree, to the observed data: Although correlations between measures of extrinsic bias and IAT measures of intrinsic bias vary widely across studies, a correlation produced in any single research study—whether high or low—tends to be interpreted in a manner that is favorable to application of the IAT. Indeed, in the seminal paper on the IAT (Greenwald, McGhee, & Schwartz, 1998), the authors report that two explicit measures of bias correlated with each other at an approximate $r$ of .60, whereas the same measures correlated with the IAT measure of implicit bias at $r = .25$. Although these findings might reasonably be interpreted as a failure to demonstrate convergent validity, the authors argue that they are instead an important demonstration of discriminant validity; however, the authors' provided evidence of criterion validity comes in the form of prediction by IAT scores controlling for explicit measures. Moreover, as described above, one major threat to the validity of a behavioral measurement instrument is irrelevant variance. In that connection, it is unfortunate that a detailed analysis of the IAT aimed at specifying an appropriate measurement model has revealed that variability in IAT scores can be attributed to a number of influences, including a cluster of variables that influences general processing speed (e.g., attention span, hand–eye coordination, mood; Blanton, Jaccard, Gonzales, & Christie, 2006). Questions have also been raised about the validity of the difference scores computed as part of the IAT. Although these issues have been addressed in revised forms of the test (e.g., Blanton et al.; Greenwald, Nosek, & Banaji, 2003; Olson & Fazio, 2004), they are relevant for a significant portion of the existing literature on IAT-assessed implicit cognition. Although a recently published meta-analysis suggests that IAT scores offer incremental validity over scores on traditional self-report measures (Greenwald et al., 2009), questions about the validity of IAT applications persist (e.g., Arkes & Tetlock, 2004).

A third set of concerns stems from the consequences of the ready public availability of the IAT. Upwards of 2,000 people per day complete an IAT at the Project Implicit website; multiple forms of the test are available on the website, including versions that purportedly reveal automatic preferences for "light-skinned" versus "dark-skinned" faces and for disabled versus abled individuals. At the conclusion of each test, the respondent receives a brief feedback statement (described as a "score;" e.g., "Your data suggest a moderate automatic preference for Young people compared to Old people") and is provided the opportunity to view a frequency table that provides the percentage of Internet respondents that received each possible "score."

As we have shown, however, the reliability and validity of IAT scores—although sufficient to support continuing research and development—are not strong. The Project Implicit team has been careful to avoid overstating the validity of the measure, stating in the website's FAQ that "these tests are not perfectly accurate by any definition of accuracy." But media reports and other websites that steer people to the Project Implicit website are not as

careful, and it seems unlikely that the average test-taker would consult the large amount of information about implicit cognition and the IAT offered by the Project Implicit team in a separate section of their website. Hence it is potentially misleading, and perhaps even harmful, for the website to communicate to the respondent that his or her IAT performance reflects the presence or absence of unconscious bias. A serious concern is the respondent's interpretation of his or her website-provided feedback and the potential consequences of that interpretation. For test-takers who prefer to view themselves as unprejudiced, an unquestioned IAT "score" that indicates otherwise could cause distress; for any test-taker, shared feedback could have social repercussions. Finally, although the IAT seeks to tap a presumably stable (and unconscious) individual difference, it is unclear how an individual's use of the Project Implicit website might impact his or her later responses to the IAT in the context of participation in a research study.

In sum, the IAT is an intriguing instrument that has captured the attention of many in the behavioral sciences research community and the interest of the media and members of the general public alike. Unfortunately, the widespread use of the IAT raises important ethical questions. Although ethical practice would suggest the use of instruments whose reliability and validity have been supported in the population tested, such evidence is not currently sufficient even for controlled research applications of the IAT, and to our knowledge, it is virtually nonexistent for the Internet-based population of IAT test-takers. Fundamental questions especially persist regarding the appropriateness of the IAT for construct-relevant interpretations. The impact of such questions on the validity of experimental findings is clear. Of equal concern are questions regarding the usefulness and potential impact, at both the individual and societal levels, of self-interpretation of IAT performances on the publicly available forms of the test.

As is illustrated here, basic concerns about how behavioral measures are described, used, and administered are more the norm than the exception. The IAT is a particularly useful case example because it illustrates these concerns as they play out in both the research context and in popular culture. Although the translation of somewhat arcane processes such as implicit cognition into a form that resonates with the public is a commendable goal, the IAT does serve as an example of the challenges involved.

## References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., … Lang, T. (2001). The Revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663–694.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35,* 33–40.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15,* 257–278.

Bailar, J. C., III, & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of Internal Medicine, 108,* 266–273.

Barbour, V., Moher, D., Sox, H., & Kahn, M. (2005). Standards of reporting biomedical research: What's new? *Science Editor, 28,* 4.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370–379.

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

Blackman, N. J.-M., & Koval, J. J. (1993). Estimating rater agreement in 2 x 2 tables: Corrections for chance and intraclass correlation. *Applied Psychological Measurement, 17,* 211–223.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42,* 192–212.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79,* 631–643.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … de Vet, H. C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Annals of Internal Medicine, 138,* 40–44.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger Publishers.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse $2^P$ tables. *British Journal of Mathematical and Statistical Psychology, 59,* 173–194.

Chedd, G. (Writer and Director). (2007). The hidden prejudice [Television series episode]. In J. Angier & G. Chedd (Executive Producers), *Scientific American Frontiers*. Public Broadcasting Corporation.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Des Jarlais, D. C., Lyles, C., & Crepaz, N. (2004). Improving the reporting quality of nonrandomized evaluations of behavior and public health interventions: The TREND statement. A*merican Journal of Public Health, 94,* 361–366.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Dunivant, N. (1981). *The effects of measurement error on statistical models for analyzing change: Final report* (Grant NIE-G-78-0071). Washington, DC: National Institute of Education (Educational Resources Information Center Document Reproduction Service No. ED223680). Retrieved from Educational Resources Information Center database.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8,* 102–109.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger Publishers.

Hägglund, G. (1982). Factor analysis by instrumental variables. *Psychometrika, 47,* 209–222.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12,* 38–47.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks, CA: Sage Publications.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Jöreskög, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57,* 239–251.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education, 9,* 355–379.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

Koocher, G. P., & Keith-Spiegel, P. (2008). *Ethics in psychology and the mental health professions: Standards and cases* (3rd ed.). New York: Oxford University Press.

Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51,* 981–983.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–548.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics, 12,* 383–394.

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., & Fenton, M. (2000). Unpublished rating scales: A major source of bias in randomized controlled trials of treatments for schizophrenia. *British Journal of Psychiatry, 176,* 249–252.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies in establishing measurement equivalence/invariance. *Organizational Research Methods, 7,* 361–388.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Messick, S. (1988a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Messick, S. (1988b). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741–749.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100,* 398–407.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86,* 653–667.

Project Implicit. (n.d.). Retrieved from http://projectimplicit.net/generalinfo.php

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Thompson, B., & Daniel, L. G. (1996). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement, 56,* 741–745.

Thompson, J. (2009). *Project Implicit: Am I racist?* Retrieved from http://www.myfoxchicago.com/dpp/news/project-implicit-am-i-a-racist-dpgo-20091029-jst1256860352012

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Tierney, J. (2008, November 17). In bias test, shades of gray. *The New York Times*. Retrieved from http://www.nytimes.com/2008/11/18/science/18tier.html

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Vedantam, S. (2005, January 23). See no bias. *The Washington Post*, p. W12.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79.

Wright, T. A., & Wright, V. P. (2002). Organizational researcher values, ethical responsibility, and the committed-to-participant research perspective. *Journal of Management Inquiry, 11,* 173–185.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger Publishers.

# 6

# *Ethics and Sample Size Planning*

**Scott E. Maxwell**
*University of Notre Dame*

**Ken Kelley**
*University of Notre Dame*

Most psychological researchers realize at least one way in which *sample size* planning is important. Namely, if their study fails to have sufficient *statistical power*, they run an increased risk of not being able to publish their work. Despite this awareness, literature reviews usually show that underpowered studies persist. In fact, underpowered studies permeate the literature not only in psychology (e.g., Bezeau & Graves, 2001; Clark-Carter, 1997; Kosciulek & Szymanski, 1993; Mone, Mueller, & Mauland, 1996; Rossi, 1990; Sedlmeier & Gigerenzer, 1989) but in other disciplines as well. For example, recent reviews in such areas of medicine as head injury (Dickinson, Bunn, Wentz, Edwards, & Roberts, 2000), orthopedics (Freedman & Bernstein, 1999), rheumatology (Keen, Pile, & Hill, 2005), stroke (Weaver, Leonardi-Bee, Bath-Hextall, & Bath, 2004), and surgery (Maggard, O'Connell, Liu, Etzioni, & Ko, 2003) show frequent instances of inadequate statistical power.

Although there are a variety of possible explanations for why underpowered studies persist (cf. Maxwell, 2004), the primary focus of this chapter pertains to the consequences of inappropriate sample size planning, especially, but not exclusively, as it relates to underpowered studies. One negative consequence of underpowered studies in a discipline is the likely preponderance of studies that appear to contradict one another. However, a number of authors (e.g., Goodman & Berlin, 1994; Hunter & Schmidt, 2004; Maxwell, 2004) have pointed out that these apparent contradictions may, in fact, reflect nothing more than sampling error. Although this state of affairs is clearly undesirable, it is not immediately clear that there are necessarily ethical ramifications. The specific purpose of this chapter is to present various aspects of the relationship between sample size planning and ethics.

Before delving into the relationship between sample size planning and ethics, it is helpful to consider the basic purpose of research in psychology and related sciences. The vast majority of empirical research in these disciplines addresses either or both of two goals: (a) establishing whether a relationship exists between variables, which also usually involves establishing the direction of any such relationship between the variables, and (b) estimating the magnitude of a relationship between variables. For example, Festinger and Carlsmith's classic 1959 study on cognitive dissonance sought to determine whether individuals would rate a boring study as more enjoyable if they were paid $1 or $20. The goal was not to estimate a parameter, but instead to differentiate between psychological theories that made different predictions about which level of payment would produce higher ratings of enjoyment. From a statistical perspective, the goal was to establish directionality, irrespective of magnitude. As an example of a study with a very different type of goal, Sternberg and Williams (1997) examined the validity of Graduate Record Examination (GRE) scores for predicting various measures of success in graduate school. In their case, it was of little interest to determine whether GRE scores relate positively or negatively to success because it was already well established that the relationship was not negative. Instead, their goal was to estimate the magnitude of the relationship.

We believe that this distinction between directionality and magnitude has important implications for how researchers should plan an appropriate sample size. Later in this chapter we will also explore possible ways in which this distinction may affect ethical implications as well. From a purely methodological perspective, researchers whose primary purpose is to establish existence and directionality of an effect should plan sample size with statistical power considerations in mind, whereas researchers whose primary purpose is to estimate magnitudes should plan sample sizes with accuracy of parameter estimation in mind.

We use the term *accuracy in parameter estimation* (often abbreviated AIPE) to emphasize that we are simultaneously considering both precision and bias. We operationalize the idea of an accurate estimate in terms of obtaining a narrow confidence interval at some specified confidence level coverage (e.g., 95%). A narrow confidence interval is desirable because it provides a tight bound on the set of plausible parameter values. Thus, holding constant confidence level coverage at a specified percentage, the narrower the range of the plausible parameter values, the more accurate is the parameter estimate.

Because the *power analysis* and the accuracy in parameter estimation approaches to sample size planning are based on a fundamentally different set of goals, the necessary sample sizes from the two approaches can be vastly different. For example, consider a researcher whose goal is to obtain an accurate estimate of the correlation between GRE scores and a

measure of success in graduate school in some clearly defined population. Suppose this researcher expects that the population correlation is around .50, which corresponds to Cohen's (1988) "large" *effect size* for a correlation. If this researcher planned his or her sample size to have statistical power of .80 to detect a correlation of this magnitude, he or she would find that a sample size of 28 would be necessary. But to what extent would a sample of size 28 yield an accurate estimate of the population correlation? For simplicity, suppose that the sample correlation coefficient turned out to be exactly .50. Then a 95% confidence interval for the population correlation coefficient would stretch from .16 to .74. This interval allows the researcher to assert that the population correlation is positive, but it hardly pinpoints the population value. In contrast, suppose the researcher desires a confidence interval with a width of only .10 instead of .58. It turns out that the sample size necessary to obtain a confidence interval whose expected width is .10 would be approximately 850, compared with the sample size of only 28 needed to obtain statistical power of .80. The point is not that either 28 or 850 is the "correct" sample size in this situation, but rather that considerations of accuracy instead of statistical power can imply different choices for sample sizes. Of course, researchers may be at least somewhat interested in both purposes, but even in such situations it is helpful to consider sample size planning from the perspective of both statistical power and accuracy in parameter estimation. This combined perspective can sometimes be operationalized in a single step (e.g., Jiroutek, Muller, Kupper, & Stewart, 2003) or using the larger of the two sample sizes derived from separate considerations of statistical power and accuracy in parameter estimation.

How does ethics enter into the picture? The primary question we will consider in this chapter is the ethical implication of designing a study that is unlikely to be able to answer its intended question as a result of inadequate sample size planning. For example, suppose a study whose goal is to determine directionality is designed with insufficient statistical power. Or suppose a study whose goal is to estimate a parameter is designed in such a way that the parameter is unlikely to be estimated accurately. In both cases, the study is unlikely to answer the scientific question it is intended to address. What are the ethical implications of such shortcomings?

An equally serious possibility is that studies can be designed with sample sizes that are too large. For example, a researcher may design a study with a sample size that would provide statistical power of .99 to detect a miniscule effect even though such a small effect would be viewed as unimportant. Similarly, a study that provided a narrower confidence interval than any reasonable person would deem to be necessary could also be considered wasteful. In both cases, sample sizes that are too large can be unethical. For example, more research participants than necessary may be exposed to a potentially harmful treatment. It may also take

longer to accrue such a large sample, in which case a beneficial treatment may not become available as rapidly as it otherwise could have. Although these are serious problems, we will not focus on them in this chapter primarily because literature reviews in psychology and many other fields show clearly that with only occasional exceptions, sample sizes suffer from being too small, not too large.

For many years, methodologists have urged substantive researchers to design studies so that statistical tests have adequate statistical power. Entire books have been written explaining the importance of adequate statistical power and delineating how to determine an appropriate sample size to ensure adequate statistical power for a wide variety of designs and analyses. In a system where studies failing to show statistically significant results may end up in the *file drawer* (Rosenthal, 1979) instead of being published, the risk of failing to publish would seem to provide sufficient motivation for researchers to learn to conduct power analysis and implement procedures to guarantee adequate statistical power in their research. Cohen (1992) describes being spurred to write a primer on power analysis by an associate editor of *Psychological Bulletin*, who believed that researchers failed to perform power analyses even though they knew they should because it was simply too difficult. However, in the intervening years, power analysis has become much more user-friendly with a number of readable primers (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Murphy & Myors, 1998) and the emergence of multiple software programs and packages (Borenstein, Rothstein, & Cohen, 1997; Elashoff, 1999; Hintze, 1996; O'Brien, 1998). Despite unmistakable progress in making power analysis more accessible, any impact this has had on progress appears to be slow or arguably even nonexistent because literature reviews generally tend to show that underpowered tests persist. Maxwell (2004) has speculated that the main reason progress has been so slow is because researchers do not need to have adequately powered statistical tests to obtain *at least some* statistically significant results. The argument behind this seeming paradox is simple, namely, that studies almost invariably involve testing *multiple hypotheses*. Even if the statistical power associated with any single test is low, the probability of obtaining at least one statistically significant result among a collection of tests can easily be large. For example, Kelley, Maxwell, and Rausch (2003) show that a researcher who tests five orthogonal comparisons in a six-group study where the levels of statistical power of the five individual tests are .5, .4, .3, .2, and .1 will have an 85% chance of obtaining at least one statistically significant result. Similarly, Maxwell (2004) shows that a researcher whose study has a statistical power of .26 for any single regression coefficient in a multiple regression analysis with five predictors may have an 84% chance of obtaining at least one statistically significant coefficient. Such examples underscore the point that the concept of the statistical power associated with a study

is vague because every individual test in a study can have low statistical power and yet the probability that the researcher obtains at least one statistically significant result can easily exceed the conventional standard of .80. Thus, what at first glance appears to be a major motivator for researchers to design their studies so their statistical tests have adequate statistical power often turns out not to be any real motivation at all whenever multiple tests are to be performed.

If an individual researcher can probably expect to obtain publishable results despite each and every one of his or her individual statistical tests being underpowered, are there other arguments that might still persuade researchers to design their studies differently, so as to ensure that the individual statistical tests themselves have adequate statistical power? In particular, should researchers be told that they have an ethical obligation to design their studies with adequate statistical power for individual statistical tests? Similarly, should researchers whose goal is to estimate a parameter be told that they have an ethical obligation to design their studies with adequate accuracy? These questions form the central focus of the current chapter.

## General Considerations of Ethics in Research

There has been increasing sensitivity into the ethical dimensions of research, especially as it relates to living organisms. For example, federal regulations in the United States stipulate that for a proposed study to be ethical, its projected benefits must outweigh the projected risks to participants (Bacchetti, Wolf, Segal, & McCulloch, 2005; The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Along similar lines, Emanuel, Wendler, and Grady (2000) proposed seven requirements to determine whether a clinical research study is ethical. Their requirements include not only a favorable *risk–benefit ratio* but also adequate scientific validity.

One aspect of scientific validity is whether a proposed study is likely to provide an appropriate answer to the scientific question of interest. The likelihood of obtaining an appropriate answer clearly depends on many aspects of a study. For example, the framework espoused by Shadish, Cook, and Campbell (2002) suggests that the quality of a study can be evaluated in terms of four types of validity: statistical conclusion, internal, external, and construct. One factor involved in statistical conclusion validity relates to statistical power and/or accuracy in parameter estimation. A study with low statistical power runs the risk of failing to provide the correct answer to the scientific question of interest. Similarly, a study

that fails to estimate a parameter accurately may leave too much doubt about the magnitude of a parameter to be truly useful.

It is difficult to see how anyone would disagree with the position that proposed studies should be scientifically valid and should have favorable risk–benefit ratios. However, it is less clear what this means or should mean in practice, especially as it relates to institutional review boards (IRBs) and data safety and monitoring boards. For example, Rosoff (2004) states that "it is not clear whether IRBs view flawed study designs or statistical analysis plans as placing subjects at increased risk" (p. 16). Should IRBs reject proposed studies that fail to provide persuasive evidence that, among other things, statistical power and/or accuracy in parameter estimation will be sufficient? Similarly, should journals reject manuscripts that fail to provide evidence of adequate statistical power and/or ability to estimate parameters accurately? And complicating matters yet further is the issue of how to handle the existence of multiple statistical tests—should authors be required to provide evidence that each individual test of a major hypothesis has adequate statistical power? Or is it enough that there be a reasonably high probability of being able to find at least one statistically significant result after performing multiple tests? To begin to address these questions, we will initially draw on perspectives developed largely in medical research.

## Perspective From Medical Research

Questions of ethical implications of statistics may have received relatively little attention in psychology, but they have been the source of heated debate in medical research (see also Fidler, Chapter 17, this volume). This presumably reflects the fact that medical research may literally involve issues of life and death, and in particular what medical treatment an individual is randomly assigned to can potentially save or greatly improve that person's life. However, many psychological studies also have the potential to affect individuals' quality of life, either directly or indirectly and either positively or negatively. Even laboratory studies where the risk to participants may seem to be low, nevertheless impose on individuals' time. In addition, participants are either explicitly or implicitly led to believe that their involvement has some plausible potential to contribute to scientific knowledge. However, if, in fact, there is little chance that scientific progress will result from such participation, ethical concerns may be appropriate because even if the risk is low, the likely reward may be even less. More generally, because issues of ethics and sample size planning have received considerable attention in medical research, it is helpful

to consider the various arguments that have been advanced in the medical literature.

As long ago as 1978, arguments have been advanced in the medical literature that sample size planning can have ethical implications. For example, Newell (1978) stated, "The ethical problem arises if the trial procedures cause additional discomfort or risk for the patient, but the trial as a whole is very unlikely to identify as significant a therapeutic effect which the clinician would regard as worth achieving" (p. 1789). Two years later, Altman (1980) echoed this sentiment in his statement that, "It is surely ethically indefensible to carry out a study with only a small chance of detecting a treatment effect unless it is a massive one, and with a consequently high probability of failure to detect an important therapeutic effect" (p. 1338). Twenty-two years later, in an influential article in the *Journal of the American Medical Association*, Halpern, Karlawish, and Berlin (2002) noted:

> Many clinical investigators continue to conduct underpowered studies and fail to calculate or report appropriate (a priori) power analyses. Not only do these scientific and ethical errors persist in the general medical literature, but 3 recent reports also highlight the alarming prevalence of these problems in more specialized fields. (Halpern et al., 2002, p. 358)

Why have these and many other authors argued that sample size planning is an ethical issue, and in particular that inadequate sample size planning is unethical? As a prelude, it is worth noting that these arguments have largely addressed issues of statistical power to the exclusion of accuracy in parameter estimation. Thus, for the moment we will also focus on statistical power, although we will return to the issue of sample size planning, accuracy, and ethics later in the chapter. The essence of the main argument is that studies whose tests are underpowered are unlikely to obtain statistically significant results. Studies without statistically significant results are unlikely to be published, a premise that has now received support in a variety of disciplines dating at least as far back as Sterling (1959) more than half a century ago. As a result, such studies run a high risk of being relegated to the "file drawer" and thus failing to contribute to cumulative knowledge. As a consequence, such studies violate the basic ethical tenet of a favorable risk–benefit ratio, as mandated by federal regulations. If the benefit is zero for unpublished studies, then the risk–benefit ratio cannot be favorable, which proponents of this position then view as implying that such studies are unethical. For example, suppose for a moment that contrary to reality, Festinger and Carlsmith's (1959) study of cognitive dissonance had led to statistically nonsignificant results. In all likelihood, such findings would not have been published because they would have failed to resolve the underlying theoretical debate about

cognitive dissonance. To the extent that such ambiguous results could have been attributed to inadequate sample size planning, the individuals who participated in the study would have been led to participate in a study that wasted their time because there was minimal likelihood that the study would have any benefit whatsoever. Its inadequate sample size planning would have all but have guaranteed that it would end up in the "file drawer" instead of being read by thousands of individuals.

Of course, in some areas of research the risk to the individual may be very low. Even so, proponents of the perspective that underpowered studies are unethical could still maintain that exposing research participants even to a presumably innocuous laboratory study is unethical if there is little chance that the study will be published because it will therefore be unlikely to make any type of positive contribution to the literature. Most important, from an ethical perspective, is that potential research participants are likely to have received a message from the researchers that their participation will advance scientific knowledge. However, if a study is underpowered and likely to end up in the "file drawer," a case can be made that such a practice is unethical. Halpern et al. (2002) have argued that researchers who design underpowered studies are ethically bound at the very least to inform potential participants that their involvement in the proposed research may have only a minimal probability of contributing to knowledge. In fact, as we will discuss later in the chapter, the situation may be even worse because publication pressures may cause researchers to "torture" their data (Mills, 1993) in the search for statistically significant results or engage in *selective reporting* or continue to recruit participants until statistical significance is obtained. These various practices may produce consequences worse than failing to add to knowledge because they run the risk of placing misleading results in the literature. These are admittedly complicated problems, which is why we will devote separate sections to them later in the chapter, but we will simply point out now that inadequate sample size planning plays a role in all of them.

The position that underpowered studies are an ethical problem has not gone unchallenged. These challenges have largely been based on three perspectives. First, as Lilford (2002) and others have pointed out, there is typically subjectivity involved in determining whether a study is underpowered. The most obvious source of subjectivity is the necessity to determine the magnitude of an appropriate effect on which to base statistical power. Lipsey (1990) has labeled effect size as the "problematic parameter" because in many situations it is difficult to determine an appropriate a priori value for the effect size on which power analyses depend. The immediate consequence is that a study that is underpowered for a small effect size may, in fact, be overpowered for a large effect size, but there may be little agreement even among experts as to whether sample size planning

for the study in question should be based on a small or a large effect size. As Schulz and Grimes (2005) state in the context of clinical trials:

> How can a process rife with subjectivity fuel a black-white decision on its ethics? With that subjectivity, basing [clinical] trial ethics on statistical power seems simplistic and misplaced. Indeed, since investigators estimate sample size on the basis of rough guesses, if deeming the implementation of low power trials as unethical is taken to a logical extreme, then the world will have no trials because sample size determination would always be open to question. (Schulz & Grimes, p. 1351)

Although their statement is in the specific context of clinical trials, the logical basis of their argument applies more generally.

Second, McCartney and Rosenthal (2000) and Prentice and Miller (1992), among others, have argued that psychologists need to be aware of the potential scientific importance of effect sizes that have conventionally been labeled as small. However, this can lead to practical problems because appropriate sample sizes to achieve adequate statistical power to detect small effect sizes may be much larger than what most psychologists regard as large, which may be beyond the resources of most investigators. As Edwards, Lilford, Braunholtz, and Jackson (1997) point out, "the requirement to seek to avoid false-negative trial results often entails surprisingly large sample sizes" (p. 804). Along these lines, Hunter and Schmidt (2004) state that "for correlational studies, 'small sample size' includes all studies with less than a thousand persons and often extends above that" and "for experimental studies, 'small sample size' begins with 3000 and often extends well beyond that" (p. 14). Edwards et al. (1997) go on to point out that:

> Clinicians do not often have ready access to the scarce resources required to mount such large-scale trials which typically involve many centres. Hence, ethics committees, by taking statistical power into account, collectively thwart many independent investigations and so may seriously diminish the stock of the world's knowledge. (Edwards et al., 1997, p. 804)

The unfortunate side effect is that investigators may be rewarded only for studying effects large enough to be detected with moderate sample sizes, irrespective of ultimate scientific importance.

Third, the recent emergence of *study registries* and data synthesis methods such as *meta-analysis* has led some authors to conclude that even if studies are underpowered they can nevertheless collectively contribute to the accumulation of knowledge. We will discuss issues involved in study registries and meta-analysis later in a separate section of the chapter.

Yet another perspective has recently been offered by Bacchetti et al. (2005). They begin with the general premise that for the design of a study to be ethical, the projected value of the study must exceed the projected risks to participants. They then develop a model of projected value and projected burden per participant. Their model implies the surprising result that small studies should never be criticized from an ethical perspective simply because they are small, and that in fact the only relevant question is whether a study is larger than can be ethically justified. Prentice (2005) and Halpern, Karlawish, and Berlin (2005) question this conclusion because they doubt some of the basic assumptions adopted by Bacchetti et al. in formulating their model.

Where does all of this leave psychology and other disciplines from the perspective of the relationship between ethics and sample size planning? To address this question, we need to consider yet one more aspect of how sample size planning often works in the real world of research. Our impression is that psychology and related disciplines take sample size planning seriously only in some very circumscribed circumstances. For example, some mention of sample size planning is routinely expected in grant applications, and some thesis and dissertation committee members may expect students to justify their proposed sample sizes before embarking on their projects. However, we are aware of no journals or IRBs that explicitly require psychologists to justify their sample sizes before conducting their research and deny research based on inadequate sample sizes. We will say more about this dichotomy later in the chapter, but for the moment, we will restrict our attention to situations such as grant writing, where some justification of sample size planning is typically expected. How does this process typically work in practice?

In theory, one of the first steps in sample size planning involves determining the magnitude of relevant effect size parameters. After deciding on the desired level of statistical power and/or desired level of accuracy in parameter estimation, a corresponding sample size is determined. Although this is the theory, actual practice often works differently. As Schulz and Grimes (2005) state, "Investigators sometimes perform a 'sample size samba' to achieve adequate power. This dance involves retrofitting of the parameter estimates (in particular, the treatment effect worthy of detection) to the available participants" (p. 1351). Goodman and Berlin (1994) concur:

> A typical sample size consultation often resembles a ritualistic dance. The investigator usually knows how many participants can be recruited and wants the statistician to justify this sample size by calculating the difference that is "detectable" for a given number of participants rather than the reverse. … The "detectable difference" that is calculated is typically larger than most investigators would consider important or even likely. (Goodman & Berlin, 1994, p. 203)

Similarly, Edwards et al. (1997) state:

> Many trials only receive approval because committees tend to accept unrealistic forecasts of recruitment, putative treatment gains much greater than those required to alter practice, and false-negative rates [Type II errors] four times greater than conventional 5% false-positive rates [Type I errors]. (Edwards et al., 1997, p. 805)

In our view, this represents a serious ethical problem, even if one is not persuaded that underpowered studies are necessarily intrinsically unethical. In the sample size "samba" described previously, researchers knowingly misrepresent the appropriate magnitude of the effect size parameter based on achieving a desired power and given available participants. An argument can be made that an unfair system forces researchers to act this way because they have no other alternative, but the fact remains that the process is deceitful and, therefore in our judgment, unethical. Why has the system evolved in such a way that many researchers are forced to act in what we regard as an unethical manner?

We believe that this unfortunate state of affairs has arisen in no small part because of a failure to distinguish sample size planning for power analysis from sample size planning for accuracy in parameter estimation. Beginning in the 1960s, largely because of Cohen's efforts, psychologists were exposed to the importance of designing studies with adequate statistical power. Interestingly, although the past decade or more has seen calls for more emphasis on effect sizes and accompanying *confidence intervals* (Wilkinson & the Task Force on Statistical Inference, 1999), there has been much less recognition that sample sizes need to be planned differently for estimating parameters than for testing hypotheses. If researchers are to take seriously the new American Psychological Association (APA) publication manual guidelines, where it is stated that "it is almost always necessary to include some measure of effect size in the Results section" and "whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size" (APA, 2010, p. 34), sample size planning for narrow confidence intervals (i.e., accuracy in parameter estimation) will continue to grow in importance. The main reason is that researchers do not wish to have an effect size with a correspondingly "embarrassingly large" confidence interval (Cohen, 1994, p. 1002). It is worth noting that Cohen is almost single-handedly responsible for calling the importance of power analysis to the attention of psychologists, beginning in the 1960s, but that he later became very critical of null hypothesis significance testing (e.g., Cohen, 1994). Sample size planning no longer depends on statistical power when the emphasis switches from hypothesis testing to parameter estimation. Foundational work on the differences between sample size

planning in terms of statistical power versus accuracy, as well as examples of methods intended for planning sample sizes to estimate parameters accurately, can be found in historical references such as Guenther (1965) and Mace (1964), with modern treatments for widely used effect sizes in psychology in such sources as Pan and Kupper (1999), Kelley and Maxwell (2003, 2008), Kelley and Rausch (2006), Kelley (2008), Maxwell, Kelley, and Rausch (2008), and Jiroutek et al. (2003).

How has the failure to distinguish sample size planning for accuracy in parameter estimation versus statistical power contributed to what we perceive as an ethical violation? First, the goal of many research projects is to estimate one or more parameters accurately, not necessarily to test a hypothesis. In such cases, requiring a power analysis is inappropriate because the power analysis addresses a different question. We should hasten to add that sample size planning remains important, but the appropriate methodology is different from the better known power analytic approach, and thus the size of an appropriate sample may also be different, as we showed earlier in the case of a correlation coefficient. Unfortunately, however, our impression is that grant review committees and journals have generally not caught up with this emerging distinction, and thus researchers are often placed in an untenable position where they are required to perform a power analysis before beginning their proposed study, even though a power analysis may not pertain to their research goals.

We believe that there is a second way in which the failure to distinguish sample size planning for accuracy in parameter estimation versus statistical power has contributed to ethical violations. In our view this second factor is even more fundamental to the overall research enterprise, but it is also somewhat more complicated to explain. To understand the connection, it is important to consider the broader issue of why there is a bias against the null hypothesis. It is well accepted that studies without statistically significant results are less likely to be published than studies with statistically significant results. On the one hand, this seems sensible when the goal of a study is to test a hypothesis. From the perspective of Jones and Tukey (2000), the major piece of information provided by a null hypothesis statistical test is the direction of an effect. A study that fails to obtain a statistically significant result when testing a hypothesis is thus unable to determine the direction of the relevant effect. In this sense, it is debatable whether it would be worth readers' time and effort to read about a study whose eventual conclusion is that the direction of the effect cannot be determined. Of course, a counterargument is that this study could provide a basis for synthesizing results across studies (i.e., be a component of a meta-analysis), so it has not necessarily literally failed. Although we acknowledge that there are situations where this perspective is sensible, we would still maintain that even in a situation such as this it is typically preferable to include such a study in a registry instead of publishing it.

On the other hand, consider a study whose main goal is to estimate a parameter accurately. The results of such a study should typically be reported in terms of a confidence interval (e.g., Cumming & Fidler, Chapter 11, this volume). In this case, the "success" of the study depends on two factors: (a) ideally the estimate of the parameter is unbiased, and (b) ideally the confidence interval is narrow. As we saw earlier when the confidence interval for a correlation coefficient stretches from .16 to .74, we have learned little about the value of the population parameter, the value of ultimate interest, except that the researcher can infer that the population correlation coefficient is positive. However, we probably already knew that GRE scores are not negatively related to measures of graduate school success, so this extremely wide confidence interval implies that the study itself almost certainly failed to meet its goal of estimating the population correlation coefficient accurately. In particular, to be successful the study would need to have provided a much narrower 95% confidence interval, which would in turn have required a much larger sample size.

There is a critical difference between this hypothesis testing situation and this parameter estimation situation. Namely, in the former, the possible answers are "negative," "positive," or "cannot be determined." Notice that the boundaries between results are categorical, at least for a fixed alpha level. However, in the parameter estimation situation, there are no hard and fast boundaries between possible study outcomes. Instead, the accuracy of estimating a parameter is a matter of degree. From this perspective, some studies are more informative than others because they provide a more accurate estimate of the parameter, such as when one study provides a narrower confidence interval than another study, holding other relevant factors constant. However, even a study whose confidence interval is wide should be judged as making a valuable contribution to the literature if the corresponding interval conveys new information about an important parameter.

What are the implications of this distinction for sample size planning and accompanying ethical considerations? An underpowered study runs a serious risk of failing to provide meaningful results. However, researchers often realize this and engage in the "samba" dance, selective reporting, data torture (Mills, 1993), and other questionable ethical practices. In such situations, the solution would appear to be honesty in establishing the effect size underlying the power analysis. However, a power analysis is generally not relevant when the goal of the study is to estimate a parameter. In such situations, ethical concerns about underpowered studies are also irrelevant. Instead, the emphasis shifts to accuracy in parameter estimation. Although we would maintain that sample size planning is appropriate whether the goal of the study is significance testing or accuracy in parameter estimation, we nevertheless believe that the ethical considerations are different in these two cases.

## Consequences of Efforts to Achieve Significance Despite Inadequate Statistical Power

How are so many studies published if (a) statistical significance is a virtual prerequisite for publication and yet (b) most studies are underpowered? Maxwell (2004) offers the simple explanation that virtually every study, at least in psychology, involves testing multiple hypotheses. Although the statistical power for any single specific hypothesis may be woefully low, if enough tests are performed, it is highly likely that the prospective author will find something to write about (i.e., a statistically significant result will be obtained). For example, Cohen's (1962) original survey that documented generally low statistical power of tests in social/personality psychology was based on 70 studies. However, these 70 studies included 4,820 statistical tests, or an average of nearly 70 tests per study! Similarly, Chan, Hròbjartsson, Haahr, Gotzsche, and Altman's (2004) survey of randomized clinical trials found a mean of 37 tests per trial and a median of 27 tests per trial. Even if the statistical power of each individual test is exceedingly low, it is highly likely that if many hypothesis tests are performed at least one will reach statistical significance, unless the Type I error rate of each test were to be adjusted for the number of tests. Of course, one complication arising from lowering the Type I error rate for each individual test is an inevitable decrease in the statistical power for that same test, often leading right back to the basic problem of underpowered tests.

Although many studies may involve many more than a single test, few if any of these studies will report the results of every test that was performed. There can be a fine line between thorough examination of possible interesting questions and what Kerr (1998) labeled as "HARKing" (hypothesizing after the results are known), which he defined as "presenting a post hoc hypothesis in the introduction of a research report as if it were an a priori hypothesis" (p. 197). Kerr reported the results of a survey suggesting that HARKing is widespread in psychological research. For many years speculation has existed about the existence of selective outcome reporting. This phenomenon is similar to publication bias. However, publication bias distinguishes between papers, whereas selective reporting occurs within papers. In particular, selective outcome reporting occurs when some types of results are more likely to be reported in a paper than other types of results.

Chan et al. (2004) found clear evidence of selective outcome reporting in the 102 randomized clinical trials they surveyed. Specifically, outcomes with statistically significant results were much more likely to be reported than outcomes with nonsignificant results. Notice that selective reporting is most likely to occur when the power associated with individual tests is low, as is often the case in psychology. The effect of such selective

reporting will tend to be similar to the effect of publication bias, and in particular will make it very difficult to create a *cumulative science* because even exact replications of one another will tend to produce seemingly contradictory results when statistical power is low. Several recent studies (e.g., Bezeau & Graves, 2001; Clark-Carter, 1997; Kosciulek & Szymanski, 1993; Mone et al., 1996) suggest that in many areas of psychology, the typical power to detect a medium effect size is probably much closer to .50 than to .80. Ironically, the probability of discrepant results across findings is maximized when the power is .50. In this situation, two independent studies will be expected to disagree with one another about statistically significant effects fully 50% of the time. Unfortunately, it is also the case that they will wrongly agree with one another 25% of time, leaving only the remaining 25% of the time when the two studies correctly agree.

To what extent is selective reporting an ethical concern? There is not necessarily a simple answer to this question, but for illustrative purposes consider a simple case of a psychotherapy outcome study intended to assess the efficacy of psychotherapy for alleviating depression. Following the advice of Shadish et al. (2002), the researcher uses multiple measures to assess the effect of the intervention. In particular, suppose that 20 outcomes are assessed. How should the researcher proceed when 1 of these 20 shows a statistically significant intervention effect at the .05 level, but the other 19 do not? What results should be included in a manuscript submitted for publication? What results should reviewers expect to see in the manuscript, and what results should they recommend for inclusion in a published article? On the one hand, journal space is often tight, and readers may not be interested in reading about results that fail to reach conventional levels of statistical significance because the direction of any effect associated with each of these results is necessarily ambiguous as a result of the failure to reject the null hypothesis. On the other hand, there is an ethical obligation not to promote this intervention as successful if, in fact, the single statistically significant result is likely to reflect nothing more than extra effort on the part of the investigator to perform multiple tests.

Yet another factor may also explain why studies are highly likely to find statistically significant results even if they are underpowered. Realizing that the eventual choice of sample size is necessarily somewhat subjective, researchers may adopt a strategy of planning a range of possible values for the sample size. For example, in a two-group independent $t$ test (with a Type I error rate of .05, two-tailed and equal group sizes), a researcher might desire to have a power of .80 to detect an effect but cannot decide whether to specify a medium effect or a large effect. The choice has an important impact on sample size because a medium effect size implies 128 participants, but a large effect requires only 52 participants. Unable to decide between such discrepant sample sizes, the researcher might collect data on 52 participants and then perform the planned $t$ test to see whether

the sample size is large enough. In particular, if the test is statistically significant, the researcher concludes that the sample was large enough, so there is no need to collect additional data. Instead, the reports of the study can be reported based on 52 participants. However, if the test was nonsignificant, the sample size is apparently not large enough, so some number (e.g., 20) of additional participants are added to the study. At this stage another check can be conducted on the adequacy of the sample size because maybe the sample really needs to be approximately 70 participants instead of either 52 or 128. This process continues if necessary until reaching the planned maximum of 128 participants.

Strube (2006) points out that the strategy of the previous paragraph might appear to be justified because at least on the surface it fine-tunes the research protocol in such a way as to avoid subjecting more individuals than necessary to participating in the study while also reducing the risks associated with underpowered studies. Unfortunately, Strube (2006) goes on to show that not only are these apparent benefits illusory but also there is a steep price to be paid for this strategy. In particular, Type I error rates can be badly inflated, sometimes even reaching levels above .30 compared with the nominal value of .05.[1] In addition, effect size estimates can also be badly biased. As a consequence, what may have started out with the best of intentions to use scarce resources efficiently can quickly turn into a dramatic distortion of the research process.

In one sense, there is often a relatively straightforward solution to the problem of "data snooping" as reflected by interim analyses to test statistical significance. Biostatisticians have developed *early stopping* rules that allow researchers to "peek" at their data at regular predetermined intervals and at the same time maintain appropriate control of the Type I error rate. Jennison and Turnbull (1999) provide an especially useful reference to the statistical aspects of this topic.

Although the statistical issues involved in early stopping are often fairly straightforward, the problem itself is much more complicated. To the best of our knowledge, psychologists have written relatively little about the complications of interim analysis and early stopping rules. In contrast, this topic has received immense attention in medical research, presumably because data from early study participants may begin to reveal that one treatment appears either to save lives or cause deaths compared with a control or another treatment. Such indications clearly raise serious questions about the ethics of continuing to assign participants to all conditions. Indeed, this was the primary motivation for the formation of data safety and monitoring boards (Slutsky & Lavery, 2004). Psychologists could enhance ethical treatment of research participants by giving more

[1] Strube's (2006) demonstration is based on correlations instead of *t* tests, but the principle is the same in both cases.

consideration to the use of interim analysis and early stopping rules where appropriate, and, in particular, should use these formalized procedures instead of the ad hoc methods that Strube (2006) has shown will often lead to badly inflated Type I error rates and biased effect sizes.

Decisions about whether to continue a study or to stop should consider three constituencies: participants currently enrolled in the study, prospective participants who will soon be recruited into the study, and the future population of individuals who will be affected by the results of the study (Slutsky & Lavery, 2004). Although early stopping might frequently seem to be the proper thing to do from an ethical perspective, the decision often involves many ethical complexities. These complications often arise because what may be best for one of the constituencies may not be what is best for one of the other constituencies. For example, Cannista (2004) provides several examples in cancer research of trials stopped early where the lack of data regarding eventual outcomes of ultimate interest may be harmful to future patients. In a similar vein, Montori et al. (2005) find that the incidence of early stopping is increasing in medical research, but authors often fail to provide appropriate justification for their decisions, leading to serious questions about what the data analyses and accompanying conclusions really mean. For reasons such as these, Mueller, Montori, Bassler, Koenig, and Guyatt (2007) state, "Stopping a randomized trial early for apparent benefit is often unethical and can be justified only under restricted circumstances" (p. 878). Readers who are interested in learning more about the complications of such decisions may want to consult DeMets, Furberg, and Friedman (2008), who provide case studies illustrating the complex issues that often emerge in evaluating the ethics of data monitoring.

Early stopping represents one example of what Brown et al. (2009) refer to as *adaptive designs*. In general, adaptive designs present various options for balancing statistical and ethical considerations. For example, another important category of adaptive designs involves adaptive response designs. The essential idea here is that instead of always assigning an equal number of participants to each treatment condition, interim data are used to assess the relative effectiveness of the different treatments. Over time, more participants begin to be assigned to whichever treatment has shown more effectiveness up to that point in time. A classic example of this design is the "play-the-winner" design (Zelen, 1969). Perhaps the most controversial example of this design was its use to assess the effect of oxygenation on premature infants. Ware (1989) and associated comments provide a variety of arguments, both statistical and ethical, on the use of this design in this study.

Both early stopping rules and adaptive methods of assigning participants to conditions illustrate that power and accuracy depend on more than just sample size. Ethical shortcomings associated with samples

that are either too small or too large could sometimes be minimized if researchers realized that there are options for increasing power and accuracy beyond simply increasing sample size. McClelland (2000), Shadish et al. (2002), and West, Biesanz, and Pitts (2000) all provide excellent advice about various approaches for increasing power or accuracy over and above increasing sample size.

## Addressing Inadequate Resources

One of the strongest arguments in defense of the position that studies with inadequate sample sizes can nevertheless still be ethical is that each individual study can eventually contribute to a synthesis of the literature via a meta-analysis. We will defer any detailed consideration of meta-analysis to Cooper and Dent's Chapter 16, this volume. However, we do briefly want to describe the possible benefits but also complications of meta-analysis with regard to ethical issues of studies with inadequate sample sizes.

Individual researchers may often discover that they do not have sufficient resources to conduct studies with adequate power and accuracy by themselves. For example, the total sample size necessary to achieve a power of .80 to detect a medium effect size between the means of two independent groups under the assumptions of a $t$ test is 128 (assuming equal sample sizes and a Type I error rate of .05, two-tailed). The corresponding value to detect a small effect size is 788. From the standpoint of accuracy, consider a researcher who anticipates a medium effect of .50 for Cohen's $d$, the standardized mean difference comparing two independent means. Suppose this researcher wants to be reasonably certain (e.g., a probability of .80) that the true effect size is really closer to medium than it is to large or small. In other words, this researcher wants to have 80% assurance that the half-width of the confidence interval he or she forms will be no wider than .15. It can be shown that the sample size necessary to achieve this goal is slightly more than 700 (Kelley & Rausch, 2006, Table 2). Of course, it could be argued that an interval with a half-width as wide as .15 is not really very precise. After all, the total width of such an interval is .30, and thus stretches as wide as the adjacent differences between Cohen's designations of small ($d = .20$), medium ($d = .50$), and large ($d = .80$). Coming to the realization that such a wide interval is less precise than should be expected of a scientific finding, suppose our intrepid researcher adjusts his or her expectations and decides that the half-width should be no more than .05. The resulting sample size turns out to be more than 3,000 per group, implying a total sample size of more than 6,000! Thus,

even for questions so simple that they involve only a $t$ test, sample sizes can stretch well beyond the resources of an individual researcher.

Researchers whose questions involve correlations will quickly discover that they are not immune to the same problems of statistical goals exceeding typical resources. For example, consider a researcher who plans to use multiple regression to identify which of five predictor variables is uniquely associated with a dependent variable of interest. Green (1991) and Maxwell (2000) argue against the use of rules of thumb for sample size planning in regression and suggest that historical rules of thumb such as a ratio of 10 participants per variable have been woefully misguided. Although Maxwell recommends that researchers calculate appropriate sample sizes based on the specific characteristics of their study, he also suggests that for researchers who insist on falling back on rules of thumb, a reasonable starting point for a multiple regression analysis with five predictors is a sample size of something over 400. His justification is that samples smaller than this will often allow researchers to discover that something is statistically significant, but the power associated with any specific test may very well be much closer to .20 than to .80. Some researchers might not find this suggested sample size to be daunting, but we suspect many other researchers will feel compelled to work with smaller sample sizes and deal with whatever consequences arise.

What about a researcher whose goal is to estimate a correlation coefficient instead of test a correlational hypothesis? The standard error of the sample correlation depends on the value of the population correlation coefficient, but for small values of the population correlation, the standard error is approximately $\sqrt{1/n}$. Suppose a researcher wants to pinpoint the true population value of the correlation coefficient to within ±.05. A 95% confidence interval for the correlation needs to be based on roughly 1,500 cases for the interval to have an expected half-width of .05 unless the correlation itself is sizable (a large correlation of .50 according to Cohen's [1988] conventions would still require more than 850 cases).

The point of this presentation of various sample sizes necessary to achieve adequate power or accuracy is simply that most psychologists have become comfortable designing studies with smaller sample sizes than would be required to have adequate power or accuracy according to commonly accepted guidelines. Many researchers would likely be shocked to discover how large their samples would need to be to have adequate power or accuracy because in many cases the necessary sample sizes will undoubtedly exceed the resources of a single researcher. It seems unfair to label any researcher who conducts a study with a sample size below these requirements as unethical. Instead, it seems more fruitful to consider possible solutions and practices that appear to solve the problem but may or may not succeed in practice at softening any ethical concerns.

Two main solutions exist for dealing with situations where the necessary sample size far exceeds the resources of a single investigator. One option is a multisite study. Such studies are common in medical research but relatively unusual in psychology. Kelley and Rausch (2006) point out that "The idea of such *multisite studies* is to spread the burden but reap the benefits of estimates that are accurate and/or statistically significant" (p. 375). At the same time, it may be useful for individual researchers to conduct studies by themselves in the initial stages of research, instead of instantly jumping into a multisite study. As Kraemer and Robinson (2005) point out, it is important to "prevent premature multicenter RCTs [randomized clinical trials] that may waste limited funding, investigator time and resources, and burden participants for little yield" (p. 528). They discuss various aspects of sample size planning in multicenter studies and also provide a model of the respective roles of multicenter studies and individual studies in creating a cumulative science.

Meta-analysis provides another mechanism for addressing situations where resources available to individual researchers are inadequate. Even studies with inadequate sample sizes can still make a useful contribution if they are viewed as a component of a meta-analysis. However, the fly in the ointment of this argument is that studies with inadequate sample sizes may not have ever been published and instead may be languishing in the researcher's file drawer. If such studies cannot be accessed and included in a meta-analysis, the absence of such studies will generally create bias because the studies in the file drawer are likely to have smaller effect sizes than the published studies available to the meta-analyst. Methodologists continue to develop new methods for assessing the presence of publication bias (i.e., the "file drawer" effect) and correcting for it in meta-analytic estimates, but there is disagreement among experts about how well such methods are likely to work in practice.

Kraemer, Gardner, Brooks, and Yesavage (1998) have recommended that the "file drawer" problem be addressed by excluding studies with inadequate a priori power from meta-analyses. They show mathematically that removing such studies from meta-analyses can substantially reduce the bias that would otherwise exist in a meta-analysis that included such studies but failed to include studies that were unpublished. They acknowledge that in theory an even better solution would be to include all the studies, unpublished and published. In practice, however, it may be difficult to obtain even a reasonable subset of the unpublished studies. In this case, excluding underpowered studies may lead to less bias.

Adopting the recommendation put forth by Kraemer et al. (1998) to exclude underpowered studies from meta-analyses clearly undermines one of the main arguments that underpowered studies are still valuable. If only sufficiently powered studies are included in meta-analyses, underpowered studies no longer have a role to play in developing a cumulative

science. Not surprisingly, Kraemer et al.'s recommendation has been controversial, but their work at least calls into question this specific justification of studies with small sample sizes. More generally, Sutton (2009) states that:

> Whatever approach is taken, the fact remains that publication bias is a difficult problem to deal with because the mechanisms causing the bias are usually unknown, and the merit of any method to address it depends on how close the assumptions the method makes are to the truth. (Sutton, 2009, p. 448)

Another promising development to deal with problems of underpowered studies and publication bias is the emergence of study registries. Chan et al. (2004) recommend that studies should be registered before the execution of a study. Consistent with this recommendation, the member journals of the International Committee of Medical Journal Editors adopted a policy in 2004 requiring registration of all clinical trials in a public trials registry as a condition of consideration for publication. Ideally such a policy makes all studies relevant to a topic available for subsequent meta-analysis. Even though some of the studies will never be published, data from these studies can nevertheless be accessed and included in any subsequent meta-analyses.

## Conclusion

Any judgment regarding the relationship between sample size planning and ethics is inherently complicated because ethical judgments themselves are often complicated. Thus, we are hesitant to make any sweeping generalizations. However, our view is that pressures to obtain statistically significant results frequently place researchers in an ethical dilemma. Resources may often be insufficient to achieve adequate statistical power for specific hypothesis tests of interest, thus leading to other strategies such as performing multiple tests, selective reporting, the sample size samba, and "HARKing." All these can be problematic from an ethical perspective because they may fill the literature with misleading results and may therefore violate basic ethical principles of risk/reward ratio to study participants. We have discussed several possible ways of addressing this problem, such as the development of study registries, increased emphasis on multisite studies, and interim analyses based on early stopping rules. An especially positive movement is the development of *reporting standards*, such as CONSORT (Moher, Schulz, & Altman, 2001) and JARS (APA Publications and Communications Board Working Group, 2008).

Another critical factor in the relationship between sample size planning and ethics is the failure to distinguish between studies whose purpose is to determine directionality versus magnitude. Sample size should be based on statistical power considerations in the former case, but sample size should proceed differently in the latter case, where accuracy in parameter estimation is the goal. We believe that failure to consider this distinction has led to many of the ethical difficulties associated with sample size planning. Finally, we want to emphasize that our focus has been specifically on ethical implications of sample size planning, although this is obviously only one of many components involved in addressing ethical aspects of research.

## References

Altman, D. G. (1980). Statistics and ethics in medical research, III: How large a sample? *British Medical Journal, 281,* 1336–1338.

American Psychological Association. (2010). *Publication manual of the American Psychological Association*. Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Bacchetti, P., Wolf, L. E., Segal, M. R., & McCulloch, C. E. (2005). Ethics and sample size. *American Journal of Epidemiology, 161,* 105–110.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal Clinical and Experimental Neuropsychology, 23,* 399–406.

Borenstein, M., Rothstein, H., & Cohen, J. (1997). *Power and precision: A computer program for statistical power analysis and confidence intervals*. Teaneck, NJ: Biostat.

Brown, C., Ten Have, T., Jo, B., Dagne, G., Wyman, P., Muthén, B., & Gibbons, R. (2009). Adaptive designs for randomized trials in public health. *American Journal of Public Health, 30,* 1–25.

Cannista, S. A. (2004). The ethics of early stopping rules: Who is protecting whom? *Journal of Clinical Oncology, 22,* 1542–1545.

Chan, A., Hròbjartsson, A., Haahr, M.T., Gotzsche, P.C., & Altman, D.G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols for published articles. *Journal of the American Medical Association, 2291,* 2457–2465.

Clark-Carter, D. (1997). The account taken of statistical power in research. *British Medical Journal, 88,* 71–83.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.
Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997–1003.
DeMets, D. L., Furberg, C. D., & Friedman, L. M. (2008). Data monitoring in clinical trials. *Journal of the Royal Statistical Society: Series A, 170,* 504–505.
Dickinson, K., Bunn, F., Wentz, R., Edwards, P., & Roberts, I. (2000). Size and quality of randomized controlled trials in head injury: Review of published studies. *British Medical Journal, 320,* 1308–1311.
Edwards, S. J. L., Lilford, R. J., Braunholtz, D., & Jackson, J. (1997). Why "underpowered" trials are not necessarily unethical. *Lancet, 350,* 804–807.
Elashoff, J. D. (1999). *NQuery Advisor version 3.0 user's guide*. Los Angeles: Statistical Solutions Ltd.
Emanuel, J., Wendler, M. D., & Grady, C. (2000). What makes clinical research ethical? *Journal of the American Medical Association, 283,* 2701–2711.
Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203–210.
Freedman, K. B., & Bernstein, J. (1999). Sample size and statistical power in clinical orthopaedic research. *The Journal of Bone and Joint Surgery, 81-A,* 1454–1460.
Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121,* 200–206.
Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499–510.
Guenther, W. C. (1965). *Concepts of statistical inference.* New York: McGraw-Hill.
Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association, 288,* 358–362.
Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2005). RE: "Ethics and sample size." *American Journal of Epidemiology, 162,* 195–196.
Hintze, J. L. (1996). *PASS 6.0 user's guide*. Kaysville, UT: NCSS.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.
Jennison, C., & Turnbull, B. W. (1999). Adaptive and nonadaptive group sequential tests. *Biometrika, 93,* 1–21.
Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics, 59,* 580–590.
Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5,* 411–414.
Keen, H. I., Pile, K., & Hill, C. L. (2005). The prevalence of underpowered randomized clinical trials in rheumatology. *Journal of Rheumatology, 32,* 2083–2088.
Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research, 43,* 524–555.
Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321.
Kelley, K., & Maxwell, S. E. (2008). Delineating the average rate of change in longitudinal models. *Journal of Educational and Behavioral Statistics, 33,* 307–332.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the Health Professions, 26,* 258–287.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217.

Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36,* 212–219.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Kraemer, H. C., & Robinson, T. N. (2005). Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? *Contemporary Clinical Trials, 26,* 518–529.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Lilford, R. J. (2002). The ethics of underpowered clinical trials. *Journal of the American Medical Association, 288,* 358–362.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Mace, A. E. (1964). *Sample-size determination*. New York: Reinhold Publishing Group.

Maggard, M. A., O'Connell, J. B., Liu, J. H., Etzioni, D. A., & Ko, C. Y. (2003). Sample size calculations in surgery: Are they done correctly? *Surgery, 134,* 275–279.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5,* 434–458.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development, 71,* 173–180.

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. *American Psychologist, 55,* 963–964.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine, 329,* 1196–1199.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet, 357,* 1191–1194.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49,* 103–120.

Montori, V. M., Devereaux, P. J., Adhikari, N. K. J., Burns, K. E. A., Eggert, C. H., Briel, M., & Guyatt, G. H. (2005). Randomized trials stopped early for benefit: a systematic review. *Journal of the American Medical Association, 294,* 2203–2209.

Mueller, P. S., Montori, V. M., Bassler, D., Koenig, B. A., & Guyatt, G. H. (2007). Ethical issues in stopping randomized trials early because of apparent benefit. *Annals of Internal Medicine, 146*, 878–881.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Erlbaum.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm

Newell, D. J. (1978). Type II errors and ethics. *British Medical Journal, 4*, 1789.

O'Brien, R. G. (1998). *A tour of UnifyPow: A SAS module/macro for sample-size analysis. Proceedings of the 23rd SAS Users Group International Conference*. Cary, NC: SAS Institute.

Pan, Z., & Kupper, L. L. (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine, 18*, 1475–1488.

Prentice, R. (2005). Invited commentary: Ethics and sample size–another view. *American Journal of Epidemiology, 161*, 111–112.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160–164.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosoff, P. M. (2004). Can underpowered clinical trials be justified? *IRB: Ethics and Human Research, 26*, 16–19.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomized trials: Mandatory and mystical. *Lancet, 365*, 1348–1353.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Slutsky, A. S., & Lavery, J. V. (2004). Data safety and monitoring boards. *New England Journal of Medicine, 350*, 1143–1147.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Sternberg, R. J., & Williams, W. W. (1997). Does the graduate record examination predict meaningful success in the graduate training of psychology? A case study. *American Psychology, 52*, 630–641.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods, 38*, 24–27.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.

Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statistical Science, 4,* 298–306.

Weaver, C. S., Leonardi-Bee, J., Bath-Hextall, F. J., & Bath, P. M. W. (2004). Sample size calculations in acute stroke trials: A systematic review of their reporting, characteristics, and relationship with outcome. *Stroke, 35,* 1216–1224.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York: Cambridge University Press.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association, 64,* 131–146.

# 7

# *Ethics and the Conduct of Randomized Experiments and Quasi-Experiments in Field Settings*

**Melvin M. Mark**
*The Pennsylvania State University*

**Aurora L. Lenz-Watson**
*The Pennsylvania State University*

In 1995, the Administration on Children, Youth and Families (ACYF) implemented the Early Head Start program at sites across the United States. Essentially a younger sibling of the long-standing Head Start program, Early Head Start was created with the primary goal of enhancing the health and development of younger children from low-income families through the provision of services to low-income families with pregnant women, infants, and toddlers, and through the training of service deliverers. One aspect of the rollout of Early Head Start was an experimental evaluation of its implementation and effectiveness. Eligible families and children from 17 communities were randomly assigned either to participate or not participate in the local Early Head Start offerings. This research design allowed the researchers to estimate the effect of participating in the program. Findings from this Early Head Start evaluation were generally positive. In particular, children who were assigned to participate in Early Head Start had higher levels of cognitive and social–emotional development and displayed a larger vocabulary than their comparison group peers (Mathematica Policy Research, 2002).

From one perspective, the Early Head Start evaluation can be viewed as a clear social good, in that it provides information that has the potential to enlighten important democratic deliberations. A study of this sort provides strong evidence of program impact (or lack thereof). When programs are shown to have positive outcomes, this evidence can be used in support of efforts to continue and expand the program. For example, evidence from a study like the Early Head Start evaluation can be cited in

legislative debates about program funding. In contrast, if the program is found to be ineffective or harmful, decision-making processes are again informed, presumably pointing to the need to revise the intervention or find other solutions. Whatever the results, useful information is injected into deliberative processes. However, from another vantage point, serious ethical questions can be raised. To test the effectiveness of Early Head Start, by design some children were randomly assigned to receive no Early Head Start services, and the findings indicate that these children in the comparison group were disadvantaged relative to the Early Head Start participants in terms of cognitive and social development and vocabulary. Is it ethical that such differences in children's performance—which could have consequences for longer-term developmental trajectories—arise as a direct result of a research study? Do the potential benefits of the study offset the withholding of potentially beneficial services at random?

In this chapter, we examine such ethical considerations as they arise with *randomized experiments* and *quasi-experiments* in field settings. Because randomized experiments often receive more criticism on ethical grounds, we address these studies more than their quasi-experiments cousins. In the next section, we define randomized experiments and quasi-experiments and examine the rationale for their use. This rationale is important because it is related to the argument that advocates of experiments provide in response to the most common ethical criticism. In a subsequent section, we discuss randomized experiments and quasi-experiments in field settings in terms of ethical considerations, reviewing both an ethical argument for and ethical arguments against such studies. We address ethical challenges in part by considering how contemporary methodological developments and practices can ameliorate ethical concerns that have been raised. Finally, we explore two issues related to ethics that we believe deserve future attention. By way of preview, a theme that emerges is that methodological quality is not simply a technical consideration, but rather can have important implications for ethics. Throughout the chapter, we return to the Early Head Start example and occasionally refer to other evaluations of social programs. However, the discussion applies to other applications of experiments and quasi-experiments in field settings as well.

## Randomized Experiments and Quasi-Experiments: A Primer

Randomized experiments and quasi-experiments are tools that can help answer a particular type of causal question (see Pearl, Chapter 15, this volume, for further discussion on establishing causality). In particular, they are relevant when one wants to know whether and to what extent

a potential causal variable makes a difference in one or more outcomes of interest. For example, in the Early Head Start example, policy makers and others were interested in whether participation in Early Head Start (a potential causal variable) makes a difference in children's school readiness and other specific measures (the outcomes of interest).

In randomized experiments, more than one "treatment" is administered. Put differently, individuals are in different "conditions" of the experiment. Sometimes the experiment compares one named treatment (e.g., Early Head Start) with a control or comparison group that receives no explicit treatment, or perhaps a placebo, or "treatment as usual" (i.e., whatever happens naturally in the absence of individuals being assigned to the named treatment). In other studies, multiple named treatments are compared (e.g., participants could be assigned to either Early Head Start or to a package of 15 home visits by a social worker). Historically, individuals are assigned to conditions, but assignment can instead take place with other units, such as classrooms, workgroups, or even communities. Which condition a given unit is in is determined by random process, such as the flip of a fair coin, the use of a random number table, or a computer program's random generator. In the case of Early Head Start, prekindergarten children were randomly assigned either to a condition in which they were enrolled in a Early Head Start program or to a treatment-as-usual comparison condition in which they were not enrolled in a Early Head Start but instead received whatever care their family provided or arranged.

Quasi-experiments are similar to randomized experiments in the sense that they compare how different treatment conditions perform on the outcome(s) of interest. Quasi-experiments differ from randomized experiments, however, in that they do not involve the random assignment of experimental units to treatment conditions. Instead, they incorporate various types of comparisons across conditions, across time, and perhaps across different kinds of outcomes and participants. Quasi-experiments often also incorporate statistical adjustments intended to correct for biases that can result from the absence of random assignment. Quasi-experiments are an option when either ethical concerns or pragmatic reasons prevent randomized assignment, but the question of a treatment's effect on outcomes is of interest (Cook & Campbell, 1979).

As noted previously, the Early Head Start evaluation involved a randomized experiment. The key benefit of a randomized experiment, using the language of Campbell and his colleagues (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), is that it minimizes *internal validity threats*. Internal validity threats are factors other than the treatment variable of interest that could influence the outcome(s). If the Early Head Start study had not used random assignment, for example, perhaps different kinds of families would have enrolled their children

in the program, relative to those families that did not. For example, the families that entered their children into Early Head Start might have been more interested in education, or they might have been more committed to their children's development in general. Or perhaps these families were more likely to have a working parent (and thereby also better off financially), or they were better connected socially and so more likely to be aware of opportunities such as Early Head Start, or less likely to be facing life challenges that hinder their ability to get the child to the Early Head Start Center. Without random assignment these or other factors might affect which children go into which condition. Moreover, the same factors might influence the outcomes measured in the study. For example, families with a greater interest in education might tend to enroll their children in Early Head Start, and the family's interest in education might also lead to greater school readiness (apart from any effect of the program). This would be an example of the internal validity threat of *selection*. Selection occurs when a preexisting difference between the treatment groups affects the outcome variable such that the true treatment effect is obscured. Random assignment renders selection statistically implausible. If children are randomly assigned to conditions, the statistical expectation is that no preexisting factors will be systematically related to condition. Only random differences between the groups are expected, and this possibility is addressed through traditional hypothesis testing statistics (Boruch, 1997).

The strengths of the randomized experiment can be contrasted with the potential strengths and weaknesses of quasi-experiments. In fact, there are numerous quasi-experimental designs, ranging from a few that are close to the randomized experiment in terms of internal validity, to ones far more susceptible to internal validity threats. We will describe a strong quasi-experimental design later, as a potential alternative to randomized experiments that may satisfy certain ethical objections. Here we review one quasi-experimental design that is typically relatively weak in terms of internal validity, the "one-group, pretest–posttest design." In this quasi-experiment design, participants are measured on the outcome of interest both before and after receipt of the treatment. Pretest and posttest scores are compared in an effort to determine the effectiveness of the treatment. For example, if children scored better after Early Head Start than before, one might be tempted to conclude that the program was effective. However, several internal validity threats other than selection would commonly apply (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002). In this hypothetical Early Head Start quasi-experiment, the threat of *maturation* would almost certainly apply. Maturation operates when the true treatment effect is obscured because of naturally occurring changes that occur in participants over time. Because children normally improve in terms of social and academic development between, say, the age of 1 and 3 years, seeing improvement from pretest to posttest would not necessarily

suggest that Early Head Start is effective. Because maturation and other internal validity threats (including history, testing, instrumentation, and statistical regression) are frequently plausible when a one-group, pretest–posttest design is used, it is generally not a good choice for research field settings (although exceptions exist, as noted later).

## Randomized Experiments and Quasi-Experiments in Field Settings: Key Ethical Issues

Many of the ethical considerations that apply to randomized experiments and quasi-experiments in field settings are common to other forms of social research. Not surprisingly then, thoughtful discussions of ethical guidelines for experimental methods (e.g., Boruch, 1997; Shadish et al., 2002) typically draw in part on general statements about research ethics, such as the *Belmont Report* (Department of Health, Education, and Welfare, 1978). The *Belmont Report* emphasizes three principles for the ethical conduct of research: beneficence, respect for participants, and justice. In practice, these three principles translate into relatively familiar practices in research ethics. In general, prospective participants should voluntarily provide informed consent before participation, where consent includes clear information about the nature of the study and its risks and benefits; fair remuneration for participation can be given but care is needed to avoid having incentives become coercive; potential risks to participants should be minimized; efforts should be taken to maximize the study's benefits for the participant; and more generally, participants' privacy should be respected, typically including confidentiality for any information gathered, especially sensitive information. Because these topics are discussed in some detail in other chapters of this volume (e.g., Gardenier, Chapter 2; Leviton, Chapter 9), we focus here on topics that apply primarily to randomized experiments and their quasi-experimental cousins.

## An Ethical Argument for Randomized Experiments

Notably, a general argument has been put forward that ethical considerations support the conduct of randomized experiments in applied research. In sum, the argument is that (a) there is a compelling need to know about the effectiveness of various treatments, and (b) the randomized experiment is especially beneficial in addressing this need. The

presumed benefit of randomized experiments is typically framed in terms of their providing the most trustworthy information about the effects of treatments, but sometimes comes in the form of a belief that findings from randomized experiments may be more influential on subsequent actions. We focus initially on the more common form of the argument, that the need for information about treatment effectiveness is best met by randomized experiments.

The basic argument was articulated decades ago by early advocates of the use of randomized trials in medicine (e.g., Gilbert, McPeak, & Mosteller, 1977). Regarding the first part of this argument, about the need for treatment effect information, it seems clear that there is a need to know whether a new treatment for lung cancer is effective relative to current best practices, or whether stents or bypass surgery are more effective for a particular type of cardiovascular blockage. Without good evidence, uncertainty prevails about the best course of action. Or, even worse, historical happenstance or persuasive advocacy by an ostensible expert, combined perhaps with anecdotal evidence, can result in a particular treatment being widely used—even though it may be ineffective or even harmful.

This need to know about effective interventions is not limited to medicine (Henry, 2009). For example, Gersten and Hitchcock (2009, p. 82) summarize an argument for randomized experiments in education "so that professional educators would have an evidence base for making decisions about selecting curricula, intervention programs for students with learning problems, selecting professional development approaches and determining effective structures for establishing tutoring programs." In the criminal justice domain, Farrington and Welsh (2005) indicate "there is a moral imperative for randomized experiments in crime and justice because of our professional obligation to provide valid answers to questions about the effectiveness of interventions" (p. 31). Similar assertions can and have been made about many other areas of practice that are studied by evaluators and applied social researchers.

A second general claim underlying the ethical argument for randomized experiments is that this method has value relative to other ways of addressing the causal question. The most common form of this argument, made by Gilbert et al. (1977) and others, involves the assertion that randomized experiments are the preferred method for obtaining an unbiased estimate of the effect of a treatment of interest. The argument that randomized experiments are needed to get the right answer typically draws on the notion that internal validity threats are more likely to apply to findings from other methods, as discussed previously (e.g., Gilbert et al., 1977; Cook, 2002). Sometimes this argument for randomized experiments includes an empirical component, showing that randomized experiments in certain areas provide different results than those obtained from quasi-experimental or other types of studies (e.g., Boruch, 1997; Cook, 2002; Mark

& Reichardt, 2009). In addition, advocates of randomized experiments may highlight the cost of incorrect conclusions, which are presumably more likely with other methods. For example, inaccurate findings can result in an ineffective (or even harmful) intervention being adopted widely, as well as in opportunity costs in terms of other potentially effective interventions not being considered; alternatively, inaccurate findings can lead to an effective program being dropped or needlessly redesigned (Bickman & Reich, 2009).

As noted previously, there is a variant on the second portion of the ethical argument presented by Gilbert et al. (1977) and others for randomized experiments. Rather than (or in addition to) arguing for the greater validity of randomized experiments, some advocates of experiments claim that findings from this method will have greater capacity to affect subsequent action. In the case of Early Head Start, the use of a randomized experiment was mandated by Congress, at the very least suggesting that legislative attention to the study findings would be lessened if an alternative method were used. More generally, it can be argued that randomized experiments add value in the sense that they are more likely to be taken seriously in policy deliberations (Henry, 2009) or in motivating practitioners to change their behavior (Gersten & Hitchcock, 2009).

In short, arguing from ethical considerations, a case can be made that it is good to carry out experimental studies of the effectiveness of various treatments. For example, consider Bickman and Reich's (2009) claim that weaker methods are more likely to provide the wrong conclusion about whether a social program is effective and that such an error can have serious negative consequences. Might there not then be an ethical mandate for researchers to try to minimize such risks? Or consider the ethical principles from the *Belmont Report*. Both beneficence (the maximization of good outcomes and the minimization of risk) and justice (the fair distribution of risks and benefits) would seem to be reduced if researchers use biased methods that lead to the wrong conclusion about treatment effects. The greater likelihood of harm from weaker methods, as well as the accompanying reduction of beneficence and justice, is greater if one included implications, not simply for study participants, but also future potential clients. For example, imagine that a weaker quasi-experiment found that Early Head Start was ineffective, but this finding resulted from selection bias. Under this scenario, children who could have benefited from the program may be relatively disadvantaged for years to come, reducing the good outcomes and attenuating the benefits that otherwise could have arisen from the study. As this example suggests, a case can be made that ethics supports the use of research methods that will give the best answer about program effectiveness because having such an answer can increase the likelihood of good outcomes, especially for those initially disadvantaged.

Randomized experiments have received more attention than quasi-experiments in terms of ethicality. In part, this is because random assignment is more intrusive. It is because randomized experiments, by definition, determine at random which treatment participants receive—and thus potentially affect important outcomes such as school readiness—that such studies are more of a target for criticism. In contrast, in quasi-experiments the force(s) that determine each participant's condition, such as self-selection or the happenstance of which program is offered in one's community, often seem more natural.

## Ethical Criticisms of Randomized Experiments, and Responses to Them

Numerous criticisms have been made regarding the use of randomized experiments and (to a lesser extent) quasi-experiments in field settings. Some of these criticisms are explicitly framed as ethical, whereas others have an implicit ethical component. In this section, we review several ethically laden critiques of randomized experiments, as well as the responses to these critiques. The ethical challenges are organized here in relation to five criteria promulgated by the *Federal Judicial Center* (1981) *Advisory Committee on Experimentation in the Law*. In short, the five criteria are that (a) the proposed study needs to address an important problem; (b) real uncertainty must exist about what the best course of action is; (c) alternatives to an experiment should be expected to be less beneficial than the experiment; (d) it should be plausible that study results will have influence, such as by informing possible changes in policy or practice; and (e) the experiment should respect participants' rights, for example, by not being coercive and maximizing benefits for participants in the experiment. From one vantage, when met in practice these five criteria can be seen as a more detailed elaboration of the ethical argument in support of conducting a randomized experiment in a field setting (Boruch, 2005; Shadish et al., 2002). Here, they help organize the discussion of ethical critiques and responses to them.

The first criterion of the Federal Judicial Center (1981) is that the proposed study addresses an important problem, that is, that the study addresses something in society that needs to be improved. A corresponding form of criticism is that, either in general or in the particular case, the question of treatment effectiveness is not of sufficient interest. Sometimes such criticisms are intertwined with concerns about whose interests are being represented in research, with this concern typically framed in terms of the interests of those who already have power versus those who are more

disadvantaged. Greene (2009, p. 157), for example, claims that "questions about the causal effects of social interventions are characteristically those of policy and decision makers, while other stakeholders have other legitimate and important questions.… This privileging of the interests of the elite in evaluation and research is radically undemocratic" and, one might surmise from Greene's comments, ethically problematic (see also Leviton, Chapter 9, this volume). The alternative position, as suggested previously, is that good answers to the question of program effectiveness may be important for potential program beneficiaries.

The second criterion from the Federal Judicial Center is that there is real uncertainty about the best course of action. For example, if extensive evidence indicates that Early Head Start is beneficial relative to treatment as usual, then it would be unethical to randomly assign children to these two conditions. Of course, after the fact, if positive effects occur it is easy with hindsight to claim that it was known all along that the treatment was effective. This tendency may be even stronger because a treatment probably would not be tested without at least some sense (even if not with sufficient evidence) that it will be beneficial. Without a strong evidence base, however, the uncertainty that exists about a given intervention's effectiveness is likely to be considerably greater than some observers might presume, especially program advocates. Indeed, some reviewers of social interventions suggest that ineffectiveness is the norm (e.g., Rossi, 1987). Reviewing earlier literature on medical interventions, Chalmers (1968, p. 910) is almost poetic in suggesting that uncertainty is commonly warranted in advance of rigorous experimentation: "One only has to review the graveyard of discarded therapies to discover how many patients might have benefited from being randomly assigned to a control group." Thus, when confronted with the criticism that an experiment involved withholding an effective treatment from members of the control condition, it is important to assess whether it was clear in advance that the other condition's treatment was effective (Burtless, 2002).

Another more specific ethical criticism also falls under the umbrella of the Federal Judicial Center's second criterion. That is, in some experiments, a treatment of interest is compared not with best practice but with a placebo or some other treatment thought to be relatively ineffective. For example, a new pain reliever might be compared with a placebo rather than an effective pain reliever already on the market. This choice of a less potent comparison will increase statistical power and increase the likelihood that a significant difference will be observed. However, the results can be misleading as to the relative effectiveness of the new treatment, and there is generally less certainty about the performance of a new treatment relative to best practice than relative to a placebo. Even worse from an ethical perspective, members of the comparison group are denied access to a more effective treatment simply for the purpose of the experiment. Thus,

good ethics often argues for the use of a "best practice" comparison group. In some cases, however, no "best practice" treatment may be known, or the reality may be that any treatment thought to be beneficial would be rarely used, so a "practice as usual" condition can be justified. In the Early Head Start example, perhaps a best practice comparison could have been identified, such as assigning children to a well-funded preschool with a good teacher:child ratio. However, absent the public funding that would occur under Early Head Start, the reality is such that the ostensible best practice option would be available to only a small minority of the disadvantaged families in the study population, if any. Thus, the practice-as-usual condition provides a policy-relevant counterfactual while not denying anyone a potentially effective treatment that they may have selected in the absence of the experiment.

The third criterion from the Federal Judicial Center is that a randomized experiment is expected to be better able to answer the causal question than are alternatives. Much criticism of randomized experiments (and some of more rigorous quasi-experiment) falls under this criterion. This criticism includes claims that alternative methods suffice for assessing the effectiveness of a treatment. Such claims are not always framed in ethical terms, but they imply an ethical criticism, for example, by suggesting that a cost–benefit assessment of a proposed experiment would tilt toward an alternative method. Critics of randomized experiments sometimes point out, quite accurately, that in everyday experience experiments are not required to determine causal impact (Scriven, 2009). For example, no controlled study is needed to learn the effect of touching a red hot electric burner. (On the other hand, one could argue that such examples implicitly involve strong quasi-experimental designs, with a long time series of data with no burning of the hand before touching the red burner. Control observations from past touching of other items perhaps even include the burner when it is not red, with a special comparison observation. That is, it was the hand that touched the burner but not the other hand that was burned, etc.).

Indeed, a case can be made that quasi-experiments, even relatively weak quasi-experiments, provide acceptable evidence about treatment effects *in certain cases*. For example, Eckert (2000) argued that a simple one-group, pretest–posttest design sufficed for evaluating the effectiveness of training programs being carried out by a unit of the World Bank. Eckert considered each of the internal validity threats that can apply to the design and argued that the threats would not plausibly apply to the studies of the training programs in question. For example, for the threat of maturation, Eckert argued that the nature of the outcomes measured was such that it was implausible for naturally occurring shifts in knowledge to occur in the short time between pretest and posttest. In contrast, for many, if not most, of the issues that might be addressed by field experiments, it will not

be so easy to rule out internal validity threats in advance. To the contrary, in many contexts, such threats are plausible, which is why, unlike with the electric burner, experimental procedures are often needed. Similarly, observation or self-report can sometimes provide accurate information about a treatment's effect. In general, however, the plausibility is substantial that internal validity threats will affect such methods, at least for the kind of treatment effects that social researchers and evaluators are likely to be called on to assess.

The standard way of stating this need for randomized experiments or strong quasi-experiments is that these methods are most needed when internal validity threats are plausible. Perhaps it is useful also to try to specify when such validity threats are likely to be plausible. Put differently, under what conditions are randomized experiments and their closest quasi-experimental approximations most likely to be needed (i.e., where will alternative methods least suffice)? Beyond the obvious requirement that one is interested in the effect of a treatment on identified outcomes, there are conditions under which experiments, rather than alternatives, are likely to be most useful (Mark, 2009). First, experimental methods will be relatively more useful when people are interested in being able to detect small or modest effects. If the only effects of interest are huge, other methods may suffice. For example, if people would support Early Head Start only if it resulted in children at age 3 performing at a fifth grade level in reading and math, simpler methods would probably suffice. When the effect of interest is so big, it could probably be detected with simpler methods than a randomized experiment—and potential internal validity threats would not be plausible for such a huge expected increase in achievement (even though they might create some bias in the estimate of the precise size of the treatment effect). In contrast, when people are interested in small effects, techniques such as random assignment are needed. Given a potentially small treatment effect, plausible validity threats would not only create bias in the estimate of the size of the treatment effect, but also could lead to completely misleading findings about whether a positive treatment effect exists at all. Second, experimental methods will be more useful when the causal field is complex, that is, when (a) multiple factors affect the outcome of interest, (b) the outcome may change over time as a result of the effects of factors other than the treatment, and (c) people naturally vary on the outcome—all of which are close to standard circumstances for the kinds of phenomena examined in randomized field trials. For example, the multiplicity of factors that can affect outcomes such as vocabulary size and the other outcomes measured in the Early Head Start evaluation, along with the routine nature of changeover time and of individual differences, especially when combined with interest in effects that are not extremely large relative to existing variation, argues against claims (e.g., by Scriven, 2009) that treatment

effects can be observed directly and without methods such as the randomized experiment.

A potentially more compelling critique of the relative value of randomized experiments is based not on internal validity considerations but rather on *external validity*. Generally speaking, external validity refers to the accuracy of attempts to apply the findings of a study to persons, settings, or times other than those examined in the study. One form of this general criticism is that the conditions which allow for random assignment may be atypical, making attempts at generalization dubious (Cook & Campbell, 1979). For example, perhaps the communities that are willing to participate in a randomized experiment of Early Head Start differ systematically from most other communities and in ways that would lead to a different treatment effect than elsewhere.

A related criticism is that the experiment enables a focus on the average effect size (i.e., the treatment effect averaged across all participants), even though the relevant processes may be contingent on the specific characteristics of the individual person, the context, and the vagaries of treatment implementation (e.g., Greene, 2009). That is, randomized experiments at least need not open the "black box" to examine the process by which the treatment has its effects. For example, an article in *The Economist* (2008) notes:

> A randomized trial can prove that a remedy works, without necessarily showing why. It may not do much to illuminate the mechanism between the lever the experimenters pull and the results they measure. This makes it harder to predict how other people would respond to the remedy or how the same people would respond to an alternative. And even if the trial works on average, that does not mean it will work for any particular individual. (*The Economist*, 2008, p. 2)

Again, even when such criticisms are not framed explicitly in terms of ethics, they have an ethical dimension; for if the findings of a randomized experiment are not valuable in terms of guiding future action, then the rationale for their conduct is diminished.

There are several ways to respond to these criticisms of randomized experiments. One is to recall, as the Federal Judicial Center's criteria made explicit, that the focus should be on the *relative* ability of randomized experiments and of alternative methods to provide useful information. Thus, in assessing the appropriateness of a randomized experiment and an alternative method, it would be necessary to argue that the combined internal and external validity of the alternative equals or surpasses that of the experiment. Notably, many alternatives, such as case studies, may have merits, but these merits are not such that case studies better facilitate generalization to other sites. A second response is to review the

representativeness or the diversity of the cases within an experiment as a way of arguing that the study's findings should inform action elsewhere. For example, if faced with external validity criticisms in the case of Early Head Start, the researchers could point to the geographical and other forms of diversity across the participating sites, perhaps examining statistically the extent to which the participating children and their families are similar to potentially eligible participants nationwide. A third response would involve going beyond the bare-bones randomized experiment by (a) testing for possible moderated effects (i.e., interactions of key characteristics with the treatment variable), (b) conducting meditational tests of possible mechanisms by which the treatment effect would occur, and/or (c) more generally, using multiple and mixed methods to complement the strengths and weaknesses of the randomized experiment.

The fourth criterion specified by the Federal Judicial Center (1981) for use of a randomized experiment is that study results should, or at least plausibly could, have influence, such as by informing possible changes in policy or practice. It appears that criticisms related to this criterion are for the most part based on another of the five criteria. For example, concern about generalizability of findings, just discussed, can contribute to an argument that the finding of, say, an Early Head Start evaluation could not fruitfully inform decisions by a prospective program site or a specific family about whether to enroll their child (*The Economist*, 2008). Beyond such concerns, it is notable that the literature on the use of research findings, although demonstrating that use occurs, does not give great confidence about a prospective prediction of the use of any particular study of treatment effects (e.g., Nutley, Walter, & Davies, 2007). Thus, we think the right threshold involves there being a reasonable possibility that the study will be influential.

The fifth requirement from the Federal Judicial Center (1981) is that a prospective experiment should respect participants' rights, for example, by not being coercive and by maximizing benefits for participants in the experiment. This requirement involves many considerations, such as informed consent, that are far from unique to field experiment and so are given little attention here. One notable consideration is highlighted by this criterion's emphasis on study *participants*: At least in some cases the risks of a study are borne by study participants, whereas the benefits may accrue largely to others after the study is over. For example, assume that the Early Head Start study shows the benefits of that treatment for eligible children and also leads to increases in funding for the program. Future generations would benefit from the study, as would have children in the Early Head Start condition, but all this would do little for the children assigned to the treatment-as-usual comparison group. Gilbert et al. (1977) suggested that the benefit to future individuals matters greatly in assessing the ethical (and pragmatic) argument for randomized trials,

alluding to the debt that a current generation inevitably owes to past ones. Nevertheless, the distribution of risks and benefits to study participants, including in a condition that proves less desirable in terms of outcomes, is an important issue.

There are several ways of increasing benefits to study participants, even if they are assigned to what proves to be the less effective condition (e.g., Shadish et al., 2002). These include comparing a new treatment option with the best available treatment rather than a no-treatment comparison; offering the more effective treatment to those in the other group after the study is over, when this is practical; providing participants with benefits that leave them better off than they would have been without participating in the study (e.g., payment for participation; health services in a study of job training), even if these are unrelated to the primary outcome variable. In addition, in assessing the benefit:risk ratio for study participants, it seems appropriate to consider what treatment opportunities would have been present if the study had not been conducted. For example, the Early Head Start study created preschool opportunities for those in the treatment group that would not have existed in the absence of the experiment, and the study did not deny access to any opportunities that comparison group members could avail themselves of. In the next section's discussion of methodological quality as an ethical consideration, we return to possible techniques for minimizing risk and increasing benefits for study participants.

## Research Quality as an Ethical Issue

Methodological quality is usually seen as a technical consideration, the subject of methods courses and technical critiques in journals and conferences, but not as an ethical matter. Rosenthal concluded that there are ethical implications of methodological quality, referring to psychological research generally, "Bad science makes for bad ethics" (1994, p. 128). When research is designed to address the impact of social and educational programs, such as Early Head Start, which have the opportunity to change children's lifelong behavioral, emotional, and occupational trajectories, an even stronger argument can be made that methodological quality has an ethical import. This was implicit in the earlier discussion of the ethics of randomized experiments. Bickman and Reich (2009) point out that getting the wrong answer about a treatment's outcomes can change a study's risks and benefits dramatically. In the context of program evaluation, Mark, Eyssell, and Campbell (1999) argued that if wrong answers could result in harm to subsequent program participants and if methodological limits can increase the risk of getting the wrong answer, then methodological

shortcomings appear to be an ethical concern. Indeed, one can make a plausible argument that the ethical implications of methodological quality are greater for research with applied implications (see Cizak & Rosenberg, Chapter 8, this volume, and Leviton, Chapter 9, this volume, on the topic of high-stakes applied settings). For example, if a methodologically flawed evaluation is used, serious costs may arise for real people; in contrast, in more traditional basic arenas such as mainstream cognitive psychology, the self-corrective mechanisms within scholarly communities are likely to correct erroneous conclusions over time.

## Implications of Methodological Advances for Meeting Ethical Challenges

A corollary of the position that methodological quality is an ethical matter is worth considering: Methodological advancements may be able to attenuate ethical challenges to the conduct of randomized experiments or quasi-experiments. Indeed, quality practices—some now familiar and some still in development—may suffice for addressing ethical critiques of randomized experiments noted in the previous section.

One illustrative methodological advance concerns the now-widespread use of power analysis to identify the minimum number of persons needed to test study hypotheses, without unnecessarily exposing extra persons to a potentially harmful treatment. Consider the risk to participants in, and the potential benefits of, the Early Head Start intervention. What if 1,000 children were assigned to the comparison group, even though the study would have been able to detect a treatment effect with only 200 per condition. This would expose far more children than needed to whatever risk exists. Conversely, imagine that only 200 children were assigned per condition, even though the study would not have reasonable statistical power to detect a meaningful treatment effect without 1,000 per condition? In this case, the study would not have a reasonable chance of providing the benefit of detecting program effects and contributing to better decision making about the program. The widespread application of power analysis attenuates these problems. Power analyses can estimate the number of participants needed to observe a treatment effect of a given size. Consequently, the number of participants used can be selected in a way that allows for the benefit of meaningful findings while minimizing the likelihood that too many participants are needlessly exposed to any risk the study might have (Maxwell & Kelley, Chapter 6, this volume).

Another methodological advance that can help minimize risk in an experiment is the "stop rule" (called *adaptive sample size planning* in

Maxwell & Kelley, Chapter 6, this volume). With a stop rule, analyses are conducted at various planned points (typically with an adjustment for the error rate from conducting multiple tests). If, say, a significant treatment effect is observed, the experiment is halted. Otherwise, it continues. Stop rules are especially likely to be used in an experiment in which participants enter over time, such as an evaluation of a surgical procedure, rather than in a study in which all participants take place at the same time, such as with a new curriculum that is implemented in randomly chosen classrooms in a given school year. Even with studies in which all participants enter at the same time, a *stop rule* can be implemented if the outcome is measured repeatedly over time. That is, the stop rule could end the study as soon as a significant effect (or effect of a prespecified size) is observed, even if additional measurement waves had been planned. In a study with stop rules, it may be possible to add a delayed treatment component, whereby participants in the less effective condition receive the more effective treatment after the original study is halted. In the case of Early Head Start, these design ancillaries would have involved (a) conducting analyses at multiple points in time and, if the significant positive effects of Early Head Start were observed midway through the study, then (b) enrolling the practice-as-usual comparison group children in Early Head Start. This approach is not always feasible (e.g., a key outcome variable might not be reliable until the children are at least 3 years old, or there may not be Early Head Start spaces available for the comparison group children). When feasible, however, stop rules minimize risk for participants and, when used in conjunction with a delayed treatment feature, allow participants who had been in the less beneficial conditions group to receive the treatment method that their participation helped demonstrate is effective.

*Adaptive randomization* is another methodological advance that holds promise for reducing the ethical concerns about assigning people at random to the less effective condition in an experiment (Hu & Rosenberger, 2006). Adaptive randomization begins by assigning equal numbers of participants to each condition. However, in an adaptive randomization scheme, interim analyses are used to adjust the probability of assignment to each condition based on the apparent effectiveness to that point. For example, if the treatment group has interim outcomes that are 1.5 times as good as the outcomes in the treatment-as-usual comparison group, then the assignment probabilities would be adjusted so that 1.5 times as many participants would be assigned to the treatment group as the study continues. At this explanation suggests, adaptive randomization applies when participants enter over time. To date, it appears that adaptive randomization has been used primarily in early-phase medical trials (Hu & Rosenberger, 2006), but the technique may find its way into the toolkit of

applied social researchers for certain kinds of field experiments, such as in legal settings where cases tend to trickle in over time.

Faced with real or potential criticism about the use of a randomized experiment, another general approach is to consider methodological advances that do not involve random assignment to condition but hold promise for giving an unbiased estimate to the treatment effect. One strong alternative of this sort is the *regression–discontinuity design* (Imbens & Lemieux, 2008; Shadish et al, 2002). In this quasi-experimental design, a set of study participants is measured on a "quantitative assignment variable" (QAV), and all individuals on one side of a cutoff score on the QAV are assigned to one condition, whereas those on the other side are assigned to the other condition. In this way, a treatment can be assigned based on need, circumstances, or merit, rather than at random. For example, in the case of Early Head Start, researchers might start by identifying children and families willing to participate in the study. Then researchers might measure the QAV, such as a measure of the children's initial cognitive development (alternatively, the QAV could be a measure of families' adjusted income, or a composite based on several indicators). A cutoff score would be established, and children with scores below the cutoff would be assigned to Early Head Start, with children scoring above the cutoff assigned to the treatment-as-usual comparison condition. In essence, a treatment effect is observed when the outcome scores of the children in the treatment group are higher than would be expected based on the trend of scores in the comparison group. Put differently, if there is a discontinuity in the regression line (with the QAV predicting the outcome) at the cutoff, the only plausible explanation in most cases is that the program made a difference. The regression–discontinuity design escapes much of the ethical criticism of the randomized experiment because it assigns the treatment to those with greater need (or in some cases, greater merit). In the past, the design has been used rarely. However, the regression–discontinuity design has received increased attention, including by economists who are advancing statistical design and validity checks (e.g., Hann, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008). Thus, the design may become a more common alternative to randomized experiments.

Another design-based methodological advance relies on random assignment but does not involve direct assignment into treatment conditions. Rather, in the random encouragement design, participants are assigned at random either to receive or not receive extensive recruitment efforts encouraging them to participate in the treatment of interest. Although not yet implemented in enough field studies to be confident about its practicality, the design appears to hold promise in avoiding or minimizing ethical criticisms about withholding potentially beneficial treatments in a randomized experiment. The *random encouragement design* was implemented by Wells et al. (2000), who used multiple forms of encouragement,

including education and reduced fees, to solicit patients at a randomly assigned set of clinics to participate in a quality improvement program for the treatment of depression. Statistical methods such as instrumental variables analysis are then used in an effort to provide good estimates of the treatment effect (see Schoenbaum et al., 2002, for an illustration of instrumental variables analysis with the Wells et al. data). In essence, these analyses assume that any effect of encouragement arises only by way of the increased program participation that the encouragement creates; this assumption facilitates statistical estimates of the effect of the program itself. The random encouragement design alleviates ethical concerns that can arise from procedures that restrict participants' access to multiple treatment options. It can reduce any concerns about coercion. These ethical benefits *may* occur with little loss of validity, although further experience with the design is needed.

In short, a set of methodological practices, including recent developments, offers promise for ameliorating some ethical concerns about randomized experiments and their quasi-experimental cousins. Power analysis, adaptive randomization, and use of the "stop rule" reduce unnecessary exposure to a potentially harmful treatment. When the "stop rule" is used with a delayed treatment component, participants in less effective conditions receive treatment that was shown to be effective. Regression–discontinuity design assigns participants who are most in need (or most deserving) of treatment to be assigned to what is hypothesized to be the more effective treatment condition. The random encouragement design reduces ethical concern about the study constraining participants' ability to make choices among treatment options. Thus, the general theme that methodological quality has ethical implications can be expanded. Methodological advances can serve to resolve ethical criticisms, if they satisfactorily address the underlying ethical problem.

## Three Topics That Warrant Future Attention in Applied Research Studies and Program Evaluations

In this section, we highlight three issues that appear to deserve future attention. For the first of these, attention is needed from methodologists and statisticians. For the second, consideration is required from those involved in the design of applied research and evaluation studies, especially those involved in the selection of measures, as well as measurement specialists. For the third issue noted in this section, a variety of parties could contribute, including scholars conducting empirical research on applied research and evaluation itself, group process researchers, and the

broader community of those interested in how decisions are to be made about the ethicality of proposed research.

## Moving Further Beyond Average Effect Sizes

As noted previously, randomized experiments and quasi-experiments can be criticized for their focus on average treatment effects, which may not provide adequate guidance for action if the effects of the treatment are moderated by undetected interactions (Greene, 2009; *The Economist*, 2008). In the face of such interactions, a program might benefit some participants but have no effect or even be harmful for others. The ethical concern seems obvious: If there is a harmful effect for a subset of participants and if the experiment fails to detect this, the harmful effects will not be ameliorated; even worse, the study results could lead to the program being administered universally in the future despite its harmful effects to some participants. Even without a harmful effect, if the treatment is ineffective for some participants, the failure to detect the differential effects could have serious opportunity costs by keeping the relevant subgroup from obtaining an alternative treatment that might be beneficial for them.

One proposed but not fully developed response is to conduct "*principled discovery*" (Mark, 2003; Mark, Henry, & Julnes, 2000). Principled discovery holds potential for addressing one form of ethical criticism of randomized experiment and quasi-experiments. More generally, it could help increase the ability of studies to guide future action, in the face of moderated relationships that limit the guidance that can be taken from average treatment effects. The basic idea of subsequent principled discovery is to engage in two phases (possibly with further iteration between the two). One would begin analyses, before principled discovery, by conducting the planned analyses to test an a priori hypothesis (e.g., that Early Head Start will lead to better outcomes than treatment as usual). In the first phase of principled discovery, the researcher would then carry out exploratory analyses. For example, the Early Head Start evaluator might examine whether the program has differential effects by looking for interaction effects using one after another of the variables on which participants have been measured (e.g., gender, race, age, family composition, etc.). A wide variety of statistical techniques can be used for the exploratory analyses of this first phase of principled discovery (Julnes & Mark, 1998; Mark, 2003). The exploration of phase 1 is not without risks, however, especially the possibility of being misled by chance. Statistical significance of course simply means that a given finding is unlikely to have arisen by chance if there really were no difference. But the conduct of many exploratory tests creates a

risk that some finding will be significant because of chance. Stigler's (1987, p. 148) admonition is apt: "Beware of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confession obtained under duress may not be admissible in the court of scientific opinion" (see also Maxwell & Kelley, Chapter 6, this volume).

If the exploratory analyses of phase 1 result in an interesting discovery, the classic admonition is to try to replicate the discovery in another study. However, this will often be infeasible in the case of field studies such as program evaluations, where any use of the study findings is likely to occur before replication is possible. Thus, the second phase of principled discovery would be called for, in which the researcher seeks one or another form of independent (or quasi-independent) confirmation of the discovery. In many instances, this will involve other tests that can be carried out within the same data set (although data might be drawn from other data sets, or new data might be collected after phase 1). For example, if an interaction were observed such that Early Head Start has a smaller effect for children in families with relatively less parental education, this could lead to another prediction that a similar and probably stronger interaction will be obtained with a composite variable (drawn from home visits) based on the amount of children's books and educational material in the children's homes. As this example illustrates, phase 2 of principled discovery will generally require an interpretation of the finding from the phase 1 exploration. This interpretation in turn gives rise to the phase 2 hypothesis. The value of the phase 2 test is that, if the original discovery is not real but instead is only the result of chance, then there is generally no reason to expect the phase 2 test to be confirmed. Future application of the approach, including further investigation of techniques for controlling for error rate in the two phases of principled discovery, seems warranted.

## Changes Over Time in Value-Based Outcomes

The outcomes that should be examined in a study are not magically revealed. Moreover, the concerns and values that drive the selection of outcomes are not historically invariant. What people care about in relation to a kind of intervention can change over time, and use of outcomes that do not reflect current values can lead to a waste of participant time and research resources and to a limited potential for the study to make a difference. Of course, in applied research the key outcomes for a study often derive rather directly from the study purpose. For example, for an evaluation of Early Head Start, measures of cognitive development seem to derive naturally from the program and its goals. However, the outcomes

that people care about and their expectations about what programs will achieve are not static. As an example, measures of social development are more common today than in the early days of preschool evaluation.

The selection of outcome variables has ethical implications, even if these are indirect. For example, if sound decision making about a preschool program would require measures of both cognitive and social development, but only cognitive development is assessed, problems can occur. The benefits of the study may be curtailed. At the extreme, a study may lead to the selection of the wrong treatment, for example, if one program has a slight benefit with respect to cognitive outcomes but performs far worse on the (unmeasured) social development outcomes. In this light, the ethical import of the value-based selection of outcomes seems evident.

A contemporary example merits attention. With growing concern about global climate change, it seems possible that environmental impact may gain in importance even for programs and policies that do not have primarily an environmental focus. For example, one can imagine a future in which a program such as Early Head Start would include procedures for assessing the environmental impact of the intervention. This would include attention to such things as power use at program sites and energy use in transportation to program sites, relative to the estimated energy impact of whatever disparate arrangements individual children in the comparison group have. Future work on tracking environmental impact for social and educational programs may be fruitful, as may more general work on social methods for identifying the outcomes that people value for a given kind of program.

## Procedures for Making Judgments About the Ethicality of a Proposed Study

Throughout this chapter we have encountered a set of questions (e.g., whether the treatment effect question is important, whether a randomized experiment will provide a better or more influential answer), the answers to which help determine the ethicality of a potential randomized experiment or of its various quasi-experimental cousins. But how are these questions to be answered? Absent a compelling protocol for researchers to judge the ethicality of a proposed study, a standard part of the answer is to rely on *institutional review boards* (IRBs) to provide answers. However, perhaps the most challenging aspect of assessing the ethicality of a randomized experiment involves the question of how to go about trying to answer questions, such as the importance of the treatment effect question and the relative benefits of a randomized experiment, in a particular case.

The inclusion of community members on IRBs in part recognizes that the matters to be judged are not only technical ones. (Regulations require both a member who is not affiliated with the institution and a member who is not a scientist, although in practice it appears these are usually represented in a single community member). However, both the political or values questions (e.g., Is the question of the treatment's effects on possible outcomes sufficiently important to justify the conduct of the study?) and the more technical ones (e.g., Is a random assignment experiment needed, relative to alternative methods?) are addressed by the same group, at least some of whom may not have the requisite skills and/or information for making thoughtful judgments about both kind of questions. Moreover, the IRB typically enters into the process after a study has been fully planned, making adjustments costly and painful.

Future research may be able to inform specific judgments related to the ethicality of future research. For example, studies could assess the extent to which potential program beneficiaries or their proxies (e.g., parents of preschool children), across a range of program types, are interested in obtaining valid answers to the question of program's treatment effect, in contrast to claims of critics such as Greene (2009). Research could also assess the practicality and worth of procedures that might expand on traditional IRB procedures, such as variants on the deliberative polling methods used in recent years by political scientists. More generally, group process researchers could fruitfully apply their expertise in an effort to improve IRB (or complementary) procedures.

## Conclusion

This chapter has addressed ethical issues related to the conduct of randomized experiments and quasi-experiments in applied field settings, including program evaluation. The ethical argument for randomized experiments and their strongest quasi-experimental cousins has been reviewed. Ethical criticisms of randomized experiments, and responses to them, have been presented. We have reviewed the potential for existing and emerging methodological advances to ameliorate certain of the ethical challenges to experiments. We have also briefly considered three topics we believe deserve further attention in the future. Although the discussion has been general, in practice ethical judgments are made about specific studies, the details of which matter. Nevertheless, we hope that the presentation of the ethical arguments for and against experiments and the other topics addressed in the chapter will help in framing more thoughtful consideration of the ethics of proposed or actual randomized experiments and quasi-experiments.

## References

Bickman, L., & Reich, S. (2009). Randomized control trials: A gold standard with feet of clay. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 51–77). Thousand Oaks, CA: Sage.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.

Boruch, R. F. (2005). Comments on 'Use of randomization in the evaluation of development effectiveness.' In G. K. Pitman, O. N. Feinstein, & G. K. Ingram (Eds.), *World Bank series on evaluation and development, Vol. 7: Evaluating development effectiveness* (pp. 205–231). New Brunswick, NJ: Transaction.

Burtless, G. (2002). Randomized field trials for policy evaluation: Why not in education? In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing: When "have nots" gain but the "haves" gain even more. *American Psychologist, 60,* 149–160.

Chalmers, T. C. (1968). Prophylactic treatment of Wilson's disease. *New England Journal of Medicine, 278,* 910–911.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24,* 175–199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Skokie, IL: Rand McNally.

Department of Health, Education, and Welfare (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* Washington, DC: U.S. Government Printing Office.

Donaldson, S., Christie, T. C., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage.

Eckert, W. A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation, 21,* 185–193.

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1,* 9–38.

Federal Judicial Center. (1981). *Experimentation in the law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law*. Washington, DC: U.S. Government Printing Office.

Gersten, R., & Hitchcock, J. (2009). What is credible evidence in education? The role of the What Works Clearinghouse in informing the process. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 78–95). Thousand Oaks, CA: Sage.

Gilbert, J. P., McPeak, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. *Science, 198,* 684–689.

Greene, J. C. (2009). Evidence as "proof" and evidence as "inkling." In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 153–167). Thousand Oaks, CA: Sage.

Hann, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69,* 200–209.

Henry, G. T. (2009). When getting it right matters: The case for high quality policy and program impact evaluations. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 32–50). Thousand Oaks, CA: Sage.

Hu, F., & Rosenberger, W. F. (2006). The theory of response-adaptive randomization in clinical trials. Hoboken, NJ: Wiley Interscience.

Imbens, G. W., & Lemieux, T. (2008). Regression-discontinuity designs: A guide to practice. *Journal of Econometrics, 142,* 615–635.

Julnes, G. J., & Mark, M. M. (1998). Evaluation as sensemaking: Knowledge construction in a realist world. In G. Henry, G. W. Julnes, & M. M. Mark (Eds.), *Realist evaluation: An emerging theory in support of practice* (pp. 33–52). San Francisco: Jossey Bass.

Mark, M. M. (2003). Program evaluation. In S. A. Schinka & W. Velicer (Eds.), *Comprehensive Handbook of Psychology* (Vol. 2, pp. 323–347). New York: Wiley.

Mark, M. M. (2009). Credible evidence: Changing the terms of the debate. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 214–238). Thousand Oaks, CA: Sage.

Mark, M. M., Eyssell, K. M., & Campbell, B. J. (1999). The ethics of data collection and analysis. In J. L. Fitzpatrick & M. Morris (Eds.), *Ethical issues in program evaluation* (pp. 47–56). San Francisco: Jossey Bass.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey Bass.

Mark, M. M., & Reichardt, C. S. (2009). Quasi-experimentation. In L. Bickman & D. Rog (Eds.), *The Sage handbook of applied social research methods* (2nd ed., pp. 182–213). Thousand Oaks, CA: Sage.

Mathematica Policy Research. (2002). *Early Head Start research: Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start*. Available at http://www.mathematica-mpr.com/publications/pdfs/ehsfinalsumm.pdf

Nutley, S. M., Walter, I., & Davies, H. T. O. (2007). *Using evidence: How research can inform public services.* Bristol, UK: Policy Press.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5,* 127–134.

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy*, 4, 3–20.

Schoenbaum, M., Unutzer, J., McCaffrey, D., Duan, N., Sherbourne, C., & Wells, K. B. (2002). The effects of primary care depression treatment on patients' clinical status and employment. *Human Service Research, 37,* 1145–1158.

Scriven, R. (2009). Demythologizing causation and evidence. In S. Donaldson, T. C. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 134–152). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stigler, S. M. (1987). Testing hypotheses or fitting models: Another look at mass extinction. In M. H. Nitecki & A. Hoffman (Eds.), *Neutral models in biology* (pp. 145–149). Oxford, UK: Oxford University Press.

The Economist. (2008, December 30). The bright young thing of economics. Retrieved from http://www.economist.com/finance/displayStory.cfm?story_id=12851150

Wells, K. B., Sherbourne, C., Schoenbaum, M., Duan, N., Merideth, L., Unutzer, J., … Rubenstein, L. V. (2000). Impact of disseminating quality improvement programs for depression in managed care: A randomized controlled trial. *Journal of the American Medical Association, 238,* 212–220.

# 8

# *Psychometric Methods and High-Stakes Assessment: Contexts and Methods for Ethical Testing Practice*

**Gregory J. Cizek**
*University of North Carolina at Chapel Hill*

**Sharyn L. Rosenberg**
*American Institutes for Research*

Psychometricians routinely use quantitative tools in test development and after test administration as part of the procedures used to evaluate the quality of the information yielded by those instruments. To some degree, nearly all those procedures play a part in ensuring that tests function in ways that promote fundamental fairness for test-takers and support the ethical use of test information by those who make decisions based on test results. In this chapter, we survey some of the quantitative methods used by testing specialists to accomplish those aims. The sections of this chapter are organized along the lines of three major phases of testing: test development, test administration, and test score reporting and use.

These topics are treated within four contexts. First, we have adopted the perspective on fairness proposed by Camilli, who has stated that "While there are many aspects of fair assessment, it is generally agreed that tests should be thoughtfully developed and that the conditions of testing should be reasonable and equitable for all students" (2006, p. 221). Further, we agree with Camilli that, although "issues of fairness involve specific techniques of analysis.… many unfair test conditions may not have a clear statistical signature" (p. 221). Thus, although the focus of this Handbook is on quantitative methods, we will occasionally allude to other methods for promoting ethical testing practice.

Second, our coverage of the psychometric methods for ethical testing practice focuses on *high-stakes* tests. Not all tests are included here, or even all standardized tests—only those tests for which important positive or negative consequences are attached. And, it is most precisely the

*decisions*—based in whole or in part—that are consequential and have stake associated with them, not strictly the tests themselves. However, high-stakes situations in which test data play a central role are increasingly common in education, psychology, occupational licensure and certification, and other contexts. Examples of high-stakes testing contexts include those of making clinical diagnoses of depression, judging the effectiveness of interventions for students with autism, counseling teenagers about career options, placing college first-year students in appropriate foreign language courses, awarding or withholding a license or credential for a given occupation, selecting or promoting civil servants, and numerous other situations. The common attribute is that the high-stakes test yields information that contributes to decisions that have meaningful consequences for individual persons, groups, or organizations. In each situation, quantitative methods can be used to promote fair and ethical decisions.

Third, high-stakes tests are not new. Miyazaki (1976) reports on the testing procedures associated with Chinese civil service examinations circa 200 B.C. An emphasis on ethical assessment has not always been a central focus of the testing profession (see, e.g., Gould, 1996). Within the past 40 years, however, increasing attention has been paid to ethical issues in high-stakes testing, and numerous standards and guidelines have been promulgated to provide direction for test developers and test users. Among these resources are:

- *Rights and Responsibilities of Test Takers: Guidelines and Expectations* (Joint Committee on Testing Practices, 1998)
- *Code of Professional Responsibilities in Educational Measurement* (National Council on Measurement in Education, 1995)
- *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004)
- *Family Educational Rights and Privacy Act* (1974)
- *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999)

Of these, the *Standards for Educational and Psychological Testing* (hereafter, *Standards*) is widely considered to be the authoritative source for best testing practices in education and psychology. The *Standards* is now in its fifth edition, a series that began with the publication of *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association, 1954). In preparing this chapter, we have relied heavily on the current edition of the *Standards*, and linkages to relevant portions of the

*Standards* will be made throughout this chapter. We have also provided citations to specific portions of other resources where appropriate.

Finally, we have chosen a standards-referenced mathematics test required for high school graduation as a context for illustrating the application of quantitative methods to promote ethical testing practice. Several states require passage of these so-called "exit" tests or "end-of-course" examinations in subject areas such as mathematics, language arts, or science for students to be awarded a high school diploma. To be sure, the graduation decision does not hinge solely on the passage of such tests; rather, they are but one of multiple measures used. In all cases, other criteria (e.g., attendance, grades, specific courses requirements, community service hours, etc.) must also be satisfied. However, the test would still be classified as "high stakes" because failing to meet the performance standard on the test would have serious consequences for students.

## Test Development

Ethical concerns arise at many junctures of the test development process, and the discipline of psychometrics has produced both qualitative and quantitative methods to promote fundamental fairness during this stage. Test development refers to "the process of producing a measure of some aspects of an individual's knowledge, skill, ability, interests, attitudes, or other characteristics," and it is "guided by the stated purpose(s) of the test and the intended inferences to be made from test scores" (AERA, APA, & NCME, 1999, p. 37). According to the *Standards*, "Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development" (p. 43).

The following subsections describe six decision points in the test development process where the application of quantitative procedures can help promote fair and ethical testing practice. The specific areas to be addressed include (a) identification of test purpose and content coverage, (b) choice of psychometric model, (c) item–task construction and evaluation, (d) test form development, (e) standard setting, and (f) validation.

### Identification of Test Purpose and Content Coverage

According to the *Standards*, test development activities should be "guided by the stated purpose(s) of the test and the intended inferences to be made from test scores" (AERA, APA, & NCME, 1999, p. 37). Thus, the first step in

producing any test is to articulate a sharp focus on the construct the test is intended to measure and the test purpose(s). Construct definition and purpose may flow from theory development, clinical needs, industrial/organizational requirements, or legislative mandates. Whether in educational achievement testing or occupational testing, the first step in test development is typically to conduct a curriculum review, job analysis, role delineation study, or task survey (see Raymond & Neustel, 2006; Webb, 2006). These activities typically result in a set of *content standards*—a collection of statements that express the knowledge, skills, or abilities that are to be included in a curriculum, the focus of instruction, and assessed by an examination. Once these clusters of essential content, prerequisites, or critical job demands that will be sampled on the test have been established, the proportions or weightings for each cluster in the examination specifications must be derived, and various quantitative procedures for doing so are used (see Raymond, 1996).

In the context of a state-mandated, high-stakes exit examination in mathematics, delineating the domain to be tested and obtaining weights for subdomains are usually accomplished via judgmental procedures that seek to balance expert input, feasibility, cost, and other factors. A large and diverse panel of mathematics teachers, curriculum specialists, mathematicians, business leaders, parents, and others might be assembled to provide recommendations on decisions such as (a) the appropriate number of items or tasks for high school students to attempt; (b) the specific subareas of mathematics to be covered (e.g., algebra, geometry, probability, and statistics) and the relative proportion of the total test devoted to each of these; (c) the appropriate item formats and contexts (e.g., multiple-choice, constructed-response); (d) the acceptable level of language load of the test items or level of writing skill necessary for constructed-response items; and (e) policies for the use of calculators, and other decisions requiring knowledge of the intended test population and test content. It should be noted that, although representative panel membership is a goal of procedures to delineate domains and develop test specifications, inequities can still result (e.g., if the opinions of the mathematicians carry the most weight in panel discussions, or if practitioners in academic settings are overrepresented in job analysis survey returns). Thus, such procedures must be constantly monitored to foster equitable results.

Another ethical issue that can arise when developing test specifications is the need to ensure that the specifications reflect two characteristics. First, as in the case of the mathematics test, the instruction provided to students would need to be aligned to the test specifications. A fundamental concept in the area of test fairness is *opportunity to learn*. In this case, opportunity to learn would reflect the extent to which examinees were provided with instruction in the knowledge, skills, and abilities to be covered on the high-stakes mathematics test. Second, if the mathematics

test were used to predict success in subsequent courses or occupations, it would be necessary to collect evidence that the content specified in the test specifications was related to performance in the courses or the skills required for safe and effective practice in the occupation. Of course, at the most fundamental level, these are issues of validity, a topic addressed later in this chapter and elsewhere in this Handbook (see Carrig & Hoyle, Chapter 5, this volume).

### Choice of Psychometric Model

Many aspects of test development, scoring, reliability, and validity are affected by the psychometric model that is used. There are two general classes of models used for building tests in education and psychology: classical test theory (CTT) and item response theory (IRT). CTT posits that an examinee's observed score is composed of a true component and a random error component, with the true score defined as the examinee's average score over an infinite number of parallel forms. With CTT, examinees' observed scores most often are calculated as a raw score (i.e., number correct) or percentage of items answered correctly, although more complicated scoring rules are possible (see Crocker & Algina, 1986).

An alternative set of models, IRT models, has become more widespread over the past few decades. IRT models invoke stronger assumptions than CTT models; they are more computationally complex; and they generally require larger sample sizes for successful use. IRT models posit that an observed score is an indicator of an underlying latent trait. They provide the probability of an examinee responding correctly to an item, with that probability dependent on the examinee's latent ability and the characteristics of the test item. IRT models require specialized software to compute estimates of examinees' standing on the latent trait; the software programs vary according to the estimation procedures used (e.g., joint maximum likelihood, marginal maximum likelihood) and the characteristics of the items (e.g., difficulty, discrimination, lower asymptote) that are estimated. There are several different types of IRT models, for example, the one-parameter logistic (1-PL) model, 2-PL model, 3-PL model, partial credit model, graded response model, and others. (For an introduction to IRT models, see Hambleton & Swaminathan, 1985.)

The choice of psychometric model has ethical implications. For example, the choice of one psychometric model over another may lead to different outcomes (e.g., pass–fail decisions, performance category classifications) for examinees. The choice of a psychometric model will affect the information that is gained about uncertainty (i.e., error) in examinees' scores. For example, one of the central features of IRT is the emphasis on a conditional standard error of measurement (CSEM). Unlike the CTT

standard error of measurement, which provides an overall, constant estimate of measurement error across the entire score range for a test, the CSEM provides an indication of the precision of an ability estimate at each score point on the test scale and varies across the test score range. In high-stakes contexts such as a high school graduation test, test construction efforts can enhance fairness by maximizing precision (i.e., minimizing the CSEM) in the regions of the score scale where cut scores are located and where classification decisions are made. In CTT, item discrimination indices and, in IRT, the *a*-parameter (in 2-PL and 3-PL models) also can be used to ensure that the most discriminating items contribute the most toward examinees' scores.

It is important to note, however, that the benefits of using a CTT, IRT, or other psychometric model only accrue to the extent that the model fits the data. Using a psychometric model that does not fit the data well in some parts of the score scale (particularly in the region where decisions are made) can compromise the fairness of those decisions. At minimum, procedures for assessing model–data fit (e.g., examination of residuals, assumptions, and fit statistics) should be used during field testing or after the first operational administration of an item.

The choice of a psychometric model often is driven by a combination of technical, practical, philosophical, and political considerations. For example, a 1-PL (Rasch) model may be chosen for developing a high school graduation test even before field test data are collected. This strategy is in sharp contrast to other contexts (e.g., structural equation modeling), where accepted practice involves comparing the fit of several alternative models and choosing the one that provides the best fit to the data (see McArdle, Chapter 12, this volume). Such a decision may be guided in part by philosophical considerations (e.g., the belief that additional parameters estimated in models accounting for item characteristics beyond item difficulty are only modeling error, a classic stance taken by Rasch model proponents such as Wright, 1997), or by political considerations (e.g., a concern that it would be difficult to explain to parents why items in the test are not weighted equally, and the possibility that students with the same raw scores can be assigned to different pass–fail or performance categories). Proponents of the Rasch model assert that "The data must fit, or else better data must be found" (Wright, 1997, p. 43). Regardless of approach, it is important to consider how examinees whose response patterns do not fit the prescribed model could be adversely impacted.

In summary, the process of choosing a psychometric model differs from the fitting of a structural equation model in the psychological literature. Model choice is not an area of psychometric methodology that has received wide attention to ethical issues, but the choice of test model can carry ethical implications. It is important to evaluate the reasons for and assumptions of choosing a particular CTT or IRT model. Examinee scores,

as well as subsequent decisions based on them, are directly related to the extent to which the psychometric model is appropriate for the test data. Unsatisfied assumptions or large modeling error can pose a serious threat to the inferences made about an examinee.

### Item–Task Construction and Evaluation

The next step after defining the domain, producing test specifications, and identifying an appropriate psychometric model is the creation of items and/or tasks and scoring guides and rubrics that will comprise the test. According to the *Standards for Educational and Psychological Testing*: "The type of items, the response formats, scoring procedures and test administration procedures should be selected based on the purpose of the test, the domain to be measured, and the intended test takers" (AERA, APA, & NCME, 1999, p. 44).

The item writing and evaluation process can pose several ethical concerns. First, it is essential that item writers have adequate knowledge of the test content and are trained on the item writing process in a consistent manner. If item writers do not have this requisite knowledge, then they are likely to produce items that may compromise fairness by failing to adequately represent the intended domain. Before pilot testing, items should undergo a preliminary bias–sensitivity review where representative stakeholders evaluate items and suggest revising or eliminating any items that have the potential to disadvantage any test-takers. The *Code of Fair Testing Practices* notes that test developers should "avoid potentially offensive content or language when developing test questions and related materials" (Joint Committee on Testing Practices, 2004, p. 4). The *Standards* requires that "To the extent possible, test content should be chosen to ensure that intended test scores are equally valid for members of different groups of test takers" and "The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats" (AERA, APA, & NCME, 1999, p. 44).

Pilot and field testing is an essential part of the measurement process and can help mitigate concerns related to test fairness. Because items are often selected for operational use based on their qualities in item tryouts, it is important that examinee samples used in this process are as large and representative as possible. Technically, IRT does not require that the pilot or field test groups be representative samples as long as they are sufficiently large and include the full range of performance in the intended population. However, given the potential for differential item functioning to occur—a sure threat to test fairness—it is desirable that pilot and field test samples are as representative as possible. Otherwise, items that appear to function well in a pilot or field test may have less validity evidence to

support their operational use in the intended population. As indicated in the *Standards*:

> When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended. (AERA, APA, & NCME, 1999, p. 44)

Likewise, the *Code of Fair Testing Practices* indicates that test developers should "obtain and provide evidence on the performance of test-takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses [and] evaluate the evidence to ensure that differences in performance are related to the skills being assessed" (Joint Committee on Testing Practices, 2004, p. 4).

It is also important that the testing conditions for a pilot or field test should be as close as possible to those of operational test administrations. If pilot or field tests are conducted as stand-alone procedures, steps should be taken to investigate and document any conditions that may affect examinee behavior and subsequent performance. For example, if items for a high school graduation test are pilot tested as a stand-alone exercise that has no consequences, low motivation is likely to affect the students' performance. This could compromise the accuracy of test results and test fairness because the item statistics generated from the pilot and field tests are typically used to select the items for the operational test that will be used for the high-stakes decisions.

To address the concern about the accuracy of item statistics from pilot or field tests, it is usually preferable to use embedded field testing procedures (i.e., where the trial items are interspersed with operational items and examinees have no knowledge of which items count toward their score). This way, testing conditions are similar to the operational administration conditions and are therefore less likely to adversely impact the results of the pilot or field tests.

There are several key purposes of item tryouts. First, pilot or field testing data can be analyzed to select the items with the best qualities that are most likely to represent the test content and minimize the potential for unfairness. Second, to maximize the precision of measurement of a test to which a cut score will be applied, it is desirable to select items that are highly discriminating in the range where a decision is made (i.e., in the area of any cut score). Third, differential item functioning (DIF) analyses can be performed to determine whether there are differences in performance on individual items where focal and reference group abilities are equivalent (Camilli, 2006). Items that are flagged for displaying statistically significant DIF routinely undergo additional review to determine

whether they should be included on an operational test, or they may simply be eliminated due to any potential for unfairness.

Finally, item tryouts permit the evaluation of scoring guides or rubrics for performance tasks or constructed-response items. Analyses are performed to ensure that each score category is functioning as intended, that the boundaries between score categories are clear, that the rubric or scoring guide can be interpreted as intended and applied consistently by any raters, and to permit adjustments to the scoring procedures when unanticipated examinee responses shed light on gaps in the category descriptions.

### Test Form Development

After the development and review of test items and tasks, test forms are created. At this juncture, ethical issues also must be addressed. If multiple forms will be developed for a single test administration, or if new forms are developed across test administrations, the forms must be developed according to the same content and statistical specifications, including targets for difficulty and reliability. Failure to develop equivalent forms can reduce confidence that equating (described later in this chapter) will correct for variations in difficulty, or that examinees' scores on different forms can be interpreted in the same way.

An additional method for promoting consistency in form development procedures and match to test specifications is found in *alignment analyses*. Alignment analyses include both judgmental review and quantitative indices of the degree to which a test matches the content standards it was intended to assess. According to Porter (2006), there are two general ways that a test may be imperfectly aligned with its content standards: (a) Some areas specified in the content standards are not measured by a test; or (b) some areas assessed on a test are not part of the content standards. The latter condition—that is, when a test includes material not specified for coverage—often accounts for examinees' informal evaluations of a test as "unfair."

Various quantitative procedures have been developed to gauge and help address concerns about alignment. Among the most commonly used are the Survey of Enacted Curriculum (Porter & Smithson, 2001) and the Webb alignment method (Webb, 1997, 2002). Each of these methods results in an index ranging from 0.0 (no alignment) to 1.0 (perfect alignment). The method proposed by Webb is the most commonly used method for gauging alignment between the content standards and assessments used by states as part of federally mandated annual student testing. The method provides quantitative summaries of various aspects of alignment, including *categorical concurrence* (i.e., the extent to which a test contains an adequate number of items measuring each content standard); *depth of*

*knowledge* (i.e., the extent to which the items or task in a test are as cognitively demanding as suggested by the content standards); *range of knowledge correspondence* (i.e., the extent to which at least half of the subobjectives for a content standard are covered by the test); and *balance of representation* (i.e., the extent to which the objectives for a content standard included in a test are addressed in an even manner). Overall, consistency in form development over time and attention to alignment help promote fairness to the extent that examinees' test results are not advantaged or disadvantaged because of the particular test form they were administered, nor because the domain covered by the test was an uneven or unrepresentative sample of the content standards to which scores on the test are referenced. This principle is reflected in the *Standards for Educational and Psychological Testing*, which states that "test developers should document the extent to which the content domain of a test represents the defined domain and test specifications" (AERA, APA, & NCME, 1999, p. 45).

## Standard Setting

Whereas the term *content standards* refers to the collections of statements regarding what examinees are expected to know and be able to do, *performance standards* refers to the levels of performance required of examinees on a test designed to assess the content standards. Although subtle and sometimes important distinctions can be made (see Cizek, 2006; Kane 1994), the term *cut score* is often used interchangeably with *performance standard*. Further, it is important to note that, although cut scores are typically derived as result of the procedures described in this section, it would be inaccurate to say that the panels of participants who engage in such procedures "set" the performance standards. Rather, such panels almost always serve as advisory to the entity with the legal or other authority to determine the cut scores that will be applied to examinees' test performances.

According to the *Standards for Educational and Psychological Testing*:

> A critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. . . . [C]ut scores embody the rules according to which tests are used or interpreted. Thus, in some situations, the validity of test interpretations may hinge on the cut scores. (AERA, APA, & NCME, 1999, p. 53)

In the context of licensure and certification testing, the *Standards* notes that "the validity of the inferences drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance" (p. 157).

The performance standards for a test are used to define various categories of performance, ranging from a simple dichotomy (e.g., *pass–fail*) used for many licensure or certification examinations, to more elaborate classifications such as *basic*, *proficient*, and *advanced* used in many student achievement testing programs. Performance standards may be expressed in a raw score metric (e.g., number correct), an IRT metric (e.g., theta value), or another metric (e.g., transformed or scaled scores).

There are five steps common to all standard-setting procedures: (1) choice of standard setting method, (2) selecting and training qualified participants, (3) providing feedback to participants, (4) calculating the cut score(s), and (5) gathering validity evidence. Each of these steps involves ethical concerns. The following portions of this chapter address steps 1, 2, 3, and 5.

Although numerous methods exist, the chosen standard setting method should be related to the purpose, format, and other characteristics of the test to which it will be applied. Detailed descriptions of the possible methods are presented elsewhere (see Cizek, 2001; Cizek & Bunch, 2007) and are beyond the scope of this chapter. Whatever method is selected, there are two primary goals—transparency and reproducibility—with the former a necessary condition for the latter. The first key goal—transparency—requires that the process for gathering judgments about cut scores should be carefully and explicitly documented. The *Standards* indicates that "when a validation rests in part on the opinions or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described" (AERA, APA, & NCME, 1999, p. 19). In addition, transparency helps to ensure that the chosen standard setting method is well aligned to the purpose of the examination. The goal of reproducibility also requires careful following and documentation of accepted procedures, but it also requires an adequate number of participants so that the standard error of participants' judgments about the cut scores is minimized. Fundamental fairness requires that any cut scores should be stable and not a statistical anomaly; if the standard-setting procedure were repeated under similar conditions, it is important to have confidence that similar cut scores would result. The *Standards* (AERA, APA, & NCME, 1999) provides some guidance on representation, selection, and training of participants (called "judges" in the *Standards*). For example, regarding the number of participants that should be used, the *Standards* indicates that "a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were replicated" (AERA, APA, & NCME, 1999, p. 54). In practice, logistical and economic factors must be considered when determining the sample size for standard-setting studies, but in general as large a group as feasible should be used to enhance reproducibility.

When selecting and training participants for a standard-setting activity, the ethical concerns center on the qualifications and representativeness of those who will participate in the process, and on how well prepared the participants were to engage in the standard-setting task. Potential participants must be knowledgeable regarding the content to be tested and the characteristics of the examinee population. Here, the *Standards* (AERA, APA, & NCME, 1999) requires that "the qualifications of any judges involved in standard setting and the process by which they are selected" (p. 54) should be fully described and included as part of the documentation of the standard setting process and that the standard setting process "should be designed so that judges can bring their knowledge and experience to bear in a reasonable way" (p. 60).

Whereas decisions about representation on standard setting panels are ultimately a policy matter for the entity responsible for the testing program, some ethical guidelines apply. The concern about appropriate qualifications can be illustrated in the contexts of achievement and credentialing tests. For the high school mathematics test used as a part of diploma-granting decisions, qualified participants would need to know about the mathematics curriculum and content covered by the test, the characteristics of the high school students who must pass the examination, and the mathematical knowledge and skill required in the variety of contexts that the students will encounter after graduation. Thus, it would be appropriate to include high school mathematics teachers on the standard-setting panel, but also representatives of higher education, business, the military, parents, and other community members. In contrast, in standard setting for licensure or certification, the primary purpose is often public protection. According to the *Standards*, "the level of performance required for passing a credentialing test should be dependent on the knowledge and skills necessary for acceptable performance in the occupation or profession" (AERA, APA, & NCME, 1999, p. 162). Thus, it would be most appropriate to include entry-level practitioners or those who already hold the credential that is the focus of the examination, as well as public representatives whose interests are ultimately served.

According to the *Standards*:

> Care must be taken to assure that judges understand what they are to do. The process must be such that well-qualified judges can apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions. (AERA, APA, & NCME, 1999, p. 54)

Thus, the training of participants in the selected standard-setting procedure is also a critical step; the method used should be one that allows

participants to make the judgments described, and the training should adequately prepare them to do so. One of the mechanisms for accomplishing this is to provide participants with various kinds of information to help them make their judgments.

There are three basic kinds of information. *Normative data* permit panelists to compare their judgments with those of other participants, and it is perhaps the most common type of feedback used in standard-setting studies. Normative data consist of information such as a distribution of item ratings or overall cut scores, minimum and maximum ratings, and mean or median ratings in the form of frequency distributions, bar graphs, or other data summaries that are easy for the participants to interpret and use. *Reality data* are provided to assist participants in generating realistic judgments. Reality information typically consists of item difficulty indices (i.e., *p* values) or theta-scale values for individual items. Reality information can be computed based on complete samples of test-takers or may be computed based on subsamples, such as examinees around a given point in the total score distribution. Panelists use this information to help them gauge the extent to which their judgments relate to the performance of examinees or test items. Finally, *impact data* (also called *consequence data*) are provided to aid panelists in understanding the consequences of their judgments. Typically, impact data consist of information about the number or percentage of examinees who would pass a test, or who would fall into a given performance level if a recommended cut score was implemented. The three types of information are typically provided across "rounds" of judgments, so that participants have opportunities to revise their judgments in response to the information provided.

The ethical implications of providing these data are clear. First, normative data are provided so that participants can gauge the extent to which their judgments concur with other qualified participants—and to make revisions as they deem appropriate. Reality data are provided so that participants' judgments are grounded in the performance of the examinees who are subject to the test, and not merely dependent on the leniency, stringency, or perspectives of the participants. Impact data are provided so that those who make cut score judgments are aware of the effect on test-takers. For example, it would not seem reasonable to deny graduation to all high school seniors (if a cut score was set too high) or to certify all examinees who take a test in brain surgery (if a cut score was set too low).

Gathering validity evidence to support the cut score recommendations is another ethical responsibility of those who engage in standard setting. Hambleton (2001) and Pitoniak (2003) have outlined several sources of potential validity evidence. These include *procedural* fidelity and appropriateness, as well as *internal* evidence (e.g., the degree to which participants

provide ratings that are consistent with empirical item difficulties, the degree to which ratings change across rounds, participants' evaluations of the process) and *external* evidence (e.g., the relationship between decisions made using the test and other relevant criteria such as grades, supervisors' ratings of job performance, performance on tests measuring similar constructs, etc.).

Evaluation of standard setting must attend to the omnipresent reality that classifications made based on the cut scores may be incorrect. To the extent that sample size, alignment, representativeness or qualifications of the participants, or other factors are compromised, the cut scores resulting from a standard-setting procedure might unfairly classify as "failing" some examinees who truly possess the knowledge, skills, or abilities deemed necessary to be classified into a certain performance category. Conversely, some examinees who do not possess the knowledge, skills, or abilities deemed necessary may be mistakenly classified as "passing." These classification errors are often referred to as *false-negative* and *false-positive* errors, respectively. Although it is true that nearly all test-based classification decisions will result in some number of classification errors, an ethical obligation of those who oversee, design, and conduct standard-setting procedures is to minimize such errors.

Finally, a specific unethical action is highlighted in the *Standards* related to setting cut scores for licensure or certification examinations. According to the *Standards*:

> The level of performance required for passing a credentialing test should be dependent on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test. (AERA, APA, & NCME, 1999, p. 162)

## Validation

Among all criteria by which tests are evaluated, validity is universally endorsed as the most important. For example, the *Standards* asserts that validity is "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999, p. 9). A necessary (although insufficient) precondition for the ethical use of any test is the collection and evaluation of adequate validity evidence.

Refining Messick's (1989) definition, in this chapter we define *validity* as the degree to which scores on an appropriately administered instrument reflect variation in the characteristic it was developed to measure and support the intended score inferences. By extension, we define *validation* as the ongoing process of gathering relevant evidence for generating

an evaluative summary of the degree of fidelity between scores yielded by an instrument and inferences about standing on the characteristic it was designed to measure. That is, validation efforts amass and synthesize evidence for the purpose of articulating the degree of confidence that intended inferences are warranted.

Validation centers on a concern about the quality of the data yielded by an instrument. That concern is heightened whenever a test is one part of procedures for making important decisions in countless situations in which the information yielded by a test has meaningful consequences for persons or systems. Because it would be unethical to use information that is inaccurate, misleading, biased, or irrelevant—that is, lacking validity—to make such decisions, validity is rightfully deemed to be the most important characteristic of test scores. The topic of validity is treated in substantial depth elsewhere in this Handbook (see Carrig & Hoyle, Chapter 5, this volume); thus, we will only briefly summarize six broadly endorsed tenets of validity here.

First among the accepted tenets is that validity pertains to the inferences that are made from test scores. Because latent traits and abilities cannot be directly observed, these characteristics must be studied indirectly via the instruments developed to measure them, and inference is required whenever it is desired to use the observed measurements as an indication of standing on the unobservable characteristic. Because validity applies to the inferences to be made from test scores, it follows that a clear statement of the intended inferences is necessary to design and conduct validation efforts.

Second, validity is not a characteristic of instruments but rather of the data generated by those instruments. Grounded in the position first articulated by Cronbach (1971), the current *Standards* notes that "it is the interpretations of test scores that are evaluated, not the test itself" (1999, p. 9).

Third, the notion of discrete kinds of validity (i.e., content, criterion, construct) has been supplanted by the realization that, ultimately, all evidence that might be brought to bear in support of an intended inference is evidence bearing on the responsiveness of the instrument to variation in the construct measured by the instrument. This conceptualization of validity is referred to as the *unified view of validity*, and validity is now generally regarded as a singular phenomenon. In describing the unified view, Messick has indicated that "What is singular in the unified theory is the kind of validity: All validity is of one kind, namely, construct validity" (1998, p. 37).

Fourth, judgments about validity are not absolute. As Zumbo has stated, "Validity statements are not dichotomous (valid/invalid) but rather are described on a continuum" (2007, p. 50). There are two reasons why this must be so. First, in a thorough validation effort, the evidence is routinely mixed in terms of how directly it bears on the intended inference,

its weight, and its degree of support for the intended inference. Second, because validation efforts cannot be considered "completed" at a specific juncture, evidence amassed at any given time must necessarily be considered tentative and a matter of degree.

Fifth, validation is an ongoing enterprise. Just as it is incorrect to say that *a test* is valid, so it is incorrect to say that the validity case for an intended inference is closed. Many factors necessitate a continuing review of the empirical and theoretical information that undergirds the inferences made from test scores. For example, replications of the original validation efforts, new applications of the instrument, new sources of validity evidence, new information from within and beyond the discipline about the construct of interest, and theoretical evolution of the construct itself all represent new information that can alter original judgments about the strength of the validity case.

The final tenet of modern validity theory is that the process of validation necessarily involves the exercise of judgment and the application of values. Because searching validation efforts tend to yield equivocal evidence, the available evidence must be synthesized, weighed, and evaluated. Kane (2001) has observed that "validity is an integrated, or unified, evaluation of the [score] interpretation" (p. 329). Validation efforts result in tentative conclusions about the degree to which evidence supports confidence that a test, administered to its intended population under prescribed conditions, yields accurate inferences about the construct it is intended to measure. Lacking such evidence, or when the evidence fails to support a desired level of confidence, the use of test data to make important decisions about examinees would be unethical.

The *Standards* lists and describes five sources of validity evidence. They include (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relations to other variables, and (e) evidence based on consequences of testing. Common threats to the validity of test score inferences include *construct underrepresentation*, in which "a test fails to capture important aspects of the construct," and *construct irrelevant variance*, in which "test scores are affected by processes that are extraneous to its intended construct" (AERA, APA, & NCME, 1999, p. 10). An example of the former would be a licensure test of automobile driving ability that involved only a written, knowledge component; an example of the latter would include a test of English composition ability for which scores were influenced by examinees' handwriting. The ethical aspects of validity in these examples are clear: It would be inappropriate to license drivers who lacked driving skill; it would be unfair if examinees of equal writing ability were assigned different scores based on a characteristic (handwriting legibility) that the test was not intended to measure.

## Test Administration and Scoring

Ethical concerns are also present when tests are administered and scored. The following subsections of this chapter describe aspects of test administration and scoring where quantitative procedures can be invoked to advance the goal of ethical testing practice. These aspects include (a) test registration and test preparation; (b) test administration conditions, accommodations, and security; and (c) scoring procedures.

### Registration and Examinee Test Preparation

When preparing to administer a test, it may be necessary first to determine whether examinees meet the eligibility guidelines for taking the test. Such guidelines may include, among other things, academic preparation requirements, age or residency requirements, and completion of required internships and supervised or independent practice. Where such qualifying criteria exist, it is an ethical obligation of the entity responsible for the testing program to ensure that only eligible candidates are permitted to take the examination.

Once it has been determined that an examinee has met the eligibility requirements, the entity should provide candidates with information about ethical and unethical test preparation activities and should follow rigorous procedures to preclude examinees from having improper prior access to test materials. According to the *Standards*, "test users have the responsibility of protecting the security of test materials at all times" (AERA, APA, & NCME, 1999, p. 64). Likewise, the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004) requires that test developers "establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting" (p. 6) and that test users should "protect the security of test materials, including respecting copyrights and eliminating opportunities for test takers to obtain scores by fraudulent means" (p. 7).

Second, where there is reason to believe that examinees may be inappropriately advantaged or disadvantaged by aspects of the test administration (e.g., computer-based mode of administration, test format), the responsible entity should take steps to address such concerns. For example, if specialized software or an unfamiliar computer interface will be used for testing, examinees should be provided with opportunities to practice with the software or interface, ideally well in advance of the day of testing. Whereas it might be reasonable to assume that test-takers are familiar with multiple-choice item formats, some tests may contain formats that would be less familiar to examinees, such as gridded-response formats, drag-and-drop completion items, or other novel response formats. In such

cases, and also to provide examinees with an opportunity to gauge the content coverage, level of difficulty, and other test-related factors, it is desirable to provide examinees who register for an examination with a practice test form that they can complete after registration but before taking the operational test.

All the major ethical standards support these recommendations. For example, they are in line with the relevant guidelines in the *Standards*, which note, among other things, that "Instructions should … be given in the use of any equipment likely to be unfamiliar to test takers [and] opportunity to practice responding should be given when equipment is involved" (AERA, APA, & NCME, 1999, p. 63). Similarly, according to the *Rights and Responsibilities of Test Takers* (Joint Committee on Testing Practices, 1998), testing professionals should "make test takers aware of any materials that are available to assist them in test preparation" (p. 8) and should "provide test takers with information about the use of computers, calculators, or other equipment, if any, used in the testing and give them an opportunity to practice using such equipment" (p. 10). Finally, the *Code of Fair Testing Practices* indicates that test users should "provide test takers with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing" (Joint Committee on Testing Practices, 2004, p. 7).

### Test Administration Conditions

A primary ethical obligation in testing is to ensure that test scores accurately reflect the true knowledge, skill, or ability of the test-taker. One way to help accomplish this goal is to establish testing conditions that do not advantage or disadvantage any test-takers. This means, among other things, ensuring that the test setting is conducive to examinees providing their best performance and configured to deter the potential for unethical behavior. Accomplishing these goals requires more than simply providing adequate lighting and seating; test security must be maintained throughout the testing process, including during test administration. According to the *Standards*, "the testing environment should furnish reasonable comfort and minimal distractions" (p. 63) and "reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means" (AERA, APA, & NCME, 1999, p. 64). The *Rights and Responsibilities of Test Takers* also indicates that testing specialists should "take reasonable actions to safeguard against fraudulent actions (e.g., cheating) that could place honest test takers at a disadvantage" (Joint Committee on Testing Practices, 1998, p. 11).

Beyond ensuring test security, it is an ethical responsibility of testing specialists to ensure that neither the testing conditions nor surface features of the test itself interfere with accurate measurement. The *Standards* notes this

goal in the first of 12 standards related to testing individuals with disabilities, noting that "test developers, test administrators, and test users should take steps to ensure that the test score inferences reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement" (AERA, APA, & NCME, 1999, p. 106).

To accomplish this, some test-takers with special physical or other needs may require adjustments to the testing conditions to demonstrate their knowledge or skills. In general, there are two broad categories of such adjustments. One category, testing *modifications*, involves an alteration in the testing conditions that also alters the construct intended to be measured by the test and reduces confidence in the validity of interpretations of the examinee's test score. The other category, testing *accommodations*, also involves altered testing conditions, but in such a way as that the construct of interest and intended score inferences are unchanged. For example, allowing an examinee to wear glasses or contact lenses would be an accommodation for a reading comprehension test because the construct of interest is reading comprehension and the use of corrective lenses is unrelated to the measurement of that construct. However, the same adjustment in testing conditions would be considered a modification if the examinee were taking a vision test. In that case, the adjustment *is* related to the characteristic being assessed and would adversely affect the accuracy of conclusions about the examinee's vision. A complete classification system for testing accommodations has been developed by Thurlow and Thompson (2004) and is shown in Table 8.1 with examples of each type of accommodation. Overall, testing specialists must carefully evaluate any alterations in testing conditions to ensure fairness; that is, to ensure that accommodations are obtained by those who need them (and not by those who do not), and to ensure that any alterations do not affect the validity of test scores.

Like test preparation, all the major ethical guidelines for testing address test conditions. According to the *Standards*:

> If the test developer indicates that the conditions of administration are permitted to vary form on test taker or group to another, permissible variation in conditions for administration should be identified and a rationale for permitting the different conditions should be documented. (AERA, APA, & NCME, 1999, p. 47)

The *Rights and Responsibilities of Test Takers* requires that test-takers, if they have a disability, should be advised that "they have the right to request and receive accommodations or modifications in accordance with the provisions of the Americans with Disabilities Act and other relevant legislation" (Joint Committee on Testing Practices, 1998, p. 10). And, according to the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004), test developers should "make appropriately modified forms of tests

**TABLE 8.1**

Categories of Accommodations

| Accommodation Type | Example |
|---|---|
| Setting | Accessible furniture; individual or small group administration |
| Timing | Extra time; frequent breaks during testing |
| Scheduling | Multiple testing sessions; different test days or times |
| Presentation | Audio, Braille, large-print, or other language version of a test |
| Response | Scribe to record student's answers; oral, pointing to indicate responses |
| Other | Highlighters, dictionaries, "reading rulers," or other aids |

*Source:* G. Walz (Ed.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators,* Pro-Ed, 2004.

or administration procedures available for test takers with disabilities who need special accommodations" (p. 4); test users should "provide and document appropriate procedures for test takers with disabilities who need special accommodations or those with diverse linguistic backgrounds" (p. 6).

## Scoring Procedures

After administration of a test, examinees' responses must be evaluated. A key fairness issue in evaluating the responses centers on the objectivity and reproducibility of the scoring. When responses are entered directly by examinees via computer or onto a form for optical scoring, the degree of objectivity and reproducibility is typically greater than if the responses involve performances, constructed responses to open-ended test items, or other response types that require human scoring. Of course, objectivity and reproducibility are issues even when scoring is automated. For example, subjective judgments must be made regarding the sensitivity settings on optical scanning equipment; judgments must be made when configuring algorithms for automated essay scoring; and so on. Thus, although it is possible to increase objectivity with these methods, it is not possible to eliminate all subjectivity in scoring.

Fairness in scoring has two aspects alluded to previously in this chapter. A necessary but insufficient condition for fair scoring is that it is consistent. That is, examinees who give the same responses should receive the same scores. Second, the scoring should be valid. That is, variation in scores assigned to responses should reflect variation in the characteristic that the instrument is intended to measure and, to the extent possible, no

other, unintended characteristics. The *Standards* (AERA, APA, & NCME, 1999) provides at least three specific recommendations related to scoring items and tasks:

- "The process of selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported." (p. 48)
- "The criteria used for scoring test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria for scoring may not be obvious to the user." (p. 46)
- "Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring." (p. 47)

Several quantitative and qualitative procedures can be implemented to facilitate the goals of reproducibility and accuracy. First, raters should be thoroughly trained in the analytical features of the response (i.e., performance, task, essay, etc.) that they will be scoring. Effective training focuses on ensuring that raters attend to the features of responses that are the intended object of measurement. For example, if scoring handwritten essays, training would focus on ensuring that raters evaluate the predetermined aspects of the essay specified in the directions to examinees (e.g., content, word choice, organization, style) and not that aspects deemed to be irrelevant (e.g., handwriting, spelling, neatness).

The process of *rangefinding* is used to help operationalize the boundaries of scoring categories. For example, suppose that a constructed-response mathematics problem appeared on the high school graduation test, and the directions required examinees to solve the problem and explain their solution. A scale might be used that assigned 0 points for missing or completely incorrect response, 1 point for an attempted but incorrect solution, 2 points for a partially correct solution lacking an explanation, 3 points for a correct response, but with missing or inadequate explanation, and 4 points for a correct solution with a complete, accurate explanation. In this scenario, the borderline between score points 3 and 4 is critical and hinges on judgments about the adequacy

of the explanation provided. Extensive work would be required to identify the range of responses that should be considered "inadequate" (and therefore receive a score of 3) versus responses that should be considered sufficiently "adequate" (and assigned a score of 4). Such judgments—admittedly subjective—would need to be made in advance of rater training, and training would need to be included in the training to ensure that similar responses were judged consistently.

*Validation samples* are used to gauge the efficacy of rater training. Validation samples are actual or model responses that exemplify the scale points in a scoring rubric or essential elements a response must contain to be assigned a certain score. After training, scorers evaluate the validation samples, and targets are established that specify the agreement rate a scorer must attain to qualify to rate operational responses. Raters who do not meet the qualification targets receive additional training or are disqualified from rating operational responses.

In the scoring of operational responses, raters continue to be monitored for their accuracy and consistency in evaluating examinees' responses. To gauge consistency, multiple raters may be assigned to rate the same responses independently, and rater agreement or rater reliability indices may be calculated (see von Eye & Mun, 2005). To monitor accuracy, typical procedures include the insertion (blind to the raters) of validation samples to assess the extent to which raters' scores agree with the (known) scores of the validation responses and assessment of the extent to which raters' exhibit errors of leniency, stringency, central tendency, or drift. These procedures reflect best practices in psychometrics and align with ethical standards of the profession, such as those found in the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004), which indicate that test developers should "provide procedures, materials and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, [test developers should] provide adequate training for scorers" (p. 6).

## Score Reporting and Use of Test Results

The third phase of testing where ethical concerns arise occurs when test scores are reported and test results are used. The following subsections of this chapter describe three aims of score reporting and use where quantitative methods are applied toward ensuring that ethical considerations are attended to when scores are calculated, reported, and used. The three aims include (a) promoting score comparability, (b) protecting confidentiality, and (c) ensuring score integrity.

## Score Comparability

A fundamental ethical issue in testing relates to the process of assigning scores to examinees who are administered different test forms (i.e., versions of a test that do not contain identical sets of test questions). Because of security concerns, most high-stakes testing programs use multiple forms. However, it is an issue of fairness that, when examinees are administered different forms, scores across the differing forms should be equivalent in meaning. As Angoff has noted regarding two test forms, X and Y, scores yielded by the forms are comparable if it is "a matter of indifference to [examinees] of every given ability level θ whether they are to take test X or test Y" (1980, p. 195). This means, for example, that it would of great ethical concern if two students of equal ability were differentially likely to pass a high school graduation test solely because they were administered different forms of the test. In general, examinees should not be penalized (or rewarded) for receiving one test form that may be slightly harder (or easier) than another.

Although equivalent forms are intended to be similar in difficulty, it is nearly impossible in practice to construct multiple test forms with exactly the same level of difficulty. *Equating* is a process used to adjust for slight differences in difficulty between test forms so that scores from all forms can be placed on the same scale and used interchangeably. For example, an equating analysis may result in performance of 24 of 30 questions on Form X being determined as equivalent to obtaining 26 of 30 questions correct on Form Y. Although examinees might perceive Form X as being slightly harder and an appearance of "unfairness," this variation in difficulty would not pose an ethical concern if scores on the two forms were properly equated.

There are several different data collection designs used for equating, detailed by Kolen and Brennan (2004) and briefly described here. The first type of equating design is the *random groups* design, in which examinees are randomly assigned to different forms, and each examinee group is assumed to be equivalent in ability. This design could pose ethical concerns if examinee groups are not equivalent, so it is necessary that the assumption of randomly equivalent groups be evaluated. The second design is the *single group with counterbalancing* design, where each examinee takes both (or all) forms of an examination, and the order of the forms is counterbalanced to control for order effects. The third design is the *common-item nonequivalent groups* design, where groups of examinees (that are not necessarily equivalent) take different forms that contain a subset of the same items; these common or "anchor" items are used to place all items on the same scale. The extent to which this design results in valid and ethical score interpretations depends in large part on the characteristics of the common items, which should be evaluated to determine the

extent to which they can be considered a representative subsample of the full test.

Technical details on equating methods are beyond the scope of this chapter, but interested readers should consult Kolen and Brennan (2004) for a thorough discussion of the appropriate methods for each equating design. Some of the most common equating methods include mean and linear equating, equipercentile methods, and IRT methods. In the simplest equating procedure, *mean equating*, scores are converted by adding a constant to account for differences in the mean scores on each form. In addition to the mean, the standard deviation is also taken into account during the transformation process when using linear equating. Equipercentile equating involves transforming the score scales by setting percentile ranks of scores on different forms to be equal. IRT methods achieve score transformation by using item parameters of established items to calibrate new items and estimate examinee ability. Whatever equating method is used, it is important to recognize that scores and subsequent decision making may be affected by both the equating process itself and the resulting error associated with the equating process. The *Standards* (AERA, APA, & NCME, 1999) addresses score comparability, including the importance of providing evidence that scores on different test forms are interchangeable and assuring that relevant assumptions for equating procedures have been satisfied. For example, the *Standards* requires that:

> "A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably" (p. 57).
>
> "When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of the equating functions" (p. 57).
>
> "In equating studies that rely on the statistical equivalence of examinee groups receiving different forms, methods of assuring such equivalence should be described in detail" (p. 58).

## Confidentiality

Because test data represent information about individuals that can be used in both beneficial and harmful ways, it is an ethical responsibility of those who report test results to ensure that scores are used appropriately. Many testing situations may require that test results be released confidentially and only with the permission of the examinee to those he or she authorizes.

According to the *Standards*, "Test results identified by the names of individual test takers, or by other personally identifying information, should be released only to persons with a legitimate, professional interest in the test taker or who are covered by the informed consent of the test taker." (AERA, APA, & NCME, 1999, p. 87). Additional ethical obligations apply even if the test scores are released appropriately to such persons: "Professionals and others who have access to test materials and test results should ensure the confidentiality of test results and testing materials." (p. 132). The *Code of Fair Testing Practices* indicates that both test developers and test users should "develop and implement procedures for ensuring the confidentiality of scores" (Joint Committee on Testing Practices, 2004, pp. 7, 8).

Confidentiality concerns apply not only to test scores but also to other aspects of testing. For example, according to the *Rights and Responsibilities of Test Takers*, those responsible for test information should also "keep confidential any requests for testing accommodations and the documentation supporting the request" (Joint Committee on Testing Practices, 1998, p. 19). Federal regulations, such as the *Family Educational Rights and Privacy Act* (FERPA, 1974) also apply to test results, which are considered "educational records" under FERPA. According to the FERPA law, except for narrow exclusions, educational records cannot be disclosed within or outside the educational institution to those who do not have a legitimate educational interest in the information. When the test scores or other records are those of a minor, a parent's or guardian's written consent for disclosure must be obtained; if the records are those of an adult, the adult's consent must be obtained.

Finally, even group reporting of test results can lead to inadvertent breaches of confidentiality. The problem of what has been called *deductive disclosure* is of increasing concern in many testing situations. Deductive disclosure occurs when an individual's identity or confidential test information can be deduced using other known characteristics of the individual. For example, suppose that a high school released individual performance results—with students' names removed—from a mathematics examination used as part of granting diplomas. In even moderately large high schools, if the individual test performance data were accompanied by collateral information about each student (e.g., sex, race/ethnicity, class level, number of test retake opportunities, middle school attended), it may be possible to determine the identity of an individual test-taker and his or her test performance.

### Ensuring Score Integrity

One of the most important aspects of the test-reporting process is communicating information about the test scores to examinees and other interested parties. A primary ethical consideration in score reporting is the importance of providing information about confidence in test scores

(and/or resulting decisions that are made). All test scores contain a certain amount of uncertainty, and error may be a result of sampling, measurement, and other sources.

Both the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1998) and the *Standards* (AERA, APA, & NCME, 1999) stress the importance of conveying information about error and precision of test results to the intended audiences. For example, the *Standards* indicates that "The standard error of measurement, both overall and conditional (if relevant), should be reported . . . in units of each derived score recommended for use in score interpretation" (p. 31); they also require that those involved in scoring tests "should document the procedures that were followed to assure accuracy of scoring [and] any systematic source of scoring errors should be corrected" (p. 64). The *Code* recommends that test developers should "provide evidence that the technical quality, including reliability and validity, of the test meets its intended purpose" (p. 4); "provide [test-takers with] information to support recommended interpretations of the results" (p. 6); and "advise test users of the benefits and limitation of test results and their interpretation" (p. 6).

Information about error and precision should include appropriate sources of error and estimates of their magnitude. In addition, the information should be most relevant to the intended uses of the test. For example, technical documentation on a high school graduation test should not be limited to an estimate of reliability or overall standard error of measurement but should also include information about decision consistency and the standard error of measurement at the cut score. The magnitude of error near a performance standard (e.g., a cut point that separates pass and fail categories) would be of primary interest. Speaking directly to the issues of accuracy and precision, respectively, the *Standards* requires that "when a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument" and that "standard errors of measurement should be reported in the vicinity of each cut score" (p. 35).

Another ethical aspect of score integrity is the importance of providing appropriate interpretive aids and avenues for appeal. According to the *Rights and Responsibilities of Test Takers*, test-takers have "the right to receive a written or oral explanation of [their] test results within a reasonable amount of time after testing and in commonly understood terms" (Joint Committee on Testing Practices, 1998, p. 6). Both the *Standards* (AERA, APA, & NCME, 1999) and the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004) stress the importance of communicating how test scores should (and should not) be interpreted, appropriate uses, and the error inherent in the scores. It is important that this information is

conveyed in simple language to all interested parties. According to the *Standards*:

> In educational testing programs and licensing and certification applications, test takers are entitled to fair consideration and reasonable process, as appropriate to the particular circumstances, in resolving disputes about testing. Test takers are entitled to be informed of any available means of recourse. (AERA, APA, & NCME, 1999, p. 89)

The *Code* requires that "test developers or test users should inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores" and that test-takers should be provided with information about their "rights to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid" (p. 10).

If subscores are reported, it is important that psychometric analyses be conducted at the subtest level. In many cases, subtests may not contain a sufficient number of items to be diagnostically useful, and the test may not have adequate validity evidence for this level of inference. For example, a high school graduation test in mathematics is likely to comprise several subtopics, including geometry and algebra. If scores for the algebra items are to be reported separately, it is important to have psychometric support for such a practice; that is, to demonstrate that the algebra items form a cohesive group and are sufficiently reliable. Comparisons between subscores should take into account the reliability of difference scores. If student performance were compared across different subtests without taking the reliability of difference scores into account, judgments of apparent differences might be entirely due to error in the scores. According to the *Standards*, when such scores are used, "any educational decision based on this comparison should take into account the extent of overlap between the two constructs and the reliability or standard error of the difference score" (p. 147).

A final aspect of score integrity that has ethical implications is the use of test scores for secondary purposes. The *Standards* (AERA, APA, & NCME, 1999) is clear that tests require evidence in support of each intended purpose, and appropriate evidence for one purpose may not support (and may even detract from) evidence needed for a different purpose. Evidence supporting a high school graduation test likely centers on analyses relating student performance on the test to skills needed outside of school. Whether it is ethical to use the results from high school graduation tests for other purposes, such as making inferences about teachers or schools, depends on the extent to which evidence has been collected for those purposes as well. No test is equally justifiable for all purposes, and the

intended inferences must be taken into account when using the scores. According to the *Standards*, "If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned against making unsupported interpretations" (p. 18).

## Conclusion

The context of testing—especially high-stakes testing—often comprises decision-making processes that result in classifications that can be consequential for people, groups, organizations, or systems. Personnel hiring and promotion decisions, psychological diagnoses, licensure and credentialing decisions, and educational admission, placement, retention, promotion, and graduation decisions are only a few examples of contexts in which test scores are used, most often in conjunction with other relevant information, to provide or withhold credentials, treatments, opportunities, and so on. All these situations are fraught with junctures at which insufficient psychometric safeguards could adversely affect those affected by test scores. It is only somewhat of an exaggeration to label the ethical concerns as life-or-death matters. In fact, the psychometric technology of standard setting was an important aspect of a U.S. Supreme Court case (*Atkins v. Virginia,* 2002) in which a convicted murderer, Daryl Atkins, had been sentenced to death. The sentence was overturned by the Supreme Court because Atkins' measured IQ of 59, derived from administration of the *Wechsler Adult Intelligence Scale*, fell below a cut score of 60. The execution of mentally retarded individuals was considered by the Court to be "cruel and unusual" and hence prohibited by the 8th Amendment (cited in Cizek & Bunch, 2007, p. 6).

Different, less dramatic, circumstances involving tests occur for individuals in many aspects of their lives, but the ethical concerns are the same. The science and practice of psychometrics has evolved and developed methods that are responsive to these concerns toward the goals of enhancing the accuracy of the information yielded by social science instruments and promoting the appropriate use of test results. The armamentarium of the assessment specialist currently comprises many quantitative tools for facilitating these goals. However, research and development efforts must continue to improve current methods and develop new ones that will equip those who develop and use tests with the tools to improve outcomes for the clients, students, organizations, and others who are the ultimate beneficiaries of high-quality test information.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

*Atkins v. Virginia.* (2002). 536 U.S. 304.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). New York: Praeger.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225–258). Mahwah, NJ: Erlbaum.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Crocker, L., & Algina, J. (1986). *An introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

*Family Educational Rights and Privacy Act.* (1974). 20 U.S.C.1232.

Gould, S. J. (1996). *The mismeasure of man.* New York: Norton.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Washington, DC: Author. Retrieved from http://www.apa.org/science/ttrr.html

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–461.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44.

Miyazaki, I. (1976). *China's examination hell: The civil service examinations of Imperial China.* New York: Weatherhill.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author.

Pitoniak, M. J. (2003). *Standard setting methods for complex licensure examinations.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–159). Mahwah, NJ: Lawrence Erlbaum.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research Report Series No. RR-048). Philadelphia: University of Pennsylvania Graduate School of Education, Consortium for Policy Research in Education.

Raymond, M., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181–224). Mahwah, NJ: Erlbaum.

Raymond, M. R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education, 9,* 237–256.

Thurlow, M. L., & Thompson, S. J. (2004). Inclusion of students with disabilities in state and district assessments. In G. Walz (Ed.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 161–176). Austin, TX: Pro-Ed.

von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement*. Mahwah, NJ: Erlbaum.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

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

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Erlbaum.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16,* 33–45.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 45–79). Amsterdam: Elsevier Science.

# 9

# *Ethics in Program Evaluation*

**Laura C. Leviton**
*The Robert Wood Johnson Foundation*

Program evaluation uses both quantitative and qualitative methods to understand the implementation and outcomes of practices, programs, and policies designed to address social problems, as well as to assess cost-effectiveness, mediators, and moderators of outcomes (Scriven, 1991; Shadish, Cook, & Leviton, 1991). In a high quality quantitative program evaluation, the objectives and measures are carefully chosen; statistical power is adequate; multiple data sources and methods are used; analysis is performed properly; and interpretation is transparent to the program staff and other interested parties.

We assert that the ethical case for methodological quality in program evaluations is particularly urgent, given the consequences for programs and people. At the same time, this case needs ongoing discussion and clarification because the answers are by no means clear. There are three general reasons for this: (a) The field is still emerging, so program evaluation has only a limited consensus on what constitutes appropriate methods. Evaluation methods are pluralistic and are certainly not confined to quantitative methods; (b) We would like to hold quantitative evaluations, in particular, to the standards of quality seen in other fields. As will be seen, however, real world constraints often preclude optimal designs and data collection. Yet the need for useful information can be acute, so the challenge is to determine what can be learned, given those constraints. Programs need oversight and improvement, so it is not acceptable to equate ethical evaluations with optimal methodology. Instead, evaluations need to be transparent about the methods that have been used, humble about their failings, and specific about the limitations of the findings; and (c) Because evaluations have important consequences, discourse about them can be inflamed by ideology or politics, resulting in disingenuous or ill-informed accusations of misconduct. Because progress in the field depends on rational discourse, it is important to describe exactly what the most salient ethical issues are and, as best we can, distinguish ethical and unethical conduct.

I first describe the general linkages between program evaluation and ethical considerations. Next, I describe the standards, guiding principles, and case examples on which the field of program evaluation relies for its ethical compass. Third, I define three ethical principles that are particularly important to social programs and policies and apply them to program evaluation. Ethical issues arise at each stage in the production of evaluations, and the next section outlines some of the most salient issues for quantitative evaluations. In the final section, I reflect on the need to distinguish between ethical, technical, and ideological considerations in the choice of methods for evaluation.

## Overview of the Intersection of Ethics and Program Evaluation

For quantitative evaluations, the ethical issues are urgent. Program and policy evaluations can produce concrete benefit or harm to real people in real time. Well-run programs consistent with the public interest have major consequences for people's ability to eat (food assistance and school meal programs), to work (job training and economic development programs), to live a healthy life (public health programs, medical care coverage), to live productively in society (education, early childhood and youth development, elder services, supportive services for people with disabilities), to live in comfort and security (income, housing, criminal justice, Homeland Security programs), and to have access to culture and personal development (museums, libraries, orchestras, zoos). All these programs have undergone evaluation. A false step for their evaluation can have far-reaching impact.

Ethical issues are intimately tied to evaluation in at least four ways. First, all definitions of evaluation include its role in assigning value to something: An object is found to be good, bad, or somewhere in between, using agreed-on criteria (Scriven, 1991). These criteria imply ethical principles at work. For example, if we value social justice, then a program is good if it reduces disparities between rich and poor. If we value efficient weapons of war, then we can evaluate whether a missile performs well or poorly. As seen in these two examples, we can evaluate many different kinds of objects, but the term *evaluation* most often refers to (a) practices, such as counseling services; (b) programs, such as those funded by federal and state governments or local nonprofit organizations; and (c) policies, whether embodied in law, regulation, or organizational procedures. We will refer to them collectively as *program evaluation*.

A second tie to ethics is the process whereby personal, professional, and societal beliefs and norms affect the choices investigators make in the

production of evaluations (Shadish et al., 1991). The production process is not objective in any sense, although it is commonly portrayed this way to laypeople. Understanding why evaluators make certain choices at each production stage can make the ethical considerations explicit. Norms and values drive choices about what gets evaluated, who gets consulted on the evaluation questions, the way the questions are framed, the design, measurement instruments and data collection, the preferred analytic methods, the interpretation of the results, the jargon or clear presentation in the reports, and the choice of audiences for dissemination. Yet these choices are often framed as technical issues in evaluation, especially when they are called into question and evaluators need to justify their choices. Technical quality is still relevant—after deciding on a general direction and evaluation questions, we still want the best technical standards to be brought to bear. If, for example, we choose to use qualitative methods, we want to use the most widely recommended procedures for data collection and analysis; if we choose to conduct a survey, we want to select valid and reliable items, ensure an adequate response rate, and handle missing data appropriately. As others in this volume point out, the choice of high technical quality is also an ethical decision. I would modify that observation in the case of program evaluation: It is ethical to choose high technical quality given the constraints of the situation, the value of the information, and a reasonable self-awareness about value judgments in the choice of methods.

The third tie to ethics lies in the consequences for people who are served by programs, as I have described, and the fourth tie lies in the consequences for the taxpayer and for government as a whole. Government funds and manages programs with the understanding that tax dollars will be used wisely. Evaluation emerged as a way to hold programs accountable for the use of public resources (Shadish et al., 1991), and that approach is extended now to encompass nonprofit and philanthropic resources (Carman & Fredericks, 2008). We would like to be sure that tax money and charitable contributions are allocated to effective, important services. Our degree of certainty about these services is strongly affected by the quality of evaluation findings (Weiss & Bucuvalas, 1980).

Mismanagement or a loss of program funding can occur for many reasons, but flawed evaluation information should not be one of them. The justification for conducting an evaluation is that it may be useful to decision makers, program staff, program beneficiaries, and other stakeholders (those who have an interest, or "stake" in a program) (Shadish et al., 1991; Weiss, 1998). Unlike other applied social research, evaluation really has no other inherent justification. Evaluations of outcome are intended to contribute to decisions about resource allocation by public and nonprofit organizations. Although it is unusual that evaluations lead directly to programs' expansion or decline, they can contribute to

such decisions along with a body of other evidence (Ginsburg & Rhett, 2003; Leviton & Boruch, 1983; Shadish et al.). More commonly, evaluations are used to improve programs (Cousins & Leithwood, 1986) or to test the assumptions of program planners and decision makers about strategy (Weiss & Bucuvalas, 1980). Many evaluations have done so at all levels of government and the nonprofit sector. For example, a recent survey of nonprofit organizations indicates that evaluations can be valued for their ability to clarify program models and direction (Carman & Fredericks, 2008). Policy makers will consider evaluation findings along with other information as they seek to redirect programs to be consistent with their vision of the public good (Leviton & Boruch, 1983; Weiss & Bucuvalas, 1980).

Unfortunately, many program evaluations are not well conducted—even using the pluralistic criteria of the evaluation field (e.g., Lipsey, 1988; Wilson & Lipsey, 2001). Their flawed conclusions can sometimes have grave consequences. Perhaps the most celebrated example is the Westinghouse evaluation of the effects of Project Head Start in the late 1960s (McGroder, 1990). The evaluation concluded that improved cognitive and language skills were seen in first grade, but that these gains disappeared by second and third grades. Citing the study, the Nixon administration tried to eliminate Head Start but was not successful. This experience and others led Weiss (1987) to worry that negative evaluations would have a chilling effect on innovations designed to address the needs of the most vulnerable groups in our society.

The Westinghouse evaluation was strongly challenged on design and analytic grounds. Because Head Start at that time covered the large majority of disadvantaged children in the United States, the evaluation had to rely on a comparison group of more advantaged children. However, the trajectories of disadvantaged children for developing cognitive, math, and language skills are known to be slower than those of more advantaged children. The evaluation used analysis of covariance, with the pretest measures of skills as the covariates, to adjust statistically for differences between the groups. Campbell and Erlebacher (1970) demonstrated how this analysis introduced a regression artifact: At posttest, each sample regressed back to its respective group mean, making the outcomes for the Head Start group look smaller than they probably were. Subsequent meta-analyses of Head Start outcome studies supported the assertion that there were a variety of program benefits (McGroder, 1990).

A newly completed randomized experiment has examined outcomes across nationally representative populations and Head Start centers, comparing 3-year-old versus 4-year-old participants, and 1 versus 2 years in Head Start. The evaluation reported improved cognitive, health, and parenting outcomes for 4-year-olds and better socioemotional outcomes for the 3-year-olds. As in the Westinghouse study, however, the

cognitive and language effects dissipated by first and second grades (Westat, Incorporated, 2010). Most importantly, the study was able to identify subgroups of children who were able to benefit more substantially from Head Start. One might argue that the consequences of the Westinghouse evaluation were not so grave after all because this new evaluation would seem to confirm the original findings. On the contrary, it would have been useful to have more definitive information 40 years ago, regardless of the policy direction in which it took the country.

Although the consequences of a mistaken conclusion can sometimes be drastic, it is more common that poor quality evaluations will simply add no information value. In these cases, there is just not much in the report to be used at all. I am collaborating in an ongoing review of sample evaluation reports to assess methods quality, using criteria adapted from several federal and international agencies (Robert Wood Johnson Foundation, 2009). As of this writing, there are over 200 reports from over 150 evaluation firms and individuals. The evaluations were conducted on behalf of international, federal, state, and local agencies in education, health, and mental health, nonprofit organizations, and philanthropies such as the United Way and private foundations. The review has identified useful evaluation reports that have used high quality research methods using our criteria. To date, however, about one quarter of the reports did not meet those criteria. Most often, methods received such a minimal description that their quality could not be assessed. Equally notable in these problematic reports, and even in many technically superior ones, was a lack of useful information. Witness the following conclusion, which I have seen in several evaluation reports: "A lesson learned was that the demonstration program was too short and there was not enough money to conduct the work." How can the reader possibly learn from this "lesson?" Clearly, stakeholders do not learn from it, given that this statement is repeated time and time again across program evaluations on the same topic. The lesson cannot be learned because it contains no guidance for the future. Should we not expect at least some indication about the length of time that would be required to complete a thorough evaluation, and some indication of the appropriate amount of funding to permit implementation of the program, or measurement of its outcomes? One does not want to draw evaluation conclusions beyond the data, but surely there was some hint, at least some concrete observations, on which to provide more useful guidance to the client. Alternatively, this "lesson" might even be that evaluators themselves should have anticipated a useless evaluation in the first place based on the inadequacy of funds and length of demonstration.

In summary, these four ties to ethics distinguish program evaluation from many other forms of behavioral and social research, and the high stakes involved in program evaluation make the ethical imperatives more urgent. At the same time, achieving methodological quality in evaluation

is often a challenge because the subjects are embedded in complex open systems, making controlled conditions difficult or impossible and introducing substantial noise into both treatments and data collection. Topping off the list of challenges, program evaluation is still emerging as a field, and for this reason it uses pluralistic criteria for what constitutes high quality methods. However, some consensus does exist, as seen in the literature linking ethics to evaluation practice.

## Prior Literature Linking Ethics to Evaluation

The field of evaluation has historically recognized some of the ethical challenges it faces. However, there are noteworthy limits on our consensus about those challenges. The Joint Committee on Standards for Educational Evaluation (1994) is a collaboration of 12 professional organizations formed in 1975 and accredited by the American National Standards Institute in 1989. In 1981 this committee produced the *Program Evaluation Standards* and revised them in 1994. Membership includes the American Psychological Association, American Evaluation Association, American Educational Research Association, and several other educational and counseling associations. Although ethical considerations certainly pertain in many of the Joint Standards, several standards are explicit about the ethical dimension under the rubric "Propriety": The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

> P3 Rights of Human Subjects: Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects.
>
> P4 Human Interactions: Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation, so that participants are not threatened or harmed.
>
> P5 Complete and Fair Assessment: The evaluation should be complete and fair in its examination and recording of strengths and weaknesses of the program being evaluated, so that strengths can be built on and problem areas addressed (Joint Committee on Standards for Educational Evaluation, 1994, inside and back covers).

By contrast, the American Evaluation Association publishes the Guiding Principles for Evaluators in each issue of journals published by

the Association. Unlike standards, guiding principles are not a requirement, and indeed their purpose is stated to be an ongoing effort to improve the ethics of evaluation practice without being premature about prescriptions. The Guiding Principles can be found at http://www.eval.org/publications/guidingprinciples.asp (American Evaluation Association, 2004). The Association's Ethics Committee first codified these principles in 1994, and they were revised in 2003. A good case book edited by Morris (2007) illustrates the principles in action. The principles explicitly recognize that methodological quality has ethical consequences, although they also cover other ethical considerations such as "Respect for People" (Section D) and "Responsibilities for General and Public Welfare" (Section E). Here are some examples of Guiding Principles pertaining to methods quality:

> Section A, Systematic Inquiry: Principle 3: Evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret, and critique their work. They should make clear the limitations of an evaluation and its results. Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses significantly affecting the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.
>
> Section B, Competence: Principle 3: Evaluators should practice within the limits of their professional training and competence, and should decline to conduct evaluations that fall substantially outside those limits. When declining the commission or request is not feasible or appropriate, evaluators should make clear any significant limitations on the evaluation that might result. Evaluators should make every effort to gain the competence directly or through the assistance of others who possess the required expertise.
>
> Section C, Integrity/Honesty: Principle 5: Evaluators should not misrepresent their procedures, data, or findings. Within reasonable limits, they should attempt to prevent or correct misuse of their work by others.

Many ethical concepts underlie the evaluation standards and guiding principles, but three concepts frame most of the issues: beneficence, autonomy, and social justice. I now turn to their definition and application.

## Definitions of Three Main Ethical Principles and Their Application to Evaluation

*Beneficence* is the ethical obligation to do good (just as the related principle of *nonmaleficence* is the obligation not to do harm) (Fisher, 2009). Beneficence justifies much of evaluation practice, if we presume that evaluation enables decision makers to allocate resources for the public good. Certainly, *formative* evaluation aims to do precisely that (i.e., the evaluation helps to form or improve the program at inception or in the early stages of its implementation; Scriven, 1991). However, there is a growing sentiment among program managers and advocates that *summative* evaluation (evaluation that sums up the achievements of the program) is problematic because the potential stakes are so high in terms of accountability to funders, and many state and local programs have such limited resources and staff expertise. Evaluation costs time and money that program staff could otherwise allocate to services, and it represents a major response burden to them. The likely evaluation findings should justify the cost, a point that has been made repeatedly and constantly ignored, as evidenced by a general failure to think through the strategic evaluation questions or to use findings (Cronbach et al., 1980; Leviton, Kettel-Khan, Rog, Dawkins, & Cotton, 2010; Shadish et al., 1991; Wholey, 2004).

In particular, some advocates perceive that outcome evaluation as conventionally practiced is the enemy of beneficence because of the rigidity of its standards for what constitutes evidence (e.g., Schorr, 2009). Evaluators who oppose these standards have proposed alternatives to experimentation or quasi-experimentation, such as using qualitative methods, relying on a theory of change, or using a phenomenological approach (Guba & Lincoln, 1989; Patton, 2004). Of course, others would argue that beneficence is at the core of rigorous evaluation, in light of the frequency with which less well-controlled designs produce overly optimistic results (Gilbert, McPeek, & Mosteller, 1977) or are not sensitive enough to detect positive outcomes that may be present (Wilson & Lipsey, 2001). Without getting into this epistemological and ideological fray, we can merely worry along with Campbell (1978) that any single way of knowing is flawed, so there is a place for methodological pluralism in program evaluation. Along these lines, many contemporary evaluators therefore prefer to triangulate by using multiple methods and measures, for example, using quantitative and qualitative methods together when possible.

*Autonomy* "refers to the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own" and not to be manipulated by external forces (Christman, 2009, p. 1). At the core of autonomy is self-rule, which presumes that individuals and groups have

the competence to judge for themselves what is best for them. *Paternalism* is interference with people's autonomy for their own good and is often justified by a belief that the people in question are not competent to judge for themselves what is beneficial to them. Challenges to the principle of autonomy can occur in at least three ways. An evaluation usually has several stakeholders (Preskill & Jones, 2009). Upholding the ethical principle of autonomy in evaluation requires that the variety of stakeholders be consulted about the questions to be asked. They may have materially important information to offer or the results may affect them in important ways. Stakeholders may sometimes have diverging views about what the program is supposed to accomplish, which can enrich or contextualize evaluation goals (Wholey, 2004). A failure to consult stakeholder groups affects their autonomy, yet this failure occurs frequently. It arguably affects beneficence as well, if stakeholder input would make a difference to improve programs.

The principle of autonomy is also challenged to the extent that evaluation planning, methods, and findings are not understandable and transparent to the program staff and other potential stakeholders. The evaluation field is rife with jargon terms that are unfamiliar to many social scientists, let alone managers and service clients. The use of such jargon may not be avoidable in some circumstances, but it limits people's ability to decide for themselves whether the evaluation is an accurate and fair reflection of program reality.

The client for evaluation is often the funder, so there is a power differential between evaluator and evaluated that affects programs' autonomy. This power differential exists whether the evaluation is conducted by the organization implementing the program (internal evaluation) or by an independent third party (external evaluation). At a minimum, an insistence on evaluation with the funder as the primary client makes program managers uneasy, and it can sometimes represent a major threat to their autonomy.

*Social justice* and fairness are at issue in the way that goods and services are distributed in society. Those who emphasize social justice argue that institutions should be structured to give the most benefit to the least advantaged people (Rawls, 2005). Social justice is at issue in evaluation in several ways. Many programs have redistributive and compensatory aims; evaluations may help decide their fate or contribute to their improvement. Social justice also presents a special case of the three challenges to autonomy: consulting stakeholders, transparency of evaluations, and power differential. Advocates for disadvantaged groups complain that program recipients and consumers are rarely consulted about evaluations of the programs that affect them—yet from a social justice perspective, they are the most interested parties and should be the primary evaluation clients (House & Howe, 1999).

Some stakeholders for disadvantaged groups are policy savvy—but their constituents are not, so transparency is especially important. All the more reason, therefore, to present in plain language the plans for evaluation, the methods to be used, the measures to be collected, the findings, and their implications. It is fundamentally unfair to use evaluation jargon when there is unequal access to the very language in which the evaluations are couched. It is entirely feasible to communicate in plain, lay terms. Examples of clear, transparent, nontechnical language in program evaluation can be found for human subjects review (Rivera et al., 2004), planning (Telfair & Leviton, 1999; W. K. Kellogg Foundation, 2004), evaluation design (SRI International, 2000), and communication of findings (Grob, 2004). Evaluators may not find it easy to communicate clearly; they may believe they do not have the time, or the primary client may not see it as a priority. However, to fulfill the principle of social justice, it is an ethical priority to do so when the programs and policies in question affect disadvantaged groups.

The power differential in evaluation reaches its most acute manifestation in programs for disadvantaged groups because the funding for such programs is most often in the hands of advantaged others. This dynamic is illustrated most graphically for grassroots programs, which tend to have relatively low resources, financial vulnerability, undifferentiated organizational structure, and fewer trained staff for many activities, including evaluation (Schuh & Leviton, 2006). The result is that funders are more likely to dictate evaluation questions and to impose outside evaluators on such organizations—simultaneously violating the autonomy principle and potentially impeding the social justice principle. The problems are compounded over time when there are (a) knee-jerk accountability requirements without a consideration of the strategic evaluation questions to be asked; (b) less-than-optimal resources and staff, making program startup longer and impairing implementation; (c) a lack of clarity on how the program is supposed to work; (d) premature evaluation of outcomes in an unrealistic timeframe; and (e) poor practice of evaluation. Evaluators who undertake assignments for programs that serve disadvantaged groups need to be mindful of social justice issues—particularly, the dynamics of the power differential between funder and program. One method that has become popular to address the challenges of social justice is *participatory evaluation* (Whitmore, 1998). Although participatory evaluation can address relevance, realism, buy-in, and transparency, it does not per se guarantee good design or analytic quality. These are different ethical considerations.

These sections serve as background on the general ethical challenges in evaluation. I now turn to the specific challenges that are presented in the production process for a quantitative evaluation.

## Ethical Issues Arising in Each Stage of the Production of Quantitative Evaluations

Once the decision is made to evaluate, each stage in the production of evaluations raises issues about methods quality. I will focus on a select few topics per evaluation stage where decision making about methods can raise ethical challenges.

### Stage 1: Consulting Stakeholders to Improve Quality and Serve Ethics

Many issues at the planning stage have ethical implications, given that quality has ethical implications. I will focus, however, on the distinctive ethical imperative to consult stakeholders. This activity may seem far removed from the methods quality of a quantitative evaluation. However, an example illustrates how consulting stakeholders can culminate in a high quality study with an ethical strategic aim. The problem being addressed is that frail elders and disabled people need personal care services, such as bathing, dressing, toileting, and feeding. They would generally prefer to choose the person who provides these services themselves. However, Medicaid pays for much of the personal care service provided in the United States, and Medicaid policy was to pay a social service agency to send a stranger to the home to provide personal care services. Working with disabled people and frail elders, their advocates, and interested health care professionals, the Robert Wood Johnson Foundation supported development of models that allowed disabled people to select their own caregivers. One model was the Cash and Counseling program: Disabled people received guidance on the selection of caregivers and a flexible budget to pay for their services. A randomized experiment of Cash and Counseling concluded that participants who were able to choose their own caregivers reported better quality of life, and there were no increased costs to Medicaid or any indication of increased fraud and abuse (Brown et al., 2007). Such "cost neutrality" is the magic phrase for health care reimbursement. Based on this study, 15 states have adopted Medicaid policies to pay for Cash and Counseling, and an additional 15 have adopted similar models. The randomized experiment was the key: Medicaid has regularly changed reimbursement policies to accommodate superior models that are cost neutral.

It is sometimes alleged that participation in decision making by service recipients and their advocates will impair the quality of evaluation studies. However, the Cash and Counseling evaluation had quantitative methods that are deemed excellent: careful random assignment of participants to conditions; implementation according to a specific service model with good quality controls; careful measurement of cost, quality of life,

and services; and appropriate analysis of outcomes and cost-effectiveness. Discussions with stakeholders allowed the issues and models to evolve over the years, which may have sharpened the focus to permit a rigorous study to convince Medicaid. It is also sometimes alleged that random assignment is somehow unethical. When random assignment compares "usual care" with an unproven innovation, it is unclear how this can be the case. On the contrary, the strategic need to persuade policy makers made random assignment the most ethical choice.

### Stage 2: Ethical Issues Involved in Measurement and Design

A number of ethical issues arise in the measurement and design of program evaluations. Because evaluators are preoccupied with fairness in evaluation, one key issue in planning is to distribute equal focus on minimizing both Type I and Type II error. Minimizing Type I error has long been a major focus in evaluation (Shadish, Cook, & Campbell, 2001). The consequences of Type I error are that a program is declared successful in improving outcomes when in fact it did not do so. This represents an opportunity cost for society: Public funds are expended on ineffective programs, and people get ineffective services when they might receive effective ones. Less attention is given to minimizing Type II error. Evaluations are prone to flaws that make Type II error likely (Wilson & Lipsey, 2001) for many reasons. In evaluation, Type II error has consequences that are at least as important from an ethical viewpoint. A Type II error means declaring that a program is unsuccessful when in fact it is successful. The consequence can be a chilling effect on innovation and the perception that "nothing works" (Weiss, 1987). This perception has hampered progress in education and criminal justice, and now we find that later studies indicate something does work (e.g., Committee on Community Supervision and Desistance From Crime, 2007). At a minimum, program development and improvement activities may be undertaken to fix something that is not broken. These represent opportunity costs for society as well, and they certainly challenge the ethical principles of beneficence and fairness.

A frequent Type II error problem in evaluation is a lack of statistical power to detect outcomes at conventional significance levels. Pilot testing, literature review, or at least serious discussions with program staff can establish likely effect sizes. Discussions with funders and policy makers can establish in advance the size of effect that is practically or clinically significant. Data collection is expensive, and funders may balk at the sample size needed to assess change. The ethical evaluator should explain the need for the sample in detail and explore options for assessment, including not conducting an evaluation that is sure to set up the program for failure. Funders and program managers understand evaluators' preoccupation for "not shooting the program in the foot."

Beyond concerns to minimize both Type I and Type II error, best practice measurement in program evaluation needs to be sensitive to detect changes where they exist. From the meta-analysis by Wilson and Lipsey (2001), we know that unreliable or invalid measurement is regularly associated with smaller effect sizes in quantitative evaluation studies. Evaluators should search for existing measures with satisfactory *reliability* and *validity* (see Carrig & Hoyle, Chapter 5, this volume, for discussion of these terms). In some fields, however, good measures may not exist yet, or measures may have been validated on populations different from the ones that the program is designed to serve. I faced this challenge in an early evaluation of HIV prevention in gay and bisexual men (Leviton et al., 1990; Valdiserri et al., 1989). There was little prior research on the reliability and validity of gay and bisexual men's self-reported HIV knowledge, attitudes, and risk behaviors, although research was emerging during the same period (Catania, Gibson, Chitwood, & Coates, 1990). We developed and pilot-tested multiple measures derived from available marketing research on gay men's preferences and attitudes. Given the logistics of the situation, however, we had to assess the measures' psychometric properties at the same time we were using the measures to assess outcomes. Fortunately, the effects of the intervention were robust, and we encountered few problems in the measures.

### Stage 3: Ethical Issues Involved in Quantitative Data Collection

Our preoccupation with Type II error means that there is an ethical imperative to obtain high quality data. Without good protocols and management and supervision of data collection staff, even the best measures and design will only produce data that are fraught with error, jeopardizing the chances to observe real change when it occurs. It is astonishing how little attention is paid to these issues in the literature on primary data collection (Nakashian, 2007). My own informal review of survey research texts found, for example, that few addressed the need for management and supervision of data collectors, despite their potential to introduce error and bias in the results. For example, one survey text entitled a "Comprehensive Guide" had, out of 304 pages, 3 to 4 pages on training of interviewers and 2 pages on their supervision (Rea & Parker, 2005). In contrast, Fowler (2001) spends 19 pages on this topic, and Lavrakas (1993) devotes half the book to training and supervision.

Quality control, in the form of protocols and supervision, is also a must for the collection of archival data such as administrative records. The data themselves can sometimes be of low quality, making them problematic as the primary data source. For the better kinds of archival records, my impression is that attention to quality control can be intense, for example, in medical records abstraction. The reason is that the problems within the

data themselves are well understood, and professionals recognize how error can be introduced at each step in the abstraction process (e.g., Reisch et al., 2003).

Protocols and supervision are necessary to maintain response rates as well. We are experiencing a crisis of lower response rates in primary data collection, especially with telephone surveys (American Association for Public Opinion Research, 2010). To gain an adequate response rate, as many as 20 call-backs may be needed, but research projects can rarely afford to do so. Low response rates reduce sample sizes and therefore power, but also may co-occur with selective responding patterns, which can bias evaluation results and also impair the credibility of findings. Incentives can improve response rates by acknowledging the fact that the investigator is taking participants' valuable time and are routinely considered best practice in school-based evaluations.

Protocols, supervision, and adequate time spent per data collection unit are labor-intensive activities. For this reason, data collection is often the most expensive component of an evaluation budget. Yet many researchers do not budget adequately for data collection, or they may experience setbacks in the field that increase the cost of data collection (Nakashian, 2007). At such times, one may be tempted to cut sample size, cut the number of call-backs for telephone surveys, or eliminate items to make the survey shorter. Yet these steps can increase the chances of Type II error or may introduce bias into the results. Sometimes in a competition among contractors to conduct an evaluation, I encounter a bid that is noticeably lower than others. Such a bid will sometimes lack quality safeguards, such as specifying the number of call-backs in a telephone survey or allowing adequate time per telephone interview, that are essential to avoiding Type II error. It takes experience to detect situations where the bid is lower because of a lack of quality safeguards.

### Stage 4: Ethical Issues Involved in Analysis

In evaluation, messy data sets are the norm. Missing data are a constant problem; distributions are not normal; and statistical assumptions are in danger of being violated right and left. As evaluators gain more experience with a particular type of program, many of these problems can be anticipated and controlled to some extent (Lipsey, 2000). The Guiding Principles specify that the evaluator will know what expertise is required and will keep up to date through formal coursework in graduate school, staying abreast of new methods, attendance at professional meetings, and discussion with peers.

A frequent analytic problem arising in program evaluation is that, in real world programs, participants are nested within settings such as neighborhoods or schools; therefore, hierarchical analysis is required (e.g.,

Goldstein, Chapter 13, this volume; Raudenbush & Bryk, 2002). Often, however, the program includes only a handful of these larger units or settings, so that appropriate analysis is a challenge; application of conventional hierarchical linear models may be impossible because of estimation problems or may produce unacceptably variable estimates (Murray, 1998). At a minimum, if such data are analyzed at the individual participant level, analysts will need to establish that the intraclass correlation is low enough that it would not inflate the significance of results.

Data may be nested within program sites, not just settings. Gathering data on across-site variations in program implementation is critical to help explain mixed results or a no-effect conclusion. Even when the overall effect is not significant or large, implementation can often explain variation in the outcomes. That is, program effects may be moderated by site characteristics. On a regular basis, monitoring implementation also leads to course corrections during an evaluation.

### Stage 5: Ethical Issues Involved in Interpretation and Reporting of Results

Clear presentations are the minimum requirement for ethical evaluations. Beyond clarity, interpretation of findings requires conservative reporting of statistically significant program effects. For example, programs may claim a wide variety of positive effects. In exploratory analysis the evaluator may encounter unintended side effects, both positive and negative. In either case, analysis of multiple outcomes needs to adjust the alpha level so there is no "fishing expedition" that would produce significant results by chance alone (see Hubert & Wainer, Chapter 4, this volume, for example procedures). Alternatively, findings from exploratory analysis might be confirmed using triangulation of methods and data sources, collection of additional data, or replicating the entire evaluation.

Beyond clarity and conservativeness, however, interpretation of findings requires framing them with contextual information, such as what else is known about the problem area (Ginsburg & Rhett, 2003; March, 2009). Humans have trouble processing numbers in isolation, and even if they did not, stand-alone conclusions are rarely appropriate. For example, how do effect sizes from the evaluation of the program compare with effect sizes in evaluations of similar programs? In what respects do mixed results represent a success or a failure? What are the concrete implications of the program for deaths and disability averted, quality of life, improvement in life chances, or social functioning? If evaluation was negative, why might this have been the case, and what can be done to improve the program? These questions appear regularly when there are negative results and can be answered at least in part by statistically and substantively exploring the implementation findings, the underlying rationale or

theory of the program, and the available research literature (Cronbach et al., 1980; Shadish et al., 1991).

Two examples support this point. Our early study of HIV/AIDS prevention in gay and bisexual men was initially presented to a mixed group of epidemiologists and infectious disease doctors (Valdiserri et al., 1989). Although they understood our statistical language regarding effect sizes, they also wanted to know what the results would concretely mean for infections and deaths avoided among the men. At that time, we did not have an answer. Since that time, a good deal more has been learned about the infectivity of various risk behaviors and the changes that would be required to stop or slow infections in a given population. The result is that a body of evidence is now available to more concretely frame these results for decision makers and potentially influence federal policy for a costly disease that continues to cripple large numbers of people (Pinkerton, Holtgrave, Leviton, Wagstaff, & Abramson, 1998).

When results are mixed, it is especially important for evaluators to participate actively in interpreting the results by offering context. A recent article describes the evaluation of a discharge planning and case management program for drug-addicted prisoners released from Rikers Island in New York City (Bunch, 2009). The program, funded by the Robert Wood Johnson Foundation, aimed to prevent crime recidivism, decrease the number of former prisoners using illicit drugs, and prevent risky behaviors that lead to the spread of HIV/AIDS. The evaluation was generally well designed and conducted (Needels, Stapulonis, Kovac, Burghardt, & James-Burdumy, 2004). According to the evaluation, the program did not succeed in its primary aims. However, several secondary aims had been also specified from the beginning, and these outcomes were positive: Former prisoners who received the services had increased participation in drug treatment programs; young men were more likely to get their GED; and women were more likely to receive gynecological services.

The Robert Wood Johnson Foundation initially framed this program as a failure. The program was ending in any case and would not be renewed, but its evaluation was viewed as disappointing. However, I argued that framing the results this way was mistaken because it did not take into account the body of knowledge about the problems involved. First, preventing recidivism from jail works under some conditions for some prisoners but is generally difficult, and our understanding of "what works" has greatly improved since the project was initiated (Committee on Community Supervision and Desistance From Crime, 2007). Case management for drug treatment does assist in linking addicts to care, although results are mixed for reduction in drug use (Hesse, Vanderplasschen, Rapp, Broekaert, & Fridell, 2007). Drug treatment tends to have modest effects at best, but more important in hindering recovery is the lack of access for poor

people and the generally low quality of drug treatment in the United States (Committee on Crossing the Quality Chasm: Adaptation to Mental Health and Addictive Disorders, 2006). Under these conditions, it was unrealistic to expect that limited discharge planning and case management services would be enough to prevent recidivism and reduce illicit drug use.

Once the context of the findings was better understood, the primary outcomes seemed less realistic, but the outcomes for the three secondary program aims became more interesting and important. Having a GED makes a difference regarding employment possibilities and rates for young men, and employment can ultimately prevent recidivism (U.S. Department of Education, 1998). Furthermore, the amount of time spent in drug treatment is correlated with treatment effectiveness (French, Zarkin, Hubbard, & Rachal, 1993; National Institute on Drug Abuse, 2009). Therefore, study participants were generally doing what was required in the long term to abstain from drugs, even if the short-term drug use results themselves were not significant. Finally, the women's greater use of gynecological services requires comment. These women were at major risk of HIV infection: (a) New York City has a very high prevalence of HIV infection; (b) they were returning from jail, where risky behaviors occur at a high prevalence with or without jailers' knowledge; (c) in or out of jail, it was therefore more likely these addicts would share drug injection equipment with HIV-infected individuals; and (d) if they had sex, it was therefore more likely to be with HIV-infected individuals. They could well become pregnant, and if so their babies would receive better prenatal care if they visited a gynecologist. Additionally, if the women became infected with HIV, early detection by a gynecologist could help prevent transmission to the infant, not to mention ensuring a longer life span for the mother. This result is of primary importance as a public health issue, so it should not be dismissed so easily just because it was a secondary program aim.

This case example illustrated that when results are mixed, it is exceptionally important from an ethical standpoint to interpret the results in context. The program at Rikers Island did not achieve everything that we would wish, but it did add to a growing body of information about prisoner reentry that the Robert Wood Johnson Foundation, as well as the country, is now pursuing (Bunch, 2009).

## Distinguishing Ethical, Technical, and Ideological Issues in Program Evaluation

As illustrated in this chapter, program evaluation is often a complex endeavor with ethical challenges at each stage of its production. Yet those

ethical challenges are often addressed in terms that are either technical or ideological—without explicit reference to the classic ethical principles that underlie the choices that must be made at each stage. That is, the evaluation field often conflates ethical, technical, and ideological considerations when separating these dimensions of the problem might be helpful. In concluding this chapter, it seems useful to draw some distinctions between ethical issues and technical competence, on the one hand, and ideology, on the other.

## Distinguishing Ethical Conduct and Evaluation Quality

Throughout this book, the authors illustrate how methodological quality raises ethical considerations because of the grave consequences for progress and understanding in many fields of inquiry. In cases such as evaluation or medicine (see Fidler, Chapter 17, this volume), there is great potential for immediate good or harm to society. In general, I endorse the idea that gatekeepers and safeguards for ethical conduct and methods quality should be brought closer together (Sterba, 2006). In evaluation, however, we need to tread with great care in doing so. Politics, fear, and disappointment over negative results can obscure what might be meant by "ethical" or "unethical" evaluations. Evaluation is not a pleasant profession. Evaluators are regularly fired or sued for telling clients what they do not want to hear (Leviton, 2001). A lack of consensus about evaluation methods exacerbates the situation, although the standards and guiding principles offer at least some protection. By discussing methods primarily on their technical and conceptual merits, I believe we can preserve rational discourse about the studies themselves. This is not to ignore the ethical considerations, but to keep them from being obfuscated by passion or deliberate manipulation. Outside the evaluation field, disagreements over methods have led to accusations of scientific misconduct that had no merit, with severe consequences for some courageous individuals (e.g., Denworth, 2009; Knoll, 1994).

## Distinguishing Evaluation Quality From Technical Quality

Evaluation methods are not a purely technical matter, however. A spirit of inquiry, transparency of procedures and reasoning, and a healthy self-doubt are not technical matters. However, methods quality is often framed in purely technical terms. This is especially true of novices in a field of inquiry, for whom rules-based procedures tend to predominate (Dreyfus & Dreyfus, 1992). More experienced evaluators generally move beyond "cookbook" approaches, adapting methods to the context at hand.

Reasoning and inquiry play a larger role as expertise is gained, for example, in the use and interpretation of quasi-experimental designs to rule out alternative explanations for findings (Shadish et al., 2001). In fact, the tendency in evaluation to rely on technical justification might be ascribed in part to the prevalence of novices and merely competent evaluators. People tend to move in and out of the evaluation field, although the number of experienced professionals is growing. Certainly, the users of evaluation tend to be unfamiliar with both the technical issues and the modes of reasoning that are needed.

### Ethics Versus Ideology

Evaluation is an inherently political endeavor (Weiss, 1998), and so political ideologies can obscure the application of ethical principles. Some ideologies are expressed tacitly, as though the people involved had never examined their assumptions. For example, programs that challenge the status quo, redistribute resources to the poor, or attempt to address disparities in health and well-being are more often subjected to evaluation than are more established practices, programs, and policies that benefit entrenched interests (Weiss, 1987). Thus, Social Security is evaluated less frequently than welfare programs or food stamps, for example, and is less subject to deep revisions or potential cuts. Bank regulation was rarely evaluated before the economic recession of 2007. The increased accountability requirements for redistributive programs relate to long-standing paternalistic attitudes about the poor and about public responsibility for them. In all fairness, it also arises from analysis of the perverse incentives that social welfare policies can introduce (Trattner, 1999).

From the political left come accusations that most evaluations are undemocratic because they did not actively consult the wishes and views of the service recipients (House & Howe, 1999). This accusation reflects an ideology that assumes the only democratic process is direct democracy and that social justice is the highest value of all values for evaluation. This is not entirely reasonable or even realistic (Shadish & Leviton, 2001). Representative democracy is also a legitimate form and more workable than direct democracies. Democratic societies have legitimated many mechanisms for the people's elected and appointed representatives to make decisions on their behalf. Thus, evaluations that do not actively consult all of those most affected cannot be said to be undemocratic. However, it is true that disadvantaged groups are marginalized even in democratic societies, as discussed earlier, so decision makers are not necessarily acting in the interest of these groups. Consulting with representatives of the affected groups is consistent with best practice in evaluation and reduces the danger of paternalism.

## Conclusion

With all the potential for inflamed discussions about evaluation results, with the pluralistic methods of the field, high-stakes studies, and many pitfalls, outsiders may wonder whether ethical conduct of program evaluation is feasible. Might it be better to leave programs and policies unexamined, if the results would be so open to debate and the potential consequences so severe? On the contrary, my conclusion is that more and better experience with these issues is essential because the public interest dictates that we cannot leave policies and programs unexamined. Discussion, transparency, and more case studies will help to guide the field into better and more ethical conduct of evaluations.

## References

American Association for Public Opinion Research. (2010). Response rates: An overview. Retrieved from http://www.aapor.org/response_rates_an_overview.htm

American Evaluation Association. (2004). Guiding principles for evaluators. Retrieved from http://www.eval.org/publications/guidingprinciples.asp

Brown, R., Lepidus Carlson, B., Dale, S., Foster, L., Phillips, B., & Schore, J. (2007). *Cash & counseling: Improving the lives of Medicaid beneficiaries who need personal care or home and community-based services*. Princeton, NJ: Mathematica Policy Research, Inc.

Bunch, W. (2009). Helping former prisoners reenter society: The Health Link Project. In S. L. Isaacs & D. C. Colby (Eds.), *To improve health and health care, Vol. XII: The Robert Wood Johnson Foundation Anthology* (pp. 165–184). San Francisco: Jossey-Bass.

Campbell, D. T. (1978). Qualitative knowing in action research. In M. Brenner, P. Marsh, & M. Brenner (Eds.), *The social context of methods* (pp. 184–209). London: Croom Helm.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluation can mistakenly make compensatory evaluation look harmful. In J. Hellmuch (Ed.), *Disadvantaged child, Vol. III: Compensatory education: A national debate* (pp. 185–210). New York: Brunner/Mazel.

Carman, J. G., & Fredericks, K. A. (2008). Nonprofits and evaluation: Empirical evidence from the field. *New Directions in Evaluation, 119,* 51–72.

Catania, J. A., Gibson, D. R., Chitwood, D. D., & Coates, T. J. (1990). Methodological problems in AIDS research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin, 108,* 339–362.

Christman, J. (2009). Autonomy in moral and political philosophy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/fall2009/entries/autonomy-moral

Committee on Community Supervision and Desistance From Crime. (2007). *Parole, desistance from crime, and community integration.* Washington, DC: National Academies Press.

Committee on Crossing the Quality Chasm: Adaptation to Mental Health and Addictive Disorders. (2006). *Improving the quality of health care for mental and substance-use conditions: Quality chasm series.* Washington, DC: National Academies Press.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56,* 331–364.

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

Denworth, L. (2009). *Toxic truth: A scientist, a doctor, and the battle over lead*. Boston: Beacon Press.

Dreyfus, H. L., & Dreyfus, S. E. (1992). *Mind over machine.* New York: Free Press.

Fisher, C. F. (2009). *Decoding the ethics code: A practical guide for psychologists* (2nd ed.). Thousand Oaks, CA: Sage.

Fowler, F. (2001). *Survey research methods.* Thousand Oaks, CA: Sage.

French, M. T., Zarkin, G. A., Hubbard, R. L., & Rachal, J. V. (1993). The effects of time in drug abuse treatment and employment on posttreatment drug use and criminal activity. *American Journal of Drug and Alcohol Abuse, 19,* 9–33.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. *Science, 198,* 684–689.

Ginsburg, A., & Rhett, N. (2003). Building a better body of evidence: New opportunities to strengthen evaluation utilization. *American Journal of Evaluation, 24,* 489–498.

Grob, G. (2004). Writing for impact. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 604–627). San Francisco: Jossy-Bass.

Guba, E. G., & Lincoln, Y. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.

Hesse, M., Vanderplasschen, W., Rapp, R., Broekaert, E., & Fridell, M. (2007). Case management for persons with substance use disorders. *Cochrane Database of Systematic Reviews, 4,* CD006265. doi: 10.1002/14651858.CD006265.pub2. Retrieved from http://www.cochrane.org/reviews/en/ab006265.html

House, E. R., & Howe, K. R. (1999). *Values in evaluation.* Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.

Knoll, E. (1994). What is scientific misconduct? *Science Communication, 14,* 174–180.

Lavrakas, P. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Thousand Oaks, CA: Sage.

Leviton, L. C. (2001). Building evaluation's collective capacity: American Evaluation Association Presidential address. *American Journal of Evaluation, 22,* 1–12.

Leviton, L. C., & Boruch, R. F. (1983). Contributions of evaluation to education programs and policy. *Evaluation Review, 7,* 563–598.

Leviton, L. C., Kettel-Khan, L., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability assessment to improve public health. *Annual Review of Public Health*, *31,* 213–233.

Leviton, L. C., Valdiserri, R. O., Lyter, D. W., Callahan, C. M., Kingsley, L. A., & Rinaldo, C. R. (1990). Preventing HIV infection in gay and bisexual men: Experimental evaluation of attitude change from two risk reduction interventions. *AIDS Education and Prevention, 2,* 95–109.

Lipsey, M. W. (1988). Practice and malpractice in evaluation research. *Evaluation Practice, 9,* 5–25.

Lipsey, M. W. (2000). Meta-analysis and the learning curve in evaluation practice. *American Journal of Evaluation, 21,* 207–212.

March, J. (2009). *A primer on decision making: How decisions happen*. New York: Free Press.

McGroder, S. M. (1990). *Head Start: What do we know about what works?* Washington, DC: U.S. Department of Health and Human Services. Retrieved from http://aspe.hhs.gov/daltcp/reports/headstar.htm

Morris, M. (2007). *Evaluation ethics for best practice: Cases and commentaries*. New York: Guilford.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Nakashian, M. (2007). *A guide to strengthening and managing research grants.* Princeton, NJ: The Robert Wood Johnson Foundation. Retrieved from http://www.rwjf.org/files/research/granteeresearchguide.pdf

National Institute on Drug Abuse. (2009). *NIDA InfoFacts: Treatment approaches for drug addiction.* Retrieved from http://www.nida.nih.gov/infofacts/treatmeth.html

Needels, K., Stapulonis, R. A., Kovac, M. D., Burghardt, J., & James-Burdumy, S. (2004). *The evaluation of Health Link: The community reintegration model to reduce substance abuse among jail inmates. Technical report.* Princeton, NJ: Mathematica Policy Research.

Patton, M. Q. (2004). *Qualitative research and evaluation methods.* Thousand Oaks, CA: Sage.

Pinkerton, S. D., Holtgrave, D. R., Leviton, L. C., Wagstaff, D. A., & Abramson, P. R. (1998). Model-based evaluation of HIV prevention interventions. *Evaluation Review*, *22*, 155–174.

Preskill, H., & Jones, N. (2009). *A practical guide for engaging stakeholders in developing evaluation questions.* Boston: FSG Social Impact Advisors. Retrieved from http://www.rwjf.org/files/research/49951.stakeholders.final.1.pdf

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rawls, J. (2005). *A theory of justice: Original edition*. Cambridge, MA: Belknap.

Rea, L. M., & Parker, R. A. (2005). *Designing and conducting survey research: A comprehensive guide.* San Francisco: Jossey-Bass.

Reisch, L. M., Fosse, J. S., Beverly, K., Yu, O., Barlow, W. E., Harris, E. L., … Elmore, J. G. (2003). Training, quality assurance, and assessment of medical record abstraction in a multisite study. *American Journal of Epidemiology, 157,* 546–551.

Rivera, R., Borasky, D., Carayon, F., Rice, R., Kirkendale, S., Wilson, W. L., & Woodsong, C. (2004). *Research ethics training curriculum for community representatives*. Research Triangle Park, NC: Family Health International. Retrieved from http://www.fhi.org/en/rh/training/trainmat/ethicscurr/retccren/index.htm

Robert Wood Johnson Foundation. (2009). Guidance on evaluation reports to the Robert Wood Johnson Foundation: A checklist for evaluators. Retrieved from http://www.rwjf.org/files/research/50349.quality.checklist.final.pdf

Schorr, L. B. (2009). To judge what will best help society's neediest, let's use a broad array of evaluation techniques. *Chronicle of Philanthropy.* Retrieved from http://philanthropy.com/article/To-Judge-What-Will-Best-Help/57351

Schuh, R. G., & Leviton, L. C. (2006). A framework to assess the development and capacity of nonprofit agencies. *Evaluation and Program Planning, 29,* 171–179.

Scriven, M. (1991). *The evaluation thesaurus*. Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theorists and their theories.* Newbury Park, CA: Sage.

Shadish, W. R., & Leviton, L. C. (2001). Descriptive values and social justice. In A. Benson, D. M. Hinn, & C. Lloyd (Eds.), *Visions of quality: How evaluators define, understand, and represent program quality.* Oxford, UK: JAI Press.

SRI International. (2000). *We did it ourselves: An evaluation guidebook.* Sacramento, CA: Sierra Health Foundation.

Sterba, S. K. (2006). Misconduct in the analysis and reporting of data: Bridging methodological and ethical agendas for change. *Ethics & Behavior, 16,* 305–318.

Telfair, J., & Leviton, L. C. (1999). The community as client: Improving the prospects for useful evaluation findings. Chapter 1 of *Evaluation of health and human services programs in community settings. New Directions in Program Evaluation, 1999*(83), 5–16.

Trattner, W. I. (1999). *From poor law to welfare state: A history of social welfare in America.* New York: Simon & Schuster.

U.S. Department of Education. (1998). Educational and labor market performance of GED recipients. Retrieved from http://www2.ed.gov/pubs/ged/lmpogr.html

Valdiserri, R. O., Lyter, D. W., Leviton, L. C., Callahan, C. M., Kingsley, L. A., & Rinaldo, C. R. (1989). AIDS prevention in gay and bisexual men: Results of a randomized trial evaluating two risk reduction interventions. *AIDS, 3,* 21–26.

Weiss, C. H. (1987). Evaluating social programs: What have we learned? *Society, 25,* 40–45.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Weiss, C. H., & Bucuvalas, M. J. (1980). *Social science research and decision-making.* New York: Columbia University Press.

Westat, Incorporated. (2010). *Head Start impact study: Final report.* Rockville, MD: Author. Retrieved from http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_final.pdf

Whitmore, E. (Ed.). (1998). Understanding and practicing participatory evaluation. *New Directions for Evaluation, 1998*(80).

Wholey, J. S. (2004). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.). *Handbook of practical program evaluation* (pp. 33–62). San Francisco: Jossey-Bass Publishers.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychological Methods*, *6*, 413–429.

W. K. Kellogg Foundation. (2004). *Logic model development guide*. Battle Creek, MI: Author. Retrieved from http://ww2.wkkf.org/DesktopModules/WKF.00_DmaSupport/ViewDoc.aspx?fld=PDFFile&CID=281&ListID=28&ItemID=2813669&LanguageID=0

# Section IV

# Ethics and Data Analysis Issues

# 10

# *Beyond Treating Complex Sampling Designs as Simple Random Samples: Data Analysis and Reporting*

**Sonya K. Sterba**
*Vanderbilt University*

**Sharon L. Christ**
*Purdue University*

**Mitchell J. Prinstein**
*University of North Carolina at Chapel Hill*

**Matthew K. Nock**
*Harvard University*

This chapter addresses two issues: (a) how the method for selecting the sample ought to be reported in observational research studies, and (b) whether and when the sample selection method needs to be accounted for in data analysis. This chapter reviews available methodological and ethical guidelines concerning each issue and considers the extent to which these recommendations are heeded in observational psychological research. Discussion focuses on potential ethical implications of the gap between available methodological recommendations and current practice. A hypothetical case example and also a real world case example involving a daily diary study are used to demonstrate some alternative strategies for narrowing this gap.

It is important to note that both of the issues taken up in this chapter (reporting and accounting for sample selection in data analysis) arise after the sampling method has already been chosen. In contrast, a chapter on ethics and sampling in observational studies might have been expected to mainly concern the sample selection method itself—particularly whether a random (probability) or nonrandom (nonprobability) sample should

be drawn.[1] The latter topic has long dominated informal discussions of ethics and sampling among social scientists, but has also often been misunderstood. Moreover, debate over choosing between probability versus nonprobability sampling has often led to an impasse, where observational researchers in particular fields (e.g., psychology) find only one sampling method pragmatically feasible (nonprobability sampling), and other fields (e.g., public health) find only one method statistically defensible (probability sampling; see Sterba, 2009). Our strategy is to begin with a brief overview of current and past perspectives on this controversial topic. The issues we address in this chapter are very general; they are relevant to whatever (probability or nonprobability) sample was selected. However, in discussing these issues in later sections, we periodically highlight relevant costs or benefits of using a probability versus nonprobability sampling method.

## Random and Nonrandom Sample Selection

When sampling was first proposed as an alternative to census taking, a distinction was drawn between two different methods for selecting samples from populations: probability (or random) sampling and nonprobability (or nonrandom) sampling (Bowley, 1906; Kaier, 1895). In probability sampling, the probability of selection for all units in the target population is known and nonzero. In nonprobability sampling, the probability of selection for some units is unknown, and possibly zero, and the finite, target population may be only loosely identified. Whereas early methodological debates sought to establish one method as superior and the other as uniformly unacceptable (Neyman, 1934; Stephan, 1948), such definitive conclusions were never reached despite extensive dialogues on the topic (see Royall & Herson, 1973; Smith, 1983, 1994; Sugden & Smith, 1984).

To summarize this debate briefly, collecting a probability sample by definition requires that key selection variables are observed and that the *selection mechanism* (i.e., the mechanism by which sampling units get from a finite population into the observed sample) is well understood. Both aspects in turn reduce the risk that selection on unmeasured, unobserved variables will bias results. Furthermore, the randomness entailed

[1] Note that the issues that arise when deciding between random versus nonrandom *assignment* in treatment settings (e.g., Mark & Lenz-Watson, Chapter 7, this volume) are meaningfully different from those that arise when deciding between random versus nonrandom *selection* in observational (or experimental) settings, although there are certain parallels (Fienberg & Tanur, 1987).

in a probability selection mechanism—specifically the fact that sampled and unsampled outcomes are assigned known probabilities—means that a distribution constructed from these probabilities can serve as the sole basis of inference to a finite population, without invoking strong modeling assumptions (e.g., Cassel, Sarndal, & Wretman, 1977). In contrast, nonprobability samples rely heavily on modeling assumptions to facilitate inference to a larger population, which is hypothetical. Nevertheless, should these modeling assumptions be met, there is a well-established statistical logic for inference from nonprobability samples (see Sterba, 2009, for a review of this logic). Hence both sampling methods have been recognized—initially at the 1903 Consensus Resolution of the International Statistical Institutes—and both are still frequently used.[2] Much attention has since turned to the two issues considered here: (a) what to report about sample selection, and (b) whether and when to account for sample selection in data analysis.

## Reporting About Sample Selection

### Methodological Guidelines

For the issue of reporting about sample selection, our review necessarily takes a historical perspective because reporting guidelines have been in existence for a long time, yet have evolved considerably. The first methodological recommendations on reporting practices appeared almost immediately after the practice of sampling was first introduced. The International Statistical Institute's 1903 Consensus Resolution called for "explicit account in detail of the method of selecting the sample" in research reports (Kish, 1996, p. 8). Similar recommendations were made in the proceedings of subsequent meetings, such as: "the universe from which the selection is made must be defined with the utmost rigour," and "exactness of definition" is needed for "rules of selection" (Jensen, 1926, pp. 62–63). The nonspecificity of these guidelines, however, led to inconsistent reporting practices.

By the 1940s, mounting dissatisfaction over inconsistent reporting practices led the United Nations (UN) Economic and Social Council to convene a Subcommission on Statistical Sampling that met throughout the decade to develop a common terminology for such reporting (UN, 1946, 1947, 1948, 1949a). This Subcommission resulted in the formalized

[2] This chapter pays specific attention to nonprobability (nonrandom) samples because they are most often used by psychologists.

"Recommendations Concerning the Preparation of Reports of Sample Surveys" (UN, 1949b, 1949c). These recommendations highlighted the importance of reporting: (a) the *sampling units*; (b) the *frame*; and (c) the method of selecting (or recruiting) units—which may include (d) whether and how the frame was *stratified* before selection, (e) whether units were selected in *clusters*, (f) whether units were selected with *equal or unequal probabilities of selection*, and (g) whether units were selected in *multiple phases*. Also highlighted were reporting (h) *sample size*; (i) rates of refusals and attrition (see Enders & Gottschall, Chapter 14, this volume); (j) suspected areas of undercoverage of the frame; (k) methods undertaken after sample selection to gain insights into reasons for refusals and attrition; and (l) how the sample composition corresponds to preexisting survey data (e.g., census data). Table 10.1 provides definitions and brief examples of the italicized terms. Taken together, when a sample involves stratification, clustering, and/or disproportionate selection probabilities, it is conventionally called a *complex* sample, and those three key features are called *complex sampling features*. Sampling designs that lack all three features can be called *simple* (hence the term *simple random sample*).

In the 4 decades after their introduction, the UN guidelines had a limited impact on reporting practices, particularly in the social sciences. Indeed, a review of reporting practices from 1940–1979 found that instead of using the concrete terminology for describing sample selection as shown in Table 10.1, researchers often simply labeled their samples "representative" with little or no empirical substantiation (Kruskal & Mosteller, 1979a, b). That is, the descriptor "representative" was often used to provide "general, unjustified acclaim for the data," which Kruskal and Mosteller (1979b) equated to stating, "My sample will not lead you astray; take my word for it even though I give you no evidence … these data just happened to come to my hand, and I have no notion of the process that led to them or of relations between the target and sampled population" (p. 114–115). Moreover, Kruskal and Mosteller (1979b) found that the application of the term *representative* was itself ambiguous. Sometimes the term was meant to imply that sampling units were "typical cases" from a population; other times the term was used to convey that the sampling method provided "adequate coverage of population heterogeneity." In contrast to simply labeling a sample representative, the terms recommended by the UN Subcommission are less value-laden and communicate more precise information about the sample selection mechanism.

## Ethical Guidelines

Guidelines for reporting about sample selection began to move from the purely methodological sphere to the ethical sphere in the 1980s.

**TABLE 10.1**

Some Terms Useful for Reporting About Sample Selection

| Term | Definition | Examples |
|---|---|---|
| Sampling units | The physical units that were selected. | Persons, schools, divorce records, accident reports. |
| Sampling frame | All sampling units that had a nonzero probability of being selected into the sample. | A list of daycare centers in a community; all persons with registered university e-mail addresses; birth records from a particular county within a 2-month period. |
| Stratified sampling | Independently selecting sampling units from mutually exclusive groups, or strata, which may be preexisting or artificially defined. | Schools could be stratified into public vs. private; patients could be stratified into inpatient vs. outpatient; Alzheimer facilities could be stratified into nursing homes vs. assisted-living centers. |
| Cluster sampling | Using entire groups as sampling units, in lieu of individual elements, at one or more stages of selection. | Schools might be sampling units at a *primary stage* of selection; classes within schools might be sampling units at a *secondary stage* of selection; students within class might be sampling units at a *tertiary stage* of selection. Here classes and schools represent clusters of the *ultimate sampling unit*: students. Or, schools might be sampling units at a primary stage of selection, but all classes and all students are included within selected schools. This constitutes one, not three, *stages of selection*. |
| Multiple phases of selection | Used when a frame containing values on desired selection variable(s) is unavailable. In a two-phase design, the first phase of selection entails collecting these values from a large sample of units, which in turn constitutes the frame for the second phase of the study. | See hypothetical case example, in a later section of the chapter, for a detailed example. (Note that phases of selection, described here, are different than stages of selection, described above.) |
| Equal or unequal probabilities of selection | Whereas equal or unequal selection probabilities can be achieved with a probability sample, in nonprobability samples, units are typically selected with unequal probabilities on observed and/or unobserved variables; the main question then becomes if the selection variable(s) are, for example, (a) independent variable(s), (b) dependent variable(s), or (c) design variables that conditionally correlate and/or interact with independent variables, while predicting the outcome; our shorthand is to refer to (b) and (c) as *disproportionate selection.* | Equal selection probabilities for mice in a one-stage cluster sample of $j = 1 \ldots J$ litters could involve selecting litter $j$ with probability = (litter $j$ size)/(total number of mice in the frame), and then including all mice within selected litters. Disproportionate selection could involve selecting parents based on parental income to study the effects of parental monitoring on child academic performance (where income is a design variable omitted from analyses, income and monitoring are correlated, and income predicts academic performance). |

Surprisingly, this shift was largely not spurred by methodologists wanting to speed the sluggish adoption of the UN guidelines, but was rather the result of external pressures. After a series of highly publicized research scandals in the late 1970s and early 1980s (chronologically reviewed by Mitcham, 2003), Congress held several hearings on research ethics. These hearings resulted in the creation of federal offices to oversee the promotion of research integrity, to facilitate the publication of research misconduct regulations, and to encourage scientific societies to pay greater attention to ethics—particularly in the form of ethics codes. Subsequently, some specifics from the UN's methodological guidelines for reporting about sample selection were incorporated into several societal standards or ethics codes (e.g., the Council of American Survey Research Organization's [CASRO] *Code of Standards and Ethics*, 2006; the American Psychological Association's [APA] *Statistical Methods in Psychology Journals: Guidelines and Explanations* [Wilkinson & the Task Force on Statistical Inference, 1999]; and the American Association for Public Opinion Research's [AAPOR] *Code of Professional Ethics & Practices*, 1991–2005). However, no specific UN guidelines were incorporated into other codes (e.g., the International Statistical Institute's [ISI] *Declaration on Professional Ethics*, 1985–2009; the APA's *Ethical Principles of Psychologists and Code of Conduct* [APA, 2002]; and the American Statistical Association's [ASA] *Ethical Guidelines for Statistical Practice*, 1983–1999).[3]

## Current Practice

Between the more thorough methodological recommendations and less thorough ethical guidelines, resources on reporting about sample selection are now quite extensive. Still, a recent review of 10 observational studies in 2006 issues per each of four highly cited psychology journals (*Developmental Psychology*, *Journal of Personality and Social Psychology*, *Journal of Abnormal Psychology*, and *Journal of Educational Psychology*) found that 50 years of international methodological guidelines regarding how to report on sample selection (plus recent ethical guidelines) were not enough to routinely ensure adequate reporting practices in top-tier psychology journals (Sterba, Prinstein, & Nock, 2008). Of the 76% of studies that were nonprobability samples, only 23% described the method of selecting units (recruitment process), and only 52% reported anything about the

[3] For example, although the ASA's 1999 guidelines mentioned the general need to "explain the sample(s) actually used" (C5), the need to "include appropriate disclaimers" "when reporting analyses of volunteer data or other data not representative of a defined population" (C11) and the need to disclose consequences of failing to follow-through on an agreed sampling plan (C12), these guidelines still lack specifics about what sample selection features should be reported, and how.

sampled population. For the 24% of studies that were probability samples, corresponding figures were better: 89% and 100%, respectively.

Hence although recommended reporting practices have been included in several societal ethics codes and standards, this has not ensured their adoption in practice. We suggest two potential reasons why. First, the presence of material on reporting about sample selection was inconsistent from one ethics code to the next. Efforts to standardize the inclusion of reporting recommendations could provide a more coherent reference source for applied researchers. Second, none of the codes or standards provided an explicit rationale for whether, and if so why, reporting is indeed an ethical issue, not simply a methodological issue. It is odd to expect an ethical imperative to improve reporting practices without providing a motivating explanation.

### Is Reporting About Sample Selection an Ethical Issue?

There are several reasons why the gaps highlighted here between applied practice and methodological recommendations go beyond a purely methodological issue and into an ethical issue (e.g., negligence). These gaps are an ethical issue because researchers have the resources and ability to do something about them, but unintentionally have not, which leads to undesirable or even harmful consequences. This thesis is consistent with what is informally called the *ought implies can principle*: establishing that someone can do something is required before holding them accountable for doing it. Psychologists presently have the means to narrow the methods–practice gap regarding reporting about sample selection. Methodological recommendations and guidelines on reporting about sample selection have been available for an extremely long period—50 years—much longer than it typically takes a methodological advance to soak into applied practice. Additionally, the effort needed to implement recommended reporting practices is slim and does not require lengthy technical training. So the *ought implies can principle* is satisfied. Further, the consequences of adequate reporting about sample selection are important. Identifying and reporting complex sampling features that were used, such as those listed in Table 10.1, is a prerequisite first step before one can move on to determine whether these features need to be accounted for in data analyses. That is, if too little attention is paid to accurately reporting about sample selection, a researcher's ability to adequately account for the sample selection mechanism in data analysis is limited. Similarly, a reviewer's ability to crosscheck whether the analysis fully accounts for sample selection is limited. In turn, adequately accounting for the sample selection mechanism is necessary to ensure the validity of statistical inferences, as explained in the next section. When the validity of statistical inferences is in question, so are substantive conclusions based on those inferences.

In the last section of this chapter we make suggestions for narrowing this methods–practice gap in reporting practices, with the aid of this ethical imperative.

## Statistically Accounting for Sample Selection

### Methodological Guidelines

In contrast to the first issue we considered (reporting), the second issue we consider (statistically accounting for sample selection in data analysis) has not been translated into accessible international methodological guidelines, nor even widely disseminated beyond the more technical statistical literature. Nonetheless, the topic of when and how to statistically account for sample selection is no less important than reporting—and arguably more so. Sample selection impacts inference because the particular sample selection mechanism chosen can constrain the population to which inferences can be made. However, analytic techniques that incorporate sample design features can broaden the population of inference. Hence this section provides an accessible introduction to existing recommendations on this topic from within the statistics literature.

The following practical guidance on when and how to account for the sample selection mechanism was gleaned from recommendations within the statistics literature. When the sample selection mechanism involves complex sampling features—(a) clustering, (b) stratification, and/or (c) disproportionate selection of sampling units (e.g., using selection variables that correlate or interact with independent variables and predict the outcome)—these features typically need to be accounted for in statistical analyses (Skinner, Holt, & Smith, 1989). To account for this kind of disproportionate selection, selection and recruitment variables can be entered as covariates in the model and allowed to interact with independent variables (and/or can be incorporated into the model estimation using sampling weights, if a probability sample was used). Biemer and Christ (2008), Pfeffermann (1993, 1996), and Sterba (2009) provide examples and procedures for this approach. Further, to account for stratification and clustering, stratum indicators can be entered as fixed effects and cluster indicators may be entered as random effects in a multilevel model (or incorporated into sandwich-type standard error estimation adjustments for a single-level model). Chambers and Skinner (2003), Lohr, (1999), and Skinner et al. (1989) give examples and procedures for this second approach.

Moreover, if these complex sampling features are not accounted for in data analysis, there can be direct consequences for the validity of

statistical inferences. When stratification is not accounted for, standard errors are typically upwardly biased, and when clustering is not accounted for, standard errors are often downwardly biased (e.g., Kish & Frankel, 1974). When disproportionate selection is unaccounted for, point estimates and standard errors can both be biased (e.g., Berk, 1983; Smith, 1983; Sugden & Smith, 1984).

In the context of the complex sample features used in a given study, researchers and journal reviewers may find it helpful to try to mentally classify a study's sample selection mechanism according to a taxonomy developed by Little (1982) and Rubin (1983).[4] This taxonomy classifies sample selection mechanisms as ignorable, conditionally ignorable, or nonignorable. Each taxon poses different implications for the validity of inferences when the sample selection mechanism is or is not accounted for.

### *Ignorable Sample Selection*

Any time the probability of selecting sampling units is proportionate to the rate at which those units appear in the frame,[5] and sampling units are neither stratified nor clustered, the sample selection mechanism is ignorable and does not need to be accounted for in the data analysis. One sampling mechanism that is always ignorable is a simple random sample.

### *Conditionally Ignorable Sample Selection*

When some complex sampling features are used, but these features are properly accounted for in data analysis, as described previously, the sampling mechanism can be thought of as conditionally ignorable. A selection mechanism rendered conditionally ignorable by the data analysis will not result in biased parameter estimates or standard errors, and thus will not affect the validity of inferences.

### *Nonignorable Sample Selection*

Consider instead the circumstance in which sampling units are again selected with (a) clustering, (b) stratification, or (c) disproportionate

[4] Closely related versions of this taxonomy have been used to describe not only sample selection mechanisms but also missing data mechanisms. All versions stem from Rubin (1976). That is, the criteria used to determine whether we need to statistically account for the process by which persons entered the sample (i.e., sample selection mechanism) are similar to the criteria used to determine whether we need to statistically account for the process by which persons or observations are missing from the sample (i.e., missing data mechanism; see Enders & Gottschall, Chapter 14, this volume).

[5] Here we are assuming no frame error (e.g., over- or undercoverage) that would make the frame systematically different than the inferential population.

selection probabilities. Furthermore, suppose that some selection variables, stratum indicators, and/or cluster indicators are partially *un*observed, or *un*recorded. This would prevent their complete incorporation into the model (and/or complete incorporation into estimation-based weighting and standard error adjustments).[6] Or, suppose that selection variables, stratum indicators, and/or cluster indicators are fully observed but are simply omitted from the model specification and/or estimation. Under either circumstance, the sample selection mechanism is *nonignorable*, meaning that it may result in biased parameter estimates or standard errors in the data analysis, and thus may affect the validity of inferences.

It can be seen from this taxonomy that classifying a sample selection mechanism as *ignorable*, *conditionally ignorable*, or *nonignorable* depends partially on how the sample was selected at the data collection stage, and partially on how the sample selection mechanism was statistically accounted for at the data analysis stage. Fully *ignorable* sample selection is rare; as previously mentioned, simple random samples and their equivalent would fall into this category. Achieving *conditionally ignorable* sample selection and avoiding *nonignorable* sample selection is the typical goal.

## Ethical Guidelines

We earlier mentioned that methodological recommendations on reporting are more widely disseminated than methodological recommendations on when and how to statistically account for sample selection. Similarly, many ethical guidelines that did describe desirable reporting practices in detail are silent on the topic of statistically accounting for the sample selection mechanism (e.g., AAPOR, 2005; APA, 2002; CASRO, 2006; ISI, 1985–2009). The ethical guidelines that do comment on when and how to statistically account for sample selection are in some cases vague, which can limit their practical use. For example, ASA's (1999) ethical guideline A2 is to "Employ data selection or sampling methods and analytic approaches that are designed to assure valid analyses," and ethical guideline B5 is to "Apply statistical sampling and analysis procedures scientifically, without predetermining the outcome." In other cases, available societal standards are misleading. For example, Wilkinson and

[6] As mentioned previously, for probability samples these estimation adjustments can involve probability-weighted point estimators and stratified, between-cluster sandwich variance estimators. Our focus here is on nonprobability samples, where probability weights are unavailable, but sandwich variance estimators are available (yet less often used). An overview of these estimation adjustments is given in du Toit, du Toit, Mels, and Cheng (2005).

the Task Force on Statistical Inference (1999, p. 595) imply that stratification and clustering need to be accounted for only in statistical models for probability (i.e., random) samples. But the same requirement applies to nonprobability samples as well. Furthermore, they made no mention of needing to statistically account for other complex sampling features besides clustering and stratification (e.g., disproportionate probabilities of selection). To be sure, when and how to statistically account for sample selection is a less straightforward topic than reporting. This fact may have discouraged the incorporation of the former topic into societal standards and/or ethics codes. Nevertheless, it seems safe to say that more concrete, less misleading statements could be made without glossing over the complexities of deciding when to account for sample selection and without oversimplifying the alternative approaches for how to account for sample selection in data analysis.

## Current Practice

A common perception is that so little is known about selection mechanisms for typical nonprobability samples in psychology that the possibility of following the aforementioned methodological guidelines is precluded (e.g., Jaffe, 2005; Peterson, 2001; Sears, 1986). That is, it is thought impossible for selection mechanisms from typical nonprobability samples in psychology to be rendered conditionally ignorable by statistically controlling for complex sample selection features. However, Sterba et al.'s (2008) article review indicated that this may not be the case. They found that 28% of studies based on nonprobability samples used one or more discernible (observed) complex sampling features (stratification, clustering, or disproportionate selection), and the authors accounted for all of them in the statistical model (potentially *conditionally ignorable sample selection*).[7] Another 58% of studies had one or more discernible complex sampling feature(s) but did not account for all of them in the statistical model (potentially *nonignorable sample selection*). The remaining 14% of studies had no discernible complex sampling features (potentially *ignorable sample selection*).[8] Corresponding percentages for probability samples were 56%, 33%, and 11%, respectively. This review tells us that there is a gap between the data available on known complex sample selection features on the one hand, and the subsequent use of those data in analyses to account for sample selection on the other.

[7] Instances of clustering solely as a result of time within person were not counted toward this total.

[8] It would have been useful if these authors had explicitly stated whether any complex sampling features were used so that their sample selection mechanisms could have been more cleanly classified.

That is, samples are being treated as if they were simple random samples despite the fact that they include complex sampling features. Put another way, researchers are often not fully capitalizing on the potential to render their sample selection mechanisms conditionally ignorable in their data analyses.

### Is Statistically Accounting for Sample Selection an Ethical Issue?

Not only are specific recommendations on statistically accounting for sample selection included in few ethics codes, but also a motivating explanation is typically absent. Without consistent inclusion and without justification, it is unsurprising that this ethical imperative seems not to have greatly affected practice. One potential two-part justification for considering accounting for sample selection an ethical issue is given here. First, psychologists often have the means to narrow the methods–practice gap regarding accounting for sample selection in data analysis. That is, more data on complex sampling features are often collected than are ultimately used in analyses (see previous section). Furthermore, multiple commercial software programs capable of accounting for complex sampling features are available; some have been available for more than 15 years. See the online appendix of Sterba (2009) for a software review. Second, the real world consequences of bias induced by unaccounted for, complex sampling features can affect substantive conclusions; this in turn misdirects scientific understanding and federal grant spending and can waste participants' time (Sterba, 2006).

But there is no denying that sometimes psychologists' means are limited; sometimes not enough is known about the sample selection mechanism in nonprobability studies to be able to fully control for it in the data analysis. This is less often the case in probability samples, where the logistics of the sampling design require that all stratum indicators, cluster indicators, and selection variable scores on the frame are observed so they can be used to assign probabilities of selection to units. This fact is certainly a strength of probability sampling and is reason to prefer it where possible. However, even in nonprobability samples, risk of biased inferences can be minimized in certain ways by recording more information on the sample selection mechanism during data collection. Also, the effects of sample selection features that were partially unobserved can sometimes still be investigated in statistical analyses to ascertain how much they may be impacting substantive conclusions. We next consider a short, hypothetical case example that illustrates how the recording of information about sample selection can be improved. We subsequently consider a longer, empirical case example that illustrates one way to investigate the effects of partially observed selection features in a common daily diary study design.

## Strategies for Narrowing the Gap Between Methodological Recommendations and Practice: Case Examples

### Hypothetical Case Example: Recording More Information About Sample Selection

For this hypothetical case example, suppose a researcher intends to collect a nonprobability, convenience sample in a community setting, and suppose the researcher wants to oversample adolescents with sleep problems. Convenience samples are often used by psychologists when a specific, nonreferred subpopulation is desired but a frame or listing of sampling units that includes the selection variable(s) (e.g., sleep problems) is unavailable. For example, to oversample adolescents with sleep problems, study advertisements typically mention the variable to be oversampled (i.e., sleep problems). Would-be participants self-select into the study based on their interest, incentives, and/or their own perceived elevation on the variables mentioned in the advertisement. They then may be included or excluded based on additional study criteria or to meet quotas of youth with and without sleep problems. The problems are that it is unclear from this design (a) what variables persons (self-)selected on and (b) at what rate persons are being over- or undersampled. If unobserved self-selection variables are correlated with independent variables in the analysis and are predictive of the outcome, parameter bias may result.

Suppose, however, that this researcher is open to collecting additional data to more fully understand the selection mechanism. Here we consider one relatively simple and inexpensive method for collecting additional data about sample selection: conversion of a convenience sample into a two-phase sample (see Pickles, Dunn, & Vazquez-Barquero, 1995, for a review of two-phase samples). After describing how this convenience sample can be converted into a two-phase sample, we describe how the two-phase sample to some extent circumvents problems (a) and (b) mentioned above.

In phase 1 of a two-phase design, a brief screening questionnaire including questions about sleep problems and any other desired inclusion and exclusion criteria would be administered cheaply to a larger number of units than the desired sample size. This phase 1 sample might be taken from an institution such as a community health clinic (e.g., all well-child visits to a community health clinic in a given month) or from public records (e.g., all marriage records in a certain county for a certain duration of time). The phase 1 screened sample becomes the frame for the phase 2 sample. That is, the point of the screen is to record scores on selection variable(s) (e.g., sleep problem scores) for units who will then constitute the phase 2 frame. This in turn allows phase 1 individuals to

be allocated to nonoverlapping strata based on their screen responses (e.g., high sleep problems stratum, low sleep problems stratum). Then, at phase 2, participants can be randomly sampled from the high sleep problems stratum at a higher rate (e.g., 80%) than the low sleep problems stratum (e.g., 20%). Furthermore, the inverse of the selection probabilities in each stratum can be used as sampling weights in the data analysis phase to ensure that the phase 2 sample is statistically generalizable to the phase 1 sample.

The key improvement of the two-phase sample over the convenience sample is that selection from phase 1 to phase 2 is now based mainly on observed variables under the control of the researcher. These observed selection variables can then be used as covariates in the analysis or entered into weight variables in the analysis. In so doing, problems (a) and (b) from the convenience sample have now been circumvented for inference from the phase 2 sample to the phase 1 sample, even if generalizability from the phase 1 sample to an undefined larger population is still uncertain.[9] The sample selection mechanism for the phase 2 sample is thus conditionally ignorable. Another way of looking at the added advantage of the two-phase sample is that, to a much greater degree, it disentangles interest in participating from eligibility to participate; the collection of screening information at phase 1 is *not* contingent on interest in participating in phase 2.

During the implementation of the two-phase sampling design, the following information needs to be collected and later reported: (a) the proportion of persons refusing the phase 1 screen and, if possible, the reasons for refusal, recruitment mode, and basic demographic information; (b) the mode of recruitment for persons completing the phase 1 screen (e.g., newspaper, e-mail, flier); (c) the proportions of persons who were excluded after phase 1 and the reasons for their exclusion; and (d) the proportion of persons recruited into phase 2 who refused and their reasons for refusal. It is often helpful to present information for items (a)–(d) in a flowchart (see Sterba, Egger, & Angold, 2007, p. 1007, for an example).

## Empirical Case Example: Investigating the Effects of Partially Observed Selection Features

The previous case example considered the circumstance where data had not yet been collected, such that the data collection method could be modified to record more detailed information about sample selection. For samples that have already been collected, this option is not available. Consider now the situation in which a study has already been completed,

[9] In a later section, we discuss procedures that could be used to gain some insight into the correspondence between the phase I sample and a particular target finite population.

but some complex selection features were partially observed or partially recorded, raising the possibility of a nonignorable selection mechanism with accompanying parameter and standard error bias. Specifically, we consider the situation in which some selection variables were unobserved, but any strata and cluster indicators used were observed. In this situation, there are several possible approaches for investigating whether the sample is systematically different from a particular target finite population of interest for inference.

One approach was briefly mentioned earlier: Find a large-scale probability sample collected from the target finite population (e.g., general population survey or census) and compare it with the sample on key variables—particularly variables that were hypothesized to be involved in selection and were included in both data sets. Another approach involves applying intensive effort to recruit a small subsample of persons in the target finite population who initially refused contact, participation, or screening. Then compare their responses with participants on key variables. Groves (2006, p. 655–656) and Stoop (2004) discuss the first and second approaches in greater detail. A third approach involves applying a model-based *sensitivity analysis* to find out the extent to which the suspected nonignorability of the sample selection mechanism impacts substantive conclusions. In this context, a sensitivity analysis involves specifying at least one alternative model in addition to the theorized model of substantive interest. These alternative model(s) relax certain assumptions about the sample selection mechanism. Those assumptions were potentially violated in the original theorized model of substantive interest. Comparing alternative model(s) with the original theorized model, the researcher can see whether their substantive conclusions are sensitive to different assumptions about the sample selection mechanism. This third approach can often involve some time and cost savings over the previous two approaches; thus, it is the one we empirically illustrate here.

Our empirical case example uses Nock and Prinstein's nonprobability experience sampling (or daily diary) study of nonsuicidal self-injury (NSSI) behaviors. In this case example, responses were solicited at repeated assessments, and our interest lies in the validity of inferences from the subset of persons selected at each repeated assessment to the full, originally recruited sample.[10] Thus, this case example differs from previously discussed examples in that sample selection occurs more than once. This case example also differs from previously discussed examples in that inference to the originally recruited sample, rather than to a wider population, is desired. Validity of inference from the

[10] Because of the manner of selection at each time point (to be described shortly), we find it more intuitive to characterize this case example in terms of a sample selection problem, but it is possible to alternatively think of it as a missing data problem.

original sample to a wider population would entail other analyses (e.g., comparisons using external finite population data) that are outside the scope of the present discussion (see Nock, Prinstein, & Sterba, 2009, for more information).

### *Sample Selection Mechanism*

In this empirical case example, the full, originally recruited sample consisted of 30 adolescents. For 14 days, these 30 adolescents were exogenously signaled to respond with their context, feelings, thoughts, and behaviors related to NSSI at several points throughout the day (called *signal-contingent selection*) and were told to also respond about these matters specifically when they were having an NSSI thought (called *event-contingent selection*). Signal-contingent selection, event-contingent selection, and their combination are widely used methods of soliciting responses at repeated assessments in daily diary studies (Bolger, Davis, & Rafaeli, 2003; Ebner-Priemer, Eid, Kleindienst, Stabenow, & Trull, 2009; Shiffman, 2007; Wheeler & Reis, 1991). Specifically, in signal-contingent selection, participants are prompted to respond by an external device that is preprogrammed to signal at fixed or varying time intervals. In contrast, in event-contingent selection, responses are solicited based on the current behavior, feelings, context, or thoughts of the participant. Event-contingent selection has been particularly recommended for rare or highly specific experiences, including interpersonal conflict, intimacy, alcohol consumption, and mood (Bolger et al., 2003; Ebner-Priemer et al., 2009). Event-contingent selection was used in the case example because NSSI is a rare experience.

In this case example, the "event" is the dependent variable itself, NSSI thought (which differs from Nock et al., 2009). Thus, the selection mechanism is suspected to be nonignorable. In this context, nonignorability practically means that the effects of independent variables on the propensity to have an NSSI thought are confounded with the effects of independent variables on the propensity to self-report. It may be the case that different covariates, or different levels of the same covariates, predict propensity to self-report versus propensity to have an NSSI thought, if we could tease those two processes apart. Yet, even in this worst-case scenario, a sensitivity analysis can be conducted to see whether this potentially nonignorable selection method meaningfully impacts results. We will see later that this sensitivity analysis capitalizes on the fact that a *combination* of event- and signal-contingent selection was used. The sensitivity analysis demonstrated here (proposed in Sterba et al., 2008, and used in Nock et al., 2009) adapts what has been termed a *shared parameter model* (Follmann & Wu, 1995; Little, 1995) or a *two-part model* (Olsen & Shafer, 2001) for the case of sample selection. These models have some similarities to traditional

cross-sectional single-level selection models (e.g., Heckman, 1979) but are less restrictive.

### *Sensitivity Analysis Step 1*

The first step in this sensitivity analysis is to specify our model of substantive theoretical interest as per usual; let us call it our outcome-generating model. This model ignores whether the response was self-selected (i.e., event-contingent) or signal-driven (i.e., signal-contingent). In our outcome-generating model, independent variables of interest at level 1 (observation level) are whether the participant was currently using drugs (*drug*), feeling rejected (*reject*), feeling sad (*sad*), feeling numb (*numb*), and whether they were with peers (*peer*). Independent variables of interest at level 2 (person level) are *age* and *sex*. In the specification of this outcome-generating model, the nesting of responses within an individual is accounted for using a multilevel model with a random intercept.[11]

Specifically, the outcome model predicting binary NSSI thoughts is:

$$\log\left[\frac{\Pr(thought_{ij}=1)}{1-\Pr(thought_{ij}=1)}\right] = \gamma^{o}_{00} + \gamma^{o}_{10} drug_{ij} + \gamma^{o}_{20} reject_{ij} + \gamma^{o}_{30} sad_{ij} + \gamma^{o}_{40} numb_{ij} + \gamma^{o}_{50} peer_{ij} + \gamma^{0}_{01} sex_{j} + \gamma^{o}_{02} age_{j} + u^{o}_{0j} \quad (10.1)$$

where the superscript *o* denotes the outcome equation, *i* denotes observation, and *j* denotes person. The γ represents a fixed effect, the *u* represents a random effect, and the random intercept variance is estimated $u^{o}_{0j} \sim N(0, \tau^{o})$. This multilevel model can also be portrayed graphically using Curran and Bauer's (2007) path diagrammatic notation, as in Figure 10.1. Drug use, rejection, sadness, and numbness were hypothesized to be positively related to NSSI thoughts, and being with peers was hypothesized to be negatively related to NSSI thoughts. Sex and age were control variables. Table 10.2, column 1, shows that only the hypotheses about rejection and sadness were supported.

### *Sensitivity Analysis Step 2*

Estimates in Table 10.2, column 1, could be biased if the sample selection mechanism is not independent from the outcome-generating mechanism (i.e., if it is nonignorable, as we suspect). That is, if the selection and

[11] Checks for autocorrelation, cyclicity, and trend were described in Nock et al. (2009), and little evidence of each was found. For simplicity, these checks are not discussed here. A three-level model (responses nested within day nested within person) encountered estimation problems as a result of little day-to-day variability in NSSI thoughts; the day level was therefore dropped.

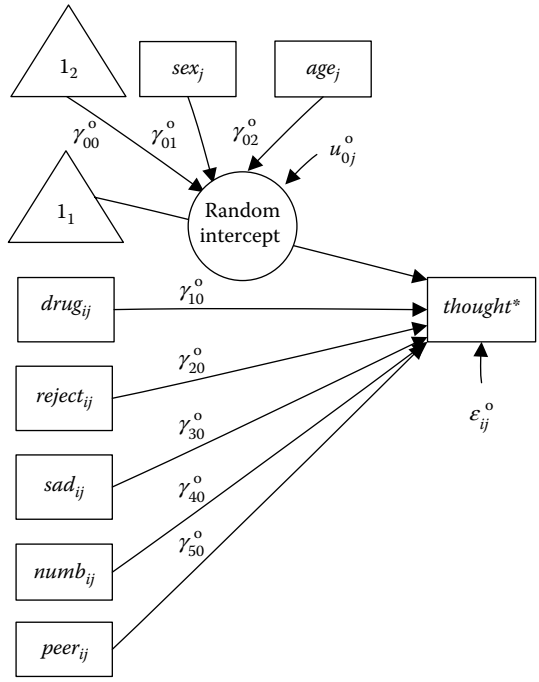



**FIGURE 10.1**
Path diagram for empirical case example: outcome model only, ignoring selection. Squares are measured variables. Circles are latent coefficients. Triangles are constants. Straight arrows are regression paths. Symbols are defined in the text equations. The multilevel model path diagram framework used here was introduced in Curran and Bauer (2007).

outcome-generating mechanisms are dependent, the effect of a predictor on the outcome is confounded with the effect of the predictor on the probability of selection. Therefore, the slope coefficients in Table 10.2, column 1, would simultaneously represent both effects. To investigate whether the potentially nonignorable selection is affecting estimates in Table 10.2, column 1, we need to specify not just an outcome-generating submodel, as per usual, but also a model for the sample selection mechanism (let us call it a selection model). Then we need to assess the extent to which these two models are interdependent. In the selection submodel, we are predicting the log odds of self-initiated response (*selection* = 1) versus a signal-initiated response (*selection* = 0). It was hypothesized that persons would be more likely to self-select if they were not with peers, were female, were feeling less numb, and were feeling more rejected.

$$\log\left[\frac{\Pr(selection_{ij} = 1)}{1 - \Pr(selection_{ij} = 1)}\right] = \gamma^{s}_{00} + \gamma^{s}_{10} reject_{ij} + \gamma^{s}_{20} numb_{ij} + \gamma^{s}_{30} peer_{ij} + \gamma^{s}_{01} sex_{j} + u^{s}_{0j} \quad (10.2)$$

**TABLE 10.2**

Empirical Case Example Results: Sensitivity Analysis for Nonignorable Selection in a Daily Diary Study

| | Model 1: Ignoring Selection | | Model 2: Accounting for Selection | |
|---|---|---|---|---|
| | **Estimate (SE)** | ***p* Value** | **Estimate (SE)** | ***p* Value** |
| Outcome (sub)model: | | | | |
| *Fixed effects* | | | | |
| Intercept | 0.764 (2.720) | .779 | 1.080 (2.539) | .671 |
| Using drugs | 0.304 (0.200) | .129 | 0.328 (0.103) | .001 |
| Feeling rejected | 1.108 (0.482) | .022 | 1.055 (0.474) | .026 |
| Feeling sad | 0.665 (0.274) | .015 | 0.671 (0.266) | .012 |
| Feeling numb | −0.561 (0.277) | .043 | −0.548 (0.273) | .044 |
| With peers | 0.014 (0.375) | .970 | 0.015 (0.350) | .965 |
| Age | −0.050 (0.112) | .658 | −0.081 (0.144) | .571 |
| Sex | 0.529 (0.636) | .406 | 0.707 (0.699) | .312 |
| *Variance components* | | | | |
| $\tau^o$ | 3.355 (1.270) | .008 | 2.726 (0.885) | .002 |
| Selection submodel: | | | | |
| *Fixed effects* | | | | |
| Intercept | | | −0.169 (0.674) | .802 |
| Feeling rejected | | | 1.340 (0.336) | .000 |
| Feeling numb | | | 0.098 (0.416) | .813 |
| With peers | | | −0.748 (0.339) | .027 |
| Sex | | | 0.814 (0.377) | .031 |
| *Variance components* | | | | |
| $\tau^s$ | | | 0.125 (0.116) | .283 |
| $\tau^{o,s}$ | | | 0.288 (0.122) | .018 |

Estimates are in the logit scale. $\tau^o$, intercept variance for the outcome (sub)model; $\tau^s$, intercept variance for the selection submodel; $\tau^{o,s}$, covariance between the random intercepts in both (sub)models.

Here the superscript *s* stands for the selection submodel and $u^s_{0j} \sim N(0, \tau^s)$. Note that the selection and outcome submodels can have the same or different covariates (Follmann & Wu, 1995). It was hypothesized that, controlling for these observed covariates, the probability that *selection* = 1 would still be dependent on the probability that *thought* = 1 because the self-initiated nature of the event-contingent responding is partially dependent on the presence of a thought. However, this dependency is now accounted for by simultaneously estimating the outcome and selection submodels, and by allowing the individual deviations on *thought* to covary with the

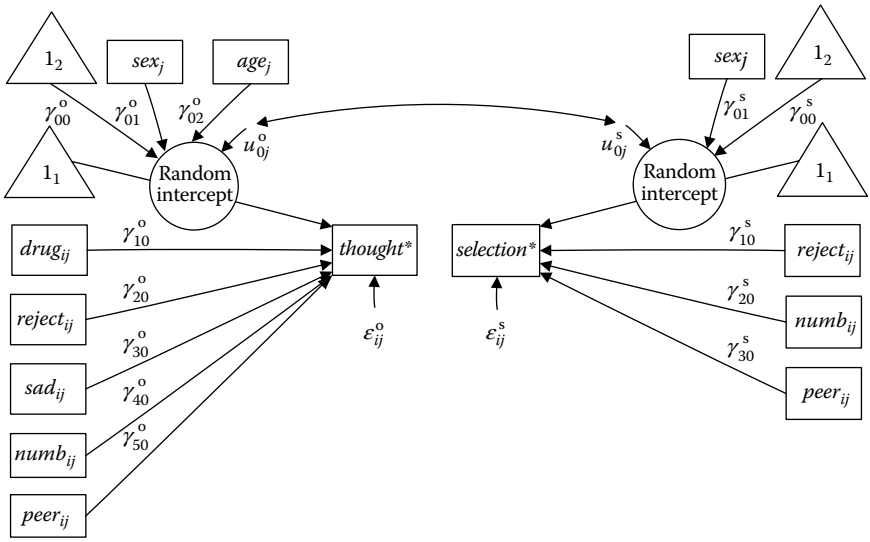



**FIGURE 10.2**
Path diagram for empirical case example: joint outcome and selection model. Curved arrows are covariances.

individual deviations on *selection*. That is, the intercept random effects for the outcome equation and the selection equation covary because of the term $\tau^{o,s}$:

$$\begin{matrix} u^{o}_{0j} \\ u^{s}_{0j} \end{matrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^{o} & \\ \tau^{o,s} & \tau^{s} \end{pmatrix}\right]. \tag{10.3}$$

This joint model assumes that, *conditional on the random effect*, the outcome and selection processes are independent.[12] This assumption is more lenient than when we just estimated the outcome-generating model, in which selection and NSSI thoughts were assumed to be *unconditionally* independent. A path diagram for this joint outcome–selection model is given in Figure 10.2.

### *Sensitivity Analysis Results and Conclusions*

Results of the joint outcome–selection model are shown in Table 10.2, column 2. Being alone, being female, and feeling rejected increased the probability of self-selection, as hypothesized. However, feeling numb did not

[12] This model is highly related to a random coefficient-dependent selection model. The latter model typically requires the correlation between the random intercept in the selection and outcome equations to be 1.0, whereas here we are freely estimating it.

decrease the probability of self-selection, as was hypothesized. In addition, the selection and outcome submodels are statistically dependent, controlling for observed covariates ($\tau^{o,s}$ = .288 [.122], $p$ = .018). Even though our selection mechanism meets the technical definition of nonignorability, we are reassured to find that most of our substantive conclusions stay the same once we allow for the nonignorable selection. The only change occurred in the effect of drug use on the log odds of NSSI thoughts, which is now significant.

It is important to underscore that we called this approach a sensitivity analysis because we are not claiming that the Table 10.2, column 2, model is the *one true model* per se. The joint outcome–selection model rests on untestable assumptions about both the selection and outcome submodels and assumes that both submodels are properly specified—even though the researcher may be less confident about specifying the selection model (Little, 1995). We recommend specifying several theoretically compelling selection models (just as one would specify competing outcome models) and then investigating whether consistent results are found across perturbations in the selection model. Additional background and rationale for using selection models in sensitivity analyses can be found in Molenberghs and Verbeke (2005). In cases like this example, we recommend reporting (a) that there is evidence of a nonignorable selection mechanism; (b) that a sensitivity analysis was conducted; (c) whether and which parameter estimates differed when nonignorable selection was accounted for; and (d) whether these changes were found across alternative theoretically driven selection models.

## Conclusion

In this chapter, we showed that publicized methodological and ethical guidelines have focused on the issue of reporting about sample selection more so than the issue of statistically accounting for sample selection in data analysis. Whereas this discrepancy may exist because the former issue is more straightforward to address, the former issue is certainly no more important than the latter. Further, we showed that a sizeable gap exists between methodological recommendations and applied practice for both issues. In response, we provided statistical rationales and ethical imperatives for why researchers ought to pay greater attention to both issues. Finally, we supplied two case examples illustrating certain ways this gap could be narrowed for particular nonprobability sampling designs.

Of course, the issues of incomplete reporting about sample selection and incomplete accounting for complex sampling features are just two

of the methodologically and ethically important issues reviewed in the chapters of this book. However, we would argue that for psychologists these two issues are overlooked more often than some others considered in this book. For example, part of the culture of our discipline is to spend comparatively much less time dealing with sample selection issues in analysis and reporting than, say, measurement issues (whereas the reverse is true in other disciplines, like epidemiology; Sterba, 2009). Our recommendation for speeding the closure of the methods–practice gap is to make doing so a proximal priority, rather than a distal aspiration. Many ethics codes, including that of the APA, are primarily aspirational in nature. In contrast, medical journals have successfully elevated a number of methodological issues to proximal priorities by forming a cohesive International Committee of Medical Journal Editors and including these issues in their "Uniform Requirements for Manuscripts Submitted to Biomedical Journals" (see also Fidler, Chapter 17, this volume). The medical journal model is certainly worth further consideration by psychological journal editors as a stimulus for improving sample selection reporting and analysis practices. In this regard, it is worth emphasizing that *exclusively improving reporting practices*, as a first step, would likely spur subsequent improvements in data analysis practices as well. That is, simply identifying the complex sampling features that were used in the sampling design would alert readers and reviewers of the features that should have been accounted for in data analysis.

## References

American Association for Public Opinion Research. (2005). *Code of professional ethics and practice*. Retrieved from http://www.aapor.org/aapor_code.htm

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Retrieved from http://www.apa.org/ethics/code/index.aspx

American Statistical Association. (1999). *Ethical guidelines for statistical practice*. Retrieved from http://www.amstat.org/about/ethicalguidelines.cfm

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review, 48,* 386–398.

Biemer, P., & Christ, S. (2008). Weighting survey data. In E. de Leeuw, J. Hox, & D. Dillman (Eds.), *International handbook of survey methodology* (pp. 317–341). New York: Erlbaum.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54,* 579–616.

Bowley, A. L. (1906). Address to the economic and statistics section of the British Association for the Advancement of Science, York, 1906. *Journal of the Royal Statistical Society, 69,* 540–558.

Cassel, C., Sarndal, C., & Wretman, J. (1977). *Foundations of inference in survey sampling*. New York: Wiley.

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data.* Chichester, UK: Wiley.

Council of American Survey Research Organizations. (2006). *Code of standards and ethics for survey research*. Retrieved from http://www.casro.org/codeofstandards.cfm

Curran, P. J., & Bauer, D. J. (2007). A path diagramming framework for multilevel models. *Psychological Methods, 12,* 283–297.

du Toit, S. H. C., du Toit, M., Mels, G., & Cheng, Y., (2005). *Analysis of complex survey data with LISREL: Chapters 1–5.* Unpublished manual. Retrieved from http://www.ssicentral.com

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology, 118,* 195–202.

Fienberg, S. E., & Tanur, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review, 55,* 75–96.

Follmann, D., & Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics, 51,* 151–168.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70,* 646–675.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153–162.

International Statistical Institute. (1985–2009). *Declaration on professional ethics*. Retrieved from http://isi.cbs.nl

Jaffe, E. (2005). How random is that? *Association for Psychological Science Observer, 18,* 9.

Jensen, A. (1926). Report on the representative method in statistics. *Bulletin of* the International *Statistical Institute, 22,* 359–380. Extensive discussion on pp. 58–69, 185–186, and 212–213.

Kaier, A. N. (1895). Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute, 9,* 176–183.

Kish, L. (1996). Developing samplers for developing countries. *International Statistical Review, 64,* 143–162.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society Series B, 36,* 1–37.

Kruskal, W., & Mosteller, F. (1979a). Representative sampling III: The current statistical literature. *International Statistical Review, 47,* 245–265.

Kruskal, W., & Mosteller, F. (1979b). Representative sampling II: Scientific literature, excluding statistics. *International Statistical Review, 47,* 111–127.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association, 77,* 237–250.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90,* 1112–1121.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole.

Mitcham, C. (2003). Co-responsibility for research integrity. *Science and Engineering Ethics, 9,* 273–290.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 109,* 558–606.

Nock, M., Prinstein, M. J., & Sterba, S. K. (2009). Revealing the form and function of self-injurious thoughts and behaviors: A real-time ecological assessment study among adolescents and young adults. *Journal of Abnormal Psychology, 118,* 816–827.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association, 96,* 730–745.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second order meta-analysis. *Journal of Consumer Research, 28,* 250–261.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61,* 317–337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research, 5,* 239–261.

Pickles, A., Dunn, G., & Vazquez-Barquero, J. L. (1995). Screening for stratification in two-phase epidemiological surveys. *Statistical Methods in Medical Research, 4,* 73–89.

Royall, R. M., & Herson, H. J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association, 68,* 880–889.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Rubin, D. B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys: Comment. *Journal of the American Statistical Association, 78,* 803–805.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Shiffman, S. (2007). Designing protocols for ecological momentary assessment. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 27–53). New York: Oxford University Press.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys*. New York: Wiley.

Smith, T. M. F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society: Series A, 146,* 394–403.

Smith, T. M. F. (1994). Sample surveys 1975–1990: An age of reconciliation? *International Statistical Review, 62,* 5–19.

Stephan, F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association, 43,* 12–39.

Sterba, S. K. (2006). Misconduct in the analysis and reporting of data: Bridging methodological and ethical agendas for change. *Ethics & Behavior, 16,* 305–318.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44,* 711–740.

Sterba, S. K., Egger, H. L., & Angold, A. (2007). Diagnostic specificity and nonspecificity in the dimensions of preschool psychopathology. *Journal of Child Psychology and Psychiatry, 48,* 1005–1013.

Sterba, S. K., Prinstein, M. J., & Nock, M. (2008). Beyond pretending complex nonrandom samples are simple and random. In A. T. Panter & S. K. Sterba (Co-chairs), *Quantitative methodology viewed through an ethical lens*. Boston: Division 5, American Psychological Association.

Stoop, I. A. (2004). Surveying nonrespondents. *Field Methods, 16,* 23–54.

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71,* 495–506.

United Nations. (1946). *Economical and social council official records: Report of the statistical commission, first year*. Lake Success, NY.

United Nations. (1947). *Economical and social council official records: Report of the statistical commission, second year*. Lake Success, NY.

United Nations. (1948). *Economical and social council official records: Report of the statistical commission, third year*. Lake Success, NY.

United Nations. (1949a). *Economical and social council official records: Report of the statistical commission, fourth year*. Lake Success, NY.

United Nations. (1949b). United Nations economic and social council sub-commission on statistical sampling: Report to the statistical commission on the second session of the sub-commission on statistical sampling I. *Sankhyā: The Indian Journal of Statistics, 9,* 377–391.

United Nations. (1949c). United Nations economic and social council sub-commission on statistical sampling: Report to the statistical commission on the second session of the sub-commission on statistical sampling II. *Sankhyā: The Indian Journal of Statistics, 9,* 392–398.

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality, 59,* 339–354.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

# 11

# *From Hypothesis Testing to Parameter Estimation: An Example of Evidence-Based Practice in Statistics*

**Geoff Cumming**
*La Trobe University*

**Fiona Fidler**
*La Trobe University*

The American Psychological Association (APA) *Publication Manual* (2010) includes several statements about hypotheses. For example, a subheading in the discussion of what an Introduction section should contain is: "State hypotheses and their correspondence to research design" (APA, 2010, p. 28). Another example is: "After presenting the results, … interpret their implications, especially with respect to your original hypotheses" (p. 35).

Such statements may seem merely benign assertions of the obvious, but the trouble is that psychology overwhelmingly interprets "hypothesis" in terms of *null hypothesis significance testing* (NHST) as a point statement usually of zero effect or zero difference. It has become the aim of statistical analysis, indeed of empirical research generally, to reject such hypotheses and conclude that an effect exists. This is, however, an impoverished, dichotomous choice: A result either is or is not statistically significant. Either we found evidence of an increase, or the change was statistically nonsignificant. This dichotomous mindset is illustrated clearly by this further statement in the *Publication Manual*: "Open the Discussion section with a clear statement of the support or non-support for your original hypotheses" (APA, 2010, p. 35).

There are three main sections to this chapter. In the first section, we make a case that NHST and *dichotomous thinking* should be replaced by statistical estimation and a more quantitative approach to theory and research planning. In the second section, we examine how advocacy of such a major change in statistical practice—from NHST to estimation—should

be framed, with particular attention to ethical considerations. We argue that *evidence-based practice* (EBP) in statistics needs to be an important goal. In the third section, we use our conclusions from the second section to guide formulation of our own argument that psychology, as well as other disciplines, should change from relying on NHST to mainly using estimation and *confidence intervals* (CIs). We draw on what cognitive evidence there is to support the change, and we identify further evidence that is needed—and that future research should seek. We then close with discussion of a simulation example.

## From NHST to Estimation: The Basic Argument

Impoverished dichotomous hypotheses have become the norm in psychology, and NHST is probably the main reason. Not only is statistical analysis centered on point null hypotheses, but psychologists also too often permit dichotomous thinking to limit their theories to mere statements of whether there is a difference or, at most, a statement of the direction of a difference. The distinguished psychologist Paul Meehl (1978) presented a devastating argument against this methodology. On NHST he stated:

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories … is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (Meehl, 1978, p. 817)

No definition of *ethics* can make such a severely criticized way to do science ethical. Meehl was scathing of psychological theories that merely made dichotomous predictions: "the usual flabby 'the boys are taller than the girls' or 'the schizophrenics are shyer than the manic depressives'" (pp. 826–827). He was scathing of research that relied on NHST: "A successful significance test of a substantive theory in soft psychology provides a feeble corroboration of the theory because the procedure has subjected the theory to a feeble risk" (p. 822). He emphasized that "I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless" (p. 823).

Meehl argued that a psychologist should instead aim for "a theory that makes precise predictions" (1978, p. 818). He blamed:

> The Fisherian tradition [NHST], with its soothing illusion of quantitative rigor, [which] has inhibited our search for stronger tests, so

> we have thrown in the sponge and abandoned hope of concocting substantive theories that will generate stronger consequences than merely "the Xs differ from the Ys." (Meehl, 1978, p. 824)

He drew a sharp contrast with other sciences: "have a look at any textbook of theoretical chemistry or physics, where one searches in vain for a statistical significance test" (p. 825), and he set as a goal for psychology that it should "generate numerical point predictions (the ideal case found in the exact sciences)" (p. 824).

Sadly, Meehl's analysis remains relevant more than 30 years later. Gigerenzer (1998) made a similar argument and blamed NHST for there being "little incentive to think hard and develop theories from which … [quantitative] hypotheses could be derived" (p. 201).

A chemist would inevitably report "the boiling point of this substance is 38.5 ± 0.2ºC" and would never dream of reporting "it is significantly greater than zero." Similarly, the physicist, astronomer, and, to some extent, the biologist expect to report and to read in journal articles experimental measurements and a statement of their degree of precision. Such researchers might read Meehl or Gigerenzer and wonder why such arguments even need to be made: Surely the main aim of most scientific research is to make and report empirical measurements about the world? Indeed it is, and so why does psychology not join other sciences and report parameter estimates with their CIs? Why does the psychologist not report "the new therapy improved depression scores by 7.5 ± 6.0 points" on some scale, rather than merely that it gave a "significant improvement"? Part of the answer may have been provided by the great statistical reformer Jacob Cohen: "I suspect that the main reason they [confidence intervals] are not reported is that they are so embarrassingly large!" (1994, p. 1002).

In the course of an extended discussion of the history of the uptake of NHST in psychology, and of the debate about its value, Fidler (2005) identified further reasons why NHST became so pervasive in psychology from around the mid-20th century:

> NHST also provided experimental psychology with the illusion of a mechanized knowledge building process: In this way, it served the ideal of objectivity. The dichotomous decision procedure of NHST seemingly removed experimenter judgment from the inferential process. It appeared no longer necessary to make subjective decisions about whether a phenomenon was real, or an effect important. "Statistical significance" became a substitute for both decisions. This rhetoric of objectivity was extremely important in psychology's struggle to be seen as a scientific discipline. (Fidler, 2005, p. 25)

For more than half a century the dominance of NHST in psychology has persisted. It continues to be taught in textbooks, implemented in

statistical software, expected by editors, and used without much reflection by researchers. It persists, despite the publication in that time of many cogent critiques of NHST and very few defenses of its use (Harlow, Mulaik, & Steiger, 1997; Kline, 2004). Meehl's "soothing illusion of quantitative rigor" (1978, p. 824) very largely continues. Dichotomous decision making seems to be deeply, deeply embedded in the thinking of most psychology researchers.

A range of better approaches to data analysis and statistical inference hold great potential for psychology. We will focus primarily on statistical estimation, meaning CIs, which provide point and *interval estimates* of population parameters. Parameter estimation not only gives a fuller picture of data but also should encourage us to ask questions that expect quantitative answers, rather than the dichotomous questions prompted by NHST. CIs also provide the basis for analysis of accuracy in parameter estimation, as discussed in Maxwell and Kelly (Chapter 6, this volume), and thus enable the planning of more efficient research.

Rodgers (2010) presented a strong case for quantitative modeling as the best—including, we would add, the most ethical—approach to epistemology in the social and behavioral sciences. He explained that even a detailed quantitative model can be regarded as a hypothesis to be tested against data by using a *p* value, but he advocated instead a model evaluation and model comparison approach. We agree: Even the goodness of fit of data to a model is better approached by avoiding dichotomous thinking and using estimation of fit indices rather than *hypothesis testing*. It is not only the testing of point null hypotheses that needs to be replaced.

CIs should provide "better answers to better questions" (Cumming & Fidler, 2009, p. 15). Researchers should be encouraged to develop, evaluate, and compare quantitative theories. A more quantitative, cumulative, and theoretically sophisticated discipline should be possible, and this should provide better, more quantitative, evidence-based guidance for practitioners. We next consider the two most recent editions of the *Publication Manual* (APA, 2010) and developments in its statistical advice that may move practice in this direction.

## The APA *Publication Manual*

In response to advocacy of *statistical reform*, the APA set up a Task Force on Statistical Inference, whose report (Wilkinson & the Task Force on Statistical Inference, 1999) is an excellent statement of good practice for the planning, conduct, and statistical analysis of research. The Task Force recommended CIs, although it did not—as some had hoped—recommend a ban on NHST. The Task Force was intended to provide guidance as to what statistical advice should be incorporated in the fifth edition of the

*Manual* (APA, 2001). The fifth edition did recommend CIs but, although it gave numerous examples and guidelines for reporting NHST, in its 439 pages gave no advice about reporting CIs and not a single CI example. From a statistical reform perspective, the fifth edition was a great disappointment (Fidler, 2002).

The statements about hypotheses that we quoted earlier from the *Publication Manual* came from the sixth edition (APA, 2010). They are typical also of many in the fifth edition, which largely followed the dichotomous decision making of NHST. However, they give a somewhat unfair impression of the sixth edition (APA, 2010), released in July 2009, which from a statistical reform perspective is a large advance over the fifth edition. For example, discussing the Introduction section of an article, the sixth edition refers to "stating your hypotheses or specific question," and the need to "examine the hypothesis or provide estimates in answer to the question" (both p. 28). The additional words referring to questions may not be noticed by most readers but for us signal a great advance and the prospect of progress beyond routine dichotomous decision making.

The sixth edition continues the recommendation of the fifth edition that "confidence intervals … are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended" (APA, 2010, p. 34). The sixth edition goes further by stating that researchers should "wherever possible, base discussion and interpretation of results on point and interval estimates" (p. 34). It also, for the first time, specifies a format for reporting a CI and includes many examples of CIs reported in text and tables. The sixth edition still includes numerous examples of NHST; therefore, although it recommends CIs and thus legitimates estimation, in our view it falls short by not giving any strong recommendation against NHST. Nevertheless, it makes important advances from the fifth edition. Its advice to base interpretation on point and interval estimates could prompt enormous and beneficial changes.

## Prospects for Change

Cumming et al. (2007) reported evidence that some statistical practices used in journal articles changed from 1998 to 2005–2006. NHST continued to dominate, appearing in 97% of empirical articles, but CI use increased from 4% to 11% of articles over the period, and inclusion of figures with error bars increased substantially, from 11% to 38% of articles. These are encouraging signs that statistical change is possible, and in some ways is occurring, even if NHST persists so overwhelmingly.

Given those signs of change, as well as the recommendations and examples in the sixth edition of the *Publication Manual* (APA, 2010), we conclude that prospects for improvement in psychology's statistical practices are

good, although certainly not guaranteed. Therefore, it is especially timely to consider the issues we discuss in this chapter.

In summary, we see CIs and an estimation approach to statistical inference as not only giving more complete information about experimental results, but also encouraging the development of more quantitative theories. In addition, research progress should be improved, and psychology should become a stronger discipline. With the imprimatur of the sixth edition of the *Publication Manual*, there is, perhaps for the first time in more than half a century, a real chance that psychology can reduce its reliance on NHST and increase its use of better techniques.

## From NHST to Estimation: How Should the Argument Be Formulated?

Distinguished statisticians, psychologists, and researchers in other disciplines have for more than half a century been putting forward cogent arguments that NHST practices are deeply flawed and that better practices, including the use of estimation based on CIs, should be preferred. The advocacy in psychology has been reviewed by Nickerson (2000) and Harlow et al. (1997) and more recently has been stated most clearly and constructively by Kline (2004, Chapter 3). The case has been made primarily in statistical terms: NHST, especially as practiced by psychologists, is flawed and damaging to research progress, whereas CIs are more informative and lead to better research decision making. Those arguments for change are strong, but we wish to emphasize an additional approach: advocacy in terms of EBP. Choice of statistical practices should be based on relevant research evidence about the effectiveness of the statistical techniques chosen.

EBP has, of course, been widely advocated and adopted in medical practice (Institute of Medicine, 2001) and more recently in the professional practice of psychology (Norcross, Beutler, & Levant, 2006). EBP in medicine can be defined as "the integration of best research evidence with clinical expertise and patient values" (Institute of Medicine, 2001, p. 147). A major reason for adopting EBP is that ethical considerations require us to do so (Hope, 1995): In any practice that claims to be based on science, practice will be most effective, safe, and beneficial for the patient or client when it can be justified in terms of empirically supported theory and research evidence relevant to the circumstances.

We believe these arguments should be extended to research practice, including especially statistical practice by researchers. Our claim is that EBP by researchers is required if research is to be conducted ethically.

The ethical frameworks and considerations put forward in the first section of this Handbook are all consistent with this claim (see Rosnow & Rosenthal, Chapter 3, this volume, and their matrix of costs and benefits of conducting or not conducting research).

Simply and most basically, to be ethical, research must use research resources, including notably the time and patience of research participants, with maximum efficiency for the acquisition of research knowledge. Theories should be detailed and accurate, and statistical techniques should be chosen to give conclusions that are as informative about the data and well-justified by the data as possible; then these must be communicated accurately, fairly, and clearly to readers of journal articles reporting the findings.

Advocacy of EBP emphasizes that ethical practice is best practice, and best practice is a shifting target. Research should continually provide further evidence, which should improve what is judged best practice. Today's best and therefore ethical practice can be tomorrow's discredited and therefore unethical practice. This applies just as much for statistical practices, as Fidler (Chapter 16, this volume) illustrates dramatically for the development of meta-analysis. Had this valuable statistical technique been available and used earlier, much better conclusions could have been drawn years earlier from then-available research evidence. A researcher now not using meta-analysis when it is appropriate is not following best statistical practice and is thus acting unethically.

What research evidence is needed to guide choice of statistical techniques? In other words, what is the evidence needed for EBP in statistics? The first kind of evidence that may spring to mind is evidence about the appropriateness of a particular statistical technique for application in a particular situation, for example, the results of statistical investigations of robustness. Evidence of robustness and appropriateness for a purpose is certainly important, but we wish to emphasize a different kind of evidence that is necessary, and perhaps even more crucial, in determining best statistical practice.

A fundamental requirement of best statistical practice is that it communicates accurately. It must summarize and present results in a way that is readily and correctly understood, and prompts justifiable conclusions. Further, it should assist researchers to conceive and carry out best practice research, as well as to analyze and communicate the findings in best practice ways. The way researchers conceive of theory and experiment and the way statistics communicate results are questions of cognition—of perception, understanding, reasoning, and decision making. So the evidence we need is cognitive evidence: How well does a statistical technique help the researcher build a good theory, design a good experiment, and write a clear journal article? How well does it present findings so that misconception is avoided and the data and conclusions are easily and successfully

understood by other researchers and, where appropriate, by others? The research field of *statistical cognition* (Beyth-Marom, Fidler, & Cumming, 2008) addresses exactly these questions.

Therefore, cognitive evidence is required if EBP of statistics is to be as good as possible. Appreciating the importance of such evidence, and carrying out research to gather it, may currently be the highest research priority for the achievement of EBP of statistics. Correspondingly, an essential component of the case for advocacy of a change in statistical practice, for example, the change from NHST to estimation, is consideration of cognition and relevant cognitive evidence. Therefore, we would consider adding cognition as an additional framework to those discussed by Gardenier in Chapter 2 of this volume.

## From NHST to Estimation: An Enhanced Argument for Reform

In addition to the statistical and other arguments that have already been put forward to justify a shift from NHST to estimation, in this third section we now consider what relevant cognitive evidence there is and what further evidence is desirable to support the identification of estimation as best, and therefore ethical practice.

### Cognitive Evidence About NHST

Consider first the criticism of NHST. Reform advocates have made statistical arguments to identify deficiencies of NHST and have presented claims that it is poorly understood and practiced. In addition, however, they have been able to cite considerable cognitive evidence that identifies particular misconceptions of NHST and $p$ values that are widely held and likely to underlie misuse of NHST and the drawing of erroneous conclusions. Kline (2004, Chapter 3) listed 13 erroneous beliefs about the meaning and use of NHST, and could cite a range of cognitive results to support his damning critique of NHST. The breadth and diversity of criticisms of NHST are notable: Reform advocates describe a range of severe problems, from fundamental flaws in the logic and theoretical basis of NHST, to deficiencies in how students learn it and researchers use it in practice. They claim it is so broken and so damaging in its consequences that it is beyond repair: Even if it could be taught and understood better, that would be insufficient to justify continuing use of NHST. Evidence of severe and persisting misconceptions about $p$ values and their meaning, even on the part

of many teachers of statistics (Haller & Krauss, 2002), is a major part of the case against NHST.

We will mention just one recent cognitive finding about $p$ values before moving on to consider cognitive evidence about CIs and estimation. Kline's (2004) Fallacy 5 about $p$ values is the incorrect belief that $p$ is the probability a result will replicate. More fundamentally, people underestimate the extent of sampling variability and so believe a *replication* is likely to give similar results. Lai, Fidler, and Cumming (2009) asked authors of journal articles in the disciplines of psychology, medicine, and statistics questions about what $p$ values are likely to be obtained on replication of an initial experiment that had given a particular $p$ value. They found that respondents in general severely underestimated the extent to which replication $p$ is likely to differ from the initial $p$ value. This result generalizes the previous conclusions that people, and notably researchers, tend to underestimate the extent of sampling variability. It also supports Kline's identification of his $p$ value Fallacy 5, the replicability fallacy. This issue of $p$ values and replication is taken further by our final simulation example below.

### Cognitive Evidence About Estimation

Turning to estimation, there are claims that CIs are more easily and accurately learned and understood than NHST. Schmidt and Hunter (1997), for example, claimed that:

> Point estimates and their associated CIs are much easier for students and researchers to understand, and as a result, are much less frequently misinterpreted. Any teacher of statistics knows that it is much easier for students to understand point estimates and CIs than significance testing with its strangely inverted logic. (Schmidt & Hunter, 1997, p. 56)

Fidler (2005, Chapter 10) reported three studies of how undergraduate students understand CIs. The students had completed at least one and as many as four statistics courses, and all had encountered CIs and NHST. She found that results presented as CIs were somewhat more likely than an NHST presentation to prompt students to consider effect size and to give, overall, a better conclusion. On the other hand, many students misunderstood CIs as giving descriptive rather than inferential information. Results like these give guidance for the design of improved teaching about CIs.

Fidler and Loftus (2009) reported two studies with graduate and senior undergraduate students that compared interpretation of results reported as CIs or NHST. They presented students with the results of a study

stated to have low power and to give a substantial effect size that was statistically nonsignificant. They found that CIs presented as error bars led to considerably better interpretation than NHST, and in particular helped students avoid the error of interpreting nonsignificance as implying zero effect.

What little cognitive study there has been of researchers and CIs has identified some reasonable understanding and also a range of misconceptions. Belia, Fidler, Williams, and Cumming (2005) studied interpretation of error bars displayed on two means in a figure. Belia et al. sent e-mails to invite researchers who had published articles in psychology, behavioral neuroscience, and medical journals to visit a website, where they saw an applet showing a figure with error bars. They were invited to note the error bars, then click on the applet to indicate the separation between the two means they judged to correspond to statistical significance at the .05 level. For different respondents the error bars were labeled as 95% CIs or as standard error (SE) bars. Judgments varied widely over respondents, suggesting the task was unfamiliar, and few responses were close to accurate. Respondents who saw 95% CIs often set them to touch end to end, whereas such error bars overlap by 25% of total CI width when $p \approx .05$ (Cumming & Finch, 2005). Conversely, respondents who saw SE error bars often set them too close together: A gap of about 1 SE is needed between the two sets of bars for $p \approx .05$. In general, researchers seemed to make little distinction between 95% CI bars and SE bars, despite the former being about twice the length of the latter. It is highly unfortunate, and a big problem, that the identical error bar graphic is used with such different meanings. In addition, few respondents realized that such comparisons of error bars on two means cannot be used to judge $p$ or statistical significance for a repeated measure design. Overall, Belia et al. identified a range of serious problems with the ways many researchers interpret even a simple figure of two means with error bars.

Cumming, Williams, and Fidler (2004) reported a similar Internet-based study that investigated how researchers think of 95% CIs in relation to replication. Researchers appreciated that CIs give information about where the means of replication studies are likely to fall and that such means are distributed approximately normally. However, the results suggested most researchers believe a 95% CI has a .95 chance of including the mean of a replication experiment, whereas that chance is actually, on average, .83 (Cumming & Maillardet, 2006). Therefore, researchers somewhat underestimated the extent of sampling variability of replication means.

Coulson, Healey, Fidler, and Cumming (2010) reported comparisons of how CIs and NHST are interpreted by researchers. They sent e-mails to authors of journal articles that asked simple questions about the interpretation of the results of two studies, one that gave $p = .02$ and the other

$p = .22$. They found somewhat better interpretation when the results were presented using CIs in a figure than as NHST, but the main finding was that CIs needed to be thought of as intervals if their advantages were to be realized. Using CIs merely to note whether a result is statistically significant or not led to misconception similar to that prompted by NHST presentation.

Finally in this brief review of research on CIs, we report a study by Faulkner, Fidler, and Cumming (2008) in which we examined statistical practices used in 193 reports of randomized control trials (RCTs) of psychological therapies published in leading journals. NHST was used in 99% of the reports, but only 31% reported and interpreted the size of the effect of the therapy under study—and only 2% used CIs. We also presented evidence that clinical psychologists want information about how large an effect a therapy is likely to yield. Therefore, we concluded that most RCTs do not provide the information that is most relevant and useful for clinical practice—they do not present the evidence needed for EBP in clinical psychology. We suggested that using CIs to analyze data and present results would give clinicians better guidance. Fidler, Faulkner, and Cumming (2008) explained with examples how RCTs could be analyzed and reported using CIs.

Our conclusion is that the balance of cognitive evidence to date is clearly in favor of CIs over NHST. The finding that many statistics teachers show NHST misconceptions (Haller & Krauss, 2002) is especially telling and suggests it would be hard to salvage NHST by attempting to improve how it is taught. Further cognitive and teaching investigation is needed, especially to improve how CIs are understood and the graphical conventions for presenting CIs. Even so, we maintain that current cognitive evidence supports our case that best practice, and thus ethical practice, is to use estimation where possible, in preference to NHST.

Researchers who wish to use CIs might consult Cumming and Finch (2005), Cumming and Fidler (2010), and Cumming (2011). As a first step, try formulating research goals as estimation questions like "How much?" and "To what extent?" rather than as null hypotheses. Second, seek options within your familiar statistical software to calculate CIs and include them as error bars in figures. In many cases this is possible, even if default settings give results in terms of NHST rather than CIs.

## Statistical Cognition Research

We mention three selected further issues on which research can help guide EBP based on estimation. First, a core issue is that the definition and basic interpretation of a CI is troublesome. The frequentist definition refers to a notional infinite set of replications of an experiment, in which 95% (if we consider a 95% CI) of the calculated intervals will include the unknown

population parameter being estimated. Given this, how should we think about and describe a single interval? Hoenig and Heisey (2001) made the interesting speculation that "imperfectly understood confidence intervals are more useful and less dangerous than imperfectly understood *p* values and hypothesis tests. For example, it is surely prevalent that researchers interpret confidence intervals as if they were Bayesian credibility regions; to what extent does this lead to serious practical problems?" (p. 23).

This suggestion accords with our teaching and research experience with CIs but needs to be investigated empirically.

Second, most discussion, including our own, about CIs focuses on simple cases and univariate dependent variables. To what extent can the case for estimation be extended to more complex designs and multivariate data sets? One basic issue is the choice of effect size measure and the study of how well people can understand and interpret values expressed in that measure. Measures of percentage of variance, which are common in multivariate situations, may be particularly difficult to picture and appreciate well. Then we need to consider CIs on such measures, so that interpretation can take account of precision of estimates. CIs on root mean square error of approximation (RMSEA) values are often reported as part of structural equation modeling (SEM), and the availability of such CIs and their value for assessing evidence about goodness of fit, and for making model comparisons, are important advantages of the RMSEA index (Stevens, 2009, p. 569). It is admirable that the value of CIs is recognized in this case. However, both statistical development and cognitive research are needed before EBP centered on estimation can be widely achieved with multivariate and other complex designs.

Finally, can the claims of Meehl and Gigerenzer that we mentioned near the start of this chapter be supported empirically? It would be especially interesting to explore how working with CIs might encourage researchers to think in terms of estimating effect sizes, rather than rejecting null hypotheses, and of quantitative theories, rather than mere dichotomous hypotheses. Such cognitive research could help guide fundamental and substantial change in how psychology conducts its theorizing and its empirical research.

Given psychology's knowledge of human cognition and its well-developed experimental methods for studying cognition, psychology is uniquely placed to provide the cognitive evidence needed for EBP in statistics. Developing statistical cognition as a research field, as well as building the evidence base needed for ethical statistical practice, is a great service that psychology can undertake to enhance statistical practice across a wide range of disciplines.

To summarize, ethical statistical practice needs to be based on the best evidence, meaning cognitive and statistical evidence as to what statistical

techniques are best. Currently, cognitive evidence supports our contention that estimation should be preferred over NHST. The evidence is not complete, and there is room for improved ways to teach and use CIs, but even now estimation can improve the efficiency and effectiveness of research in psychology and help development of a more sophisticated, quantitative, and cumulative discipline.

## NHST, Estimation, and Replication: An Example

We close with a replication experiment to give one further example of how estimation is superior to NHST and of the cognitive research needed to support EBP in statistics. Replication is fundamental to science: A research finding is rarely considered established until it has been replicated, typically by more than one researcher working under somewhat different conditions. Replication can overcome sampling variability—the initial finding was not just a chance fluctuation—and also, if replication is observed over a variety of conditions, establishes some robustness of an effect (see McArdle, Chapter 12, this volume, on the importance of replication).

Given this importance of replication, it is reasonable to ask of any technique for statistical inference what information it gives about replications of an original experiment. Here we refer to replication experiments assumed to be identical with the original but with new, independent samples of participants. Thus, we are concerned with replication to overcome sampling fluctuations, rather than replication to establish robustness via variations in experimental conditions.

We consider a simple experiment that compares two independent groups of participants: an experimental group that has experienced our new pep-up therapy and a control group that spent the same amount of time discussing hobbies. The dependent variable is a measure of well-being taken after the therapy or hobbies discussion. Each group is of size $N = 32$. We assume the two populations, of control and experimental well-being scores, are normally distributed. For simplicity we also assume each has known standard deviation $\sigma = 20$. Figure 11.1 shows a dot plot of 32 scores for each group, the two sample means, and their 95% CIs. It shows the $p$ value for the $z$ test of the null hypothesis of no difference. It also shows the difference between the sample means and the 95% CI on that difference. NHST draws a conclusion based on the $p$ value, whereas estimation interprets the CI on the difference between the means. Figure 11.1 shows all the information a researcher has: the data for one experiment and the $p$ value and CIs calculated from those data.

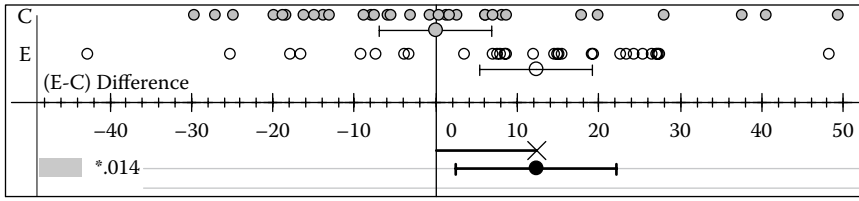



**FIGURE 11.1**
Simulated data for two independent groups, showing all the information a researcher typically has. The upper dot plot shows the $N = 32$ data values for the control (C) condition, with its sample mean and 95% CI just below. The lower dot plot similarly shows 32 data values, mean, and CI for the experimental (E) condition. A floating difference axis for the (E – C) difference is shown, with its zero lined up with the control sample mean. The difference between the sample means is marked by the cross and solid horizontal line and also the black dot—which is shown with the 95% CI for that difference. At left is the $p$ value for testing the null hypothesis of no difference between E and C.

Figure 11.2 also shows the populations from which the scores were sampled and indicates with a dashed vertical line the true difference between the population means, which we assume is 10 points, or 0.5σ. Thus, the population effect size is Cohen's δ = 0.5, a medium-sized effect. The researcher, of course, never knows the populations or that difference, but assumes some such populations exist—the aim of the experiment is to draw a conclusion about them, usually to estimate the difference between the means, which is the effect of the therapy. NHST leads either to rejection

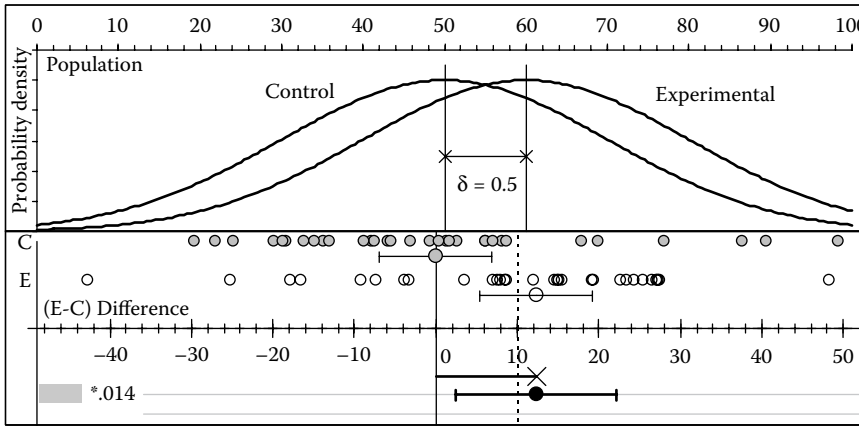



**FIGURE 11.2**
The simulated data shown in Figure 11.1, but with the addition of the underlying control (C) and experimental (E) populations. These are assumed to be normally distributed, with means of 50 and 60, respectively, and each with an SD of 20, assumed known. The true difference of 10, or 0.5σ, is shown and is indicated on the floating difference axis by the vertical dashed line. Thus, the population difference is Cohen's δ = 0.5, a medium-sized effect.

of the null, and a conclusion the population means differ significantly, or to nonrejection. Estimation, by contrast, uses the difference between the sample means as our best point estimate of the effect of the therapy and provides information (the margin of error, i.e., the length of one arm of the CI) about the precision of that estimate.

We would encourage the researcher to observe Figure 11.1 and imagine, or visualize, what is likely to happen on replication. How widely is the sample mean likely to vary from replication to replication? How widely are the *p* values likely to vary? Note that in our simplified example, with $\sigma = 20$ assumed known, the CI width will be the same for every replication, whereas in the more usual situation in which population standard deviation (SD) is estimated from sample SD in each experiment, CI width will also vary from replication to replication.

Figure 11.3 illustrates 24 replications of the original experiment, which appears as the first experiment at the bottom. Note how the 25 means and their 95% CIs bounce around on either side of the dashed line. This is the diagram used in textbooks to explain what the 95% means: In the long run, 95% of such intervals will include the population parameter, marked here by the dashed line. In Figure 11.3, 3 of the 25 intervals happen to just miss the line. The figure is a familiar depiction of the extent of sampling variability.

The most striking aspect of Figure 11.3 is the enormous variation in *p* values over replications: from less than .001 to .848. It seems that *p* can take more or less any value at all! Note that our experiment, which uses two independent groups of size 32 to investigate a medium-sized true effect, has power of .52, which makes it typical of published research in many fields of psychology (Maxwell, 2004). Therefore, the astonishingly wide variation in *p* cannot be attributed to any weirdness of our chosen example. There is a dramatic contrast between the familiarity of sampling variability of sample means and CIs and the unfamiliarity of any mention of variation in *p* over replication. Every statistics textbook takes pains to explain sampling distributions and the SE, but we know of no textbook that even mentions the corresponding sampling variability of *p*; instead, the focus is usually on calculating *p* precisely and then basing decisions on the exact *p* value.

Thinking of replication emphasizes that our single result, as in Figure 11.1, is one chosen at random from an infinite set of potential results, 25 of which appear in Figure 11.3. Consider Figure 11.1 again: Does the single *p* value give information about the whole set of potential results—or potential *p* values? Given the enormous variation in *p* with replication, surely the answer must be that it gives virtually *no* information about the whole set of potential results? What about estimation: Does the mean difference and CI on that difference shown in Figure 11.1 give any idea of the infinite set of potential results? Yes, it does because the width of

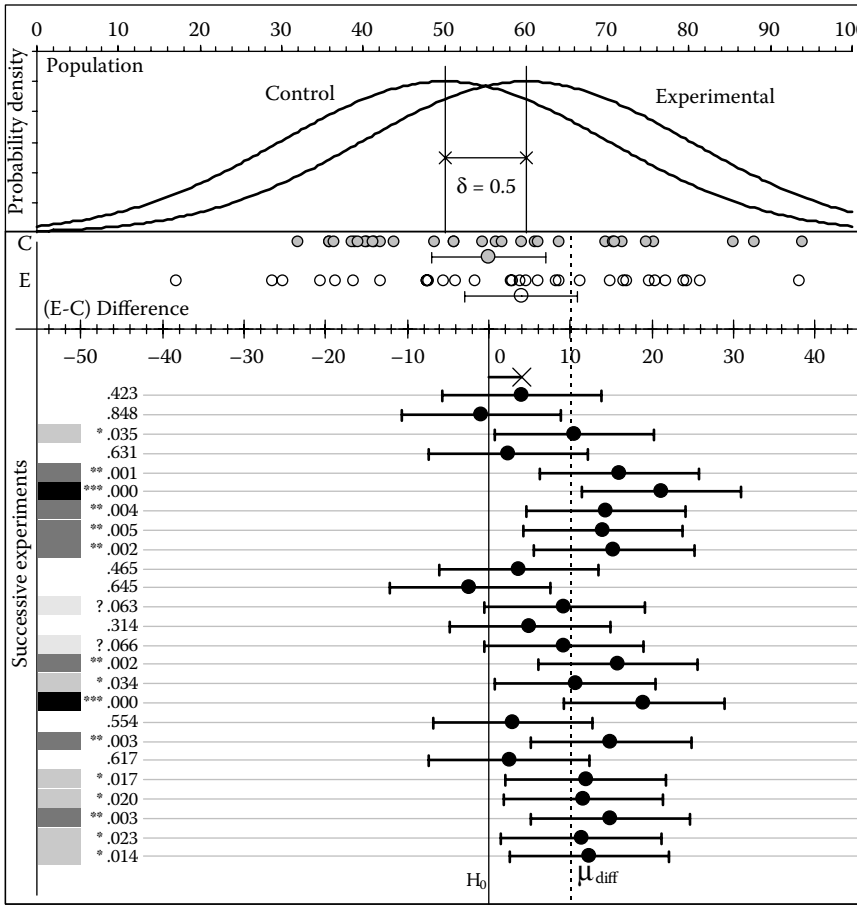



**FIGURE 11.3**
Simulated results of a further 24 replications of the experiment shown in Figures 11.1 and 11.2. The original experiment appears at the bottom. Each experiment is identical, except that for each a new independent random sample is drawn from each population. The figure illustrates how the sample mean differences, as well as the CIs on these differences, bounce around over the replications. It also illustrates how dramatically the $p$ values vary with replication. The $p$ values are marked with conventional asterisks (*** for $p < .001$, ** for $.001 < p < .01$, and * for $.01 < p < .05$) and with "?" for $.05 < p < .10$. The patches also indicate $p$ values by varying from black for ***, through shades of grey, to white for $p > .10$.

the CI gives some indication of how widely the mean differences bounce around over replications.

Given the single result of Figure 11.1, would you prefer to be told the difference between means and in addition just the $p$ value, or in addition the CI on that difference? Surely the latter is much more informative about the whole set of potential results and about what is likely to happen on

replication? We conclude that thinking about replication gives one further reason for regarding estimation as more informative than $p$ values. We put this forward as an additional reason for making the shift from NHST to estimation.

Cumming et al. (2004) and Cumming and Maillardet (2006) explained that the probability is .83 that a replication mean will fall within the 95% CI found by the initial experiment. Cumming (2008) investigated the distribution of the $p$ value and illustrated how greatly $p$ values vary over replications. Cumming concluded that researchers rely much too heavily on particular $p$ values and that anything but very small values ($p < .001$ or, just possibly, $p < .01$) conveys very little useful information indeed.

Is there cognitive evidence available to support this reform argument based on replication? As mentioned earlier, Cumming et al. (2004) reported evidence that researchers seem to have a reasonable understanding of CIs in relation to replication, although they somewhat underestimate the extent replication means vary. Also as mentioned earlier, Lai et al. (2009) found researchers in general severely underestimate the extent $p$ values vary with replication. These findings suggest researchers may have a somewhat more accurate appreciation of what CIs rather than $p$ values indicate about replications. However, we know of no attempt to study researchers' thinking about the infinite set of replication results, expressed in $p$ value or in CI form as illustrated in Figure 11.3. Our additional reason for supporting the shift from NHST to estimation would be stronger if there were cognitive evidence to reinforce our argument above based on Figures 11.1, 11.2, and 11.3. Such evidence is required if our replication argument is to meet our standards for contributions that seek to shape good and therefore ethical EBP in statistics.

## Conclusion

Widespread adoption of evidence-based medicine (EBM) was largely justified by its proponents on ethical grounds. The extension of the same reasoning to promotion of EBP in psychology and other disciplines relied similarly on ethical considerations. EBM and EBP in psychology are now widely accepted and expected around the world as best professional practice. We argued here that researchers should, correspondingly, be able to justify their choices of research methods, including statistical techniques, in terms of evidence. Evidence about statistical appropriateness and performance is important, but cognitive evidence of effectiveness is especially important. Results from statistical cognition research should guide better learning about statistical concepts, reduction in misconception, design

of better graphics, and choice of statistical techniques that readers can understand more readily and accurately.

We illustrated our argument for EBP in statistics in the context of criticisms of NHST and advocacy of estimation based on CIs. We focused on the potential for estimation to give better representations of results and more justified conclusions from data, and also to encourage a shift from dichotomous decision making and the generation of richer, more quantitative theories in psychology. Further cognitive evidence is needed, especially to guide how estimation can be taught better and estimation practices improved. The cognitive evidence to date, however, supports our argument that estimation should be preferred wherever possible to NHST. Thus, using estimation is, we contend, best practice and therefore ethical practice.

## References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396.

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7,* 20–39.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement, 1*(26). doi:10.3389/fpsyg.2010.00026. Retreived from http://www.frontiersin.org/psychology/quantitativepsychologyandmeasurement/paper/10.3389/fpsyg.2010.00026

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286–304.

Cumming, G. (2011). *Introduction to the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge. Manuscript in preparation.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology, 217,* 15–26.

Cumming, G., & Fidler, F. (2010). Effect sizes and confidence intervals. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 107–124). New York: Routledge.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., …Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60,* 170–180.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11,* 217–227.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research and Therapy, 46,* 270–281.

Fidler, F. (2002). The fifth edition of the APA *Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62,* 749–770.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology*. Unpublished PhD thesis, University of Melbourne. Retrieved from http://www.botany.unimelb.edu.au/envisci/docs/fidler/fidlerphd_aug06.pdf

Fidler, F., Faulkner, S., & Cumming, G. (2008). Analyzing and presenting outcomes: Focus on effect size estimates and confidence intervals. In A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 315–334). New York: OUP.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie /Journal of Psychology, 217,* 27–37.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8,* 195–204.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1–20.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Hope, T. (1995). Evidence based medicine and ethics. *Journal of Medical Ethics, 21,* 259–260.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington DC: National Academy Press.

John, I. D. (1992). Statistics as rhetoric in psychology. *Australian Psychologist, 27,* 144–149.

Kline, R.B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington DC: American Psychological Association.

Lai, J., Fidler, F., & Cumming, G. (2009). *Subjective* p *intervals: Researchers underestimate the variability of* p *values over replication*. Manuscript submitted for publication.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Norcross, J. C., Beutler, L. E., & Levant, R. F. (Eds.). (2006). *Evidence-based practices in* mental *health: Debate and dialogue on the fundamental questions.* Washington, DC: American Psychological Association.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling. A quiet methodological revolution. *American Psychologist, 65,* 1–12.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–63). Mahwah, NJ: Lawrence Erlbaum.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Routledge.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

# 12

# *Some Ethical Issues in Factor Analysis*

**John J. McArdle**
*University of Southern California*

## Methodological Issues

This is a book about ethics in data analysis. So we might begin by asking, "Why worry about ethics in data analysis? Isn't this already taken care of by good science training?" The answer of course is, "Yes." Very early in our careers, as early as in elementary school, we are taught to follow and respect the so-called "scientific method" as a guide to obtaining such useful and replicable results. We can all agree that sturdy scientific results require sturdy scientific principles. A key reason we worry about this topic is because we have to trust one another in the creation of sturdy scientific results. But because we are all so trustworthy, what could possibly be the problem?

Unfortunately, we are also well aware of publicized violations of this trust: We know we should not simply graft cancer-free tails onto otherwise sickly rats, and we should not claim to have created a device that creates useful energy from nothing, and we know we should not publish algebraic proofs developed by others as if they were our own invention. We usually assume these are charades posing as good science and believe that we would never knowingly create such problems ourselves. Never! But then we run our key hypotheses about group differences using a one-way *analysis of variance* (ANOVA) and find probability values that are just larger than the arbitrary $p < .05$. We consider using multiple $t$ tests instead of the one-way ANOVA, or using one-tailed tests, but our early statistics training makes us shudder at this obvious violation of statistical laws (see Scheffe, 1959). So we start to say this result is "approaching significance," essentially creating our own new level of probability that is without bounds. Or we eliminate some offending data (i.e., possibly true outliers), or we try a transformation of the dependent variable (DV), and rerun the ANOVA to see whether we can achieve the seemingly magic numbers required for publication.

The ethical basis of this scenario is really no different than when in a more complex modeling analysis, we see that a model fitted using our a priori logic does not seem to fit by acceptable standards (i.e., root mean square error of approximation, $e_a < .05$; see Browne & Cudeck, 1993). In this case, we try hard to find another model that is very close to our original model and does meet the arbitrary standards of good fit (see Brown, 2006). Because the second one will serve our purposes, we report it and, unfortunately, we use standard statistical tests and treat the model as though it were our starting point. In the thrill of a publishing moment, we may completely forget about our starting point model—and our ethical virtues.

As in many areas of life, the reason we cross these ethical boundaries in data analysis is because we desire novel and reasonable results, and we often simply blame the rules of publication for having absurd criteria. We have also become aware that the search for connections and dynamic influences and causes is fairly complex and not easy to describe, so we conclude that our little deception will do no real harm in the long run. We come to realize a good description of what we are doing is much more like we are making a "principled argument" (Abelson, 1995). So we are encouraged to stretch our ethical boundaries, and, unfortunately, they become less clear.

But almost any self-evaluation will lead us to be rightfully concerned that we may be carrying out science without a firm ethical compass. Since the turn of the 20th century, there has been a collective effort to develop helpful ethical principles in all kinds of empirical research studies. Ethical principles were important in the development of the cooperation between the scientist, producer of results, and the consumers of results—and we hope the newspaper reporters do not criticize the scientists. At the same time, the need for accurate, replicable, and reliable information flow is needed for the accumulation of studies and "replicable results" in the "soft-fact" sciences.

In this chapter we will highlight some ethical dilemmas of one widely used technique—*factor analysis* (FA; see McDonald, 1985; Mulaik, 2009). The history of psychological statistics shows a great respect for a priori testing of formal hypothesis (Fisher, 1925) and has led to many organized and successful research programs. Unfortunately, this also led to skepticism and disdain for exploratory data analysis procedures (e.g., Tukey, 1962, 1977), although not all of this criticism is warranted. The previous divisions between confirmation and exploration are apparent in FA as well, but some confusion has led researchers to state that they used confirmatory methods when, in fact, their work was largely exploratory in nature.

To resolve these problems, we try to show how a *structural factor analysis* (SFA; Albert, Blacker, Moss, Tanzi, & McArdle, 2007; Bowles, Grimm, &

McArdle, 2005; Cattell, 1966; McArdle, 1996; McArdle & Cattell, 1994) approach to FA allows us to use the full continuum and avoid the artificial semantic differences of confirmation and exploration. This approach to FA relies on both "confirmation" and "exploration" and is consistent with a "functionalist" view of psychological research (as in McArdle, 1994a, 1994b; McArdle & Lehman, 1992). Further definitions are listed in Table 12.1, and we return to this table at various points in this discussion.

In this chapter, some technical issues of SFA are presented first, but not in great detail, and these are quickly followed by a case study example using real cognitive data. This leads to a discussion of what others have done about ethical problems in FA, as well as five suggestions for future work. My hope is that this approach will lead us to think that ethical principles can always be followed in data analysis. This also leads us to see that the main ethical problem we face in SFA is what and how much should we tell others about what we have done. The ethical answer is clear—we should document all our SFA work and tell others whatever we can. In practice, this is not so easy.

## Statistical Background

### The Statistical Basis of Factor Analysis

The statistical and psychometric history of FA is long and contains many specialized techniques and colorful concepts (see McDonald, 1985, 1999; Mulaik, 2009). Most of the older techniques will not be used here, and we will only discuss techniques based on the contemporary principles of *maximum likelihood estimation* (MLE; see Lawley & Maxwell, 1971). This approach allows us to carry out both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) using *structural equation modeling*

**TABLE 12.1**

A Continuum of Factor Analysis Techniques

| | | |
|---|---|---|
| Confirmatory | ..... | Exploratory |
| More theory | ..... | Less theory |
| More restrictions | ..... | Few restrictions |
| Overidentified | ..... | Exactly identified |
| More *df*s | ..... | Less *df*s |
| Less absolute fit | ..... | Greater absolute fit |
| Ample stat tests | ..... | Few stat tests |
| Seemingly strong | ..... | Seemingly weak |

(SEM) computer algorithms (e.g., AMOS, LISREL, semR, OpenMx, M+). This approach used here also allows us to distinguish the most important features of EFA–CFA differences—how many parameter "restrictions" are placed on the data on an a priori basis. Repeatedly, we note that the a priori nature of this selection is still needed for appropriate statistical tests based on the *chi-square* ($\chi^2$) distribution. A *degree of freedom* (*df*) is a model expectation that can be incorrect (i.e., a way to go wrong), so the number of *df*s is used in many indices of model parsimony (to be described).

The techniques of *exploratory factor analysis* are used in most classical FA. A set of unobserved common factors is thought to be responsible for the observed correlations among observed variables ($V$). In the EFA approach, we propose a specific number of common factors $k$, possibly with a specific hypothesis, but we almost always explore several models, from no common factors ($k = 0$) to as many common factors as possible ($k = v/2$). In contemporary terms, the number of common factors is predetermined, but the specific set of common factor regression coefficients, termed *factor loadings*, is "exactly identified." This means the factors can be "rotated" to a simpler, possibly more meaningful solution, with no change of common variance and no change in overall fit. A statistical test for the number of common factors in EFA is typically conducted as a sequence of nested chi-square tests, formally based on the size of the residual correlations, and various approaches and indices have been suggested to determine an adequate number of factors (Browne & Cudeck, 1993; Cattell, 1978; Lawley & Maxwell, 1971; McDonald, 1985).

The term *confirmatory factor analysis* was popularized by Jöreskog (1966, 1969, 1977) and used by Tucker and Lewis (1973) to describe the new SEM-based approach to FA. Here we follow their lead and fit "overidentified models" with specific restrictions on the factor loadings. It turned out that classical test statistics (e.g., chi-square) could be applied to this kind of a problem, so CFA fit in very well with ANOVA and other pure forms of statistical inquiry. As a result, many data analysts started to search for clear and a priori factor patterns. Unfortunately, the required level of precision seemed to be lacking. In response, many CFA researchers "trimmed" their data sets and/or parameters or relied on exploratory "modification indices" and "correlated errors," so models appeared to be CFA and benefited from the statistical tests (e.g., Brown, 2006). For similar reasons, Cattell (1978) suggested that we substitute the term proofing FA for MLE–CFA, although this insightful terminology never became popular.

## Initial Structural Factor Analysis Models

For the purposes of this discussion, let us assume we have measured six different variables ($v = 6$) on a number of different individuals ($N > 10*v$). When we apply the techniques of FA to this kind of data, we are trying

to understand the best way to represent the observed variables in terms of unobserved factors. A series of alternative models is presented in the path diagrams of Figures 12.1 and 12.2. In these diagrams, the observed variables are drawn in squares, and the unobserved variables are drawn as circles. One-headed arrows represent a direction influence, typically

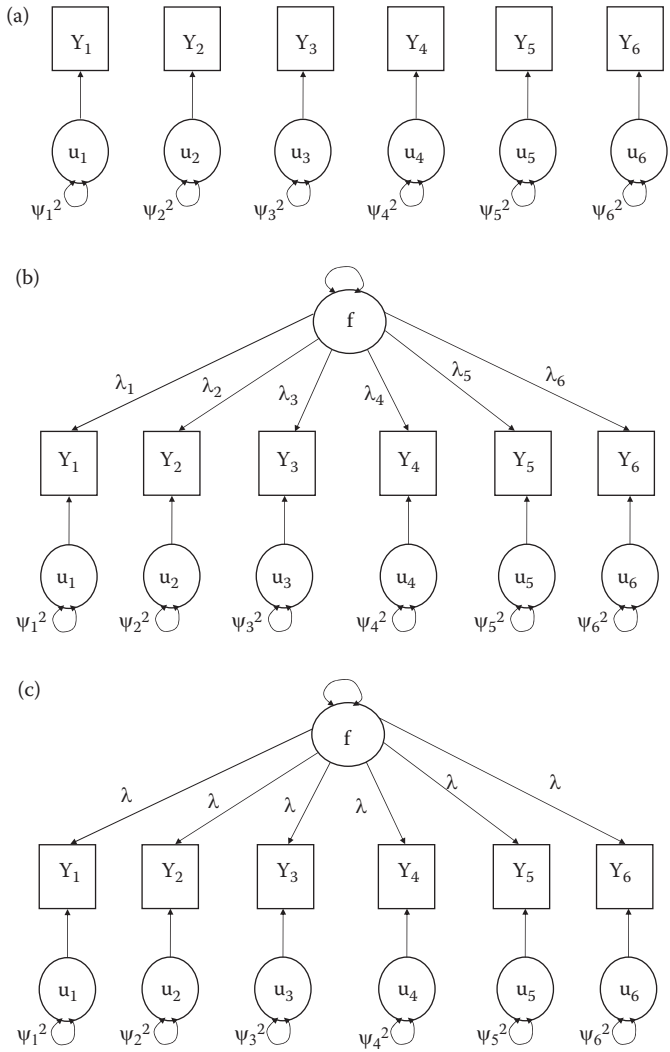



**FIGURE 12.1**
Alternative common factor models. (a) Six variables—zero common factors but six unique factors ($df = 15$). (b) Spearman-type (1904) one common factor model ($df = 9$). (c) Rasch-type (1961) one common factor model ($df = 14$).

termed *factor loadings*, and two-headed arrows represent nondirectional influences, such as "variance" or "covariance" terms.

The zero-factor model is almost always useful as a starting point or baseline model, and this is presented as a path diagram in Figure 12.1a. Here we assume each of the observed variables is composed of the influence of only one unique factor, labeled $u_v$, with fixed loadings (unlabeled) but free unique variances, labeled $Y_v^2$. In this model, the unique latent scores

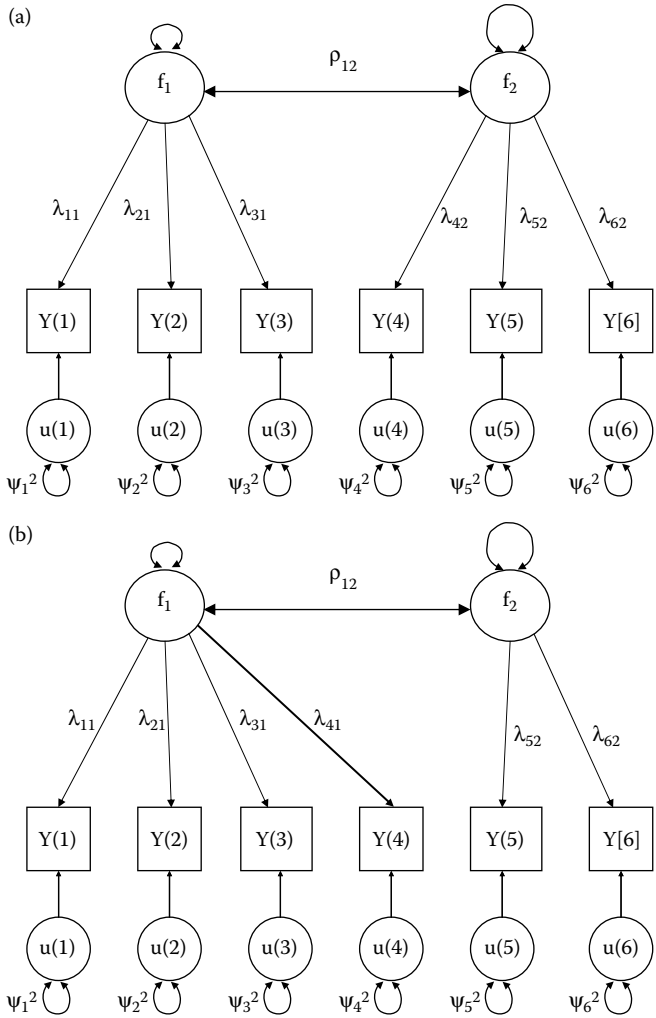



**FIGURE 12.2**
Alternative two common factor models. (a) "Simple structure" two common factor model (*df* = 8). (b) "Non-nested" two common factor model (*df* = 8). (c) "Exactly identified" two common factor model with oblique constraints (*df* = 4).

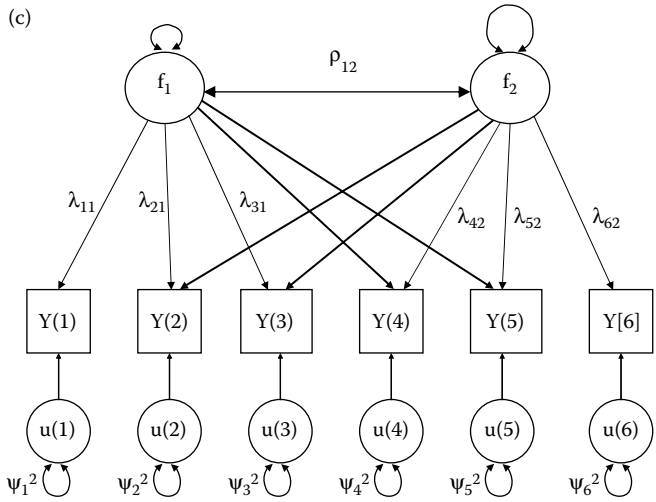



**FIGURE 12.2 (Continued)**
Alternative two common factor models. (a) "Simple structure" two common factor model (*df* = 8). (b) "Non-nested" two common factor model (*df* = 8). (c) "Exactly identified" two common factor model with oblique constraints (*df* = 4).

are thought to produce the variation we observe. This model restricts the unique variables to have zero correlations, and each restricted correlation counts as *df* = 15, so this is our simplest, most parsimonious model.

If the model of "no correlation" is true, we typically state that "this zero-factor model fits the observed data," and we have no need to go further with data analysis. However, if there are significant correlations among the observed scores, then this simple model does not completely capture the observed correlations, and we then typically say this simple latent variable model does not fit the data. One common statistical test used here is based on the likelihood ratio test (LRT), formed from the *likelihood of the* original data matrix compared with the likelihood of the model estimated matrix (i.e., a diagonal). It is often briefly stated that "under certain regularity conditions" such as "the unique factor scores are normally distributed," the LRT is distributed as a chi-square index that can be used to evaluate the model misfit—that is, with low chi-square relative to the *df*s taken as an indication of good fit (for details, see Browne & Cudeck, 1993; Lawley & Maxwell, 1971; McDonald, 1985).

The next theoretical model is based on the one common factor model, and this can be seen in the path diagram of Figure 12.1b. In this diagram, the observed variables are drawn in squares, and the unobserved variables from our theory are drawn as circles. In Figure 12.1b, we assume each observed variable is composed of the influence of both its own unique factor as before but also that there is one common factor (labeled *f*), each

with its own loadings (labeled $\lambda_v$). In this model, the two latent scores are thought to produce the variation we observed, but only the latent common factor is thought to produce the covariance of the observed scores by a simple pattern of expectations (i.e., without details, $\sigma_{ij} = \lambda_i \times \lambda_j$). This model also restricts the unique variables to have zero correlation but requires six additional factor loadings to do so (so $df = 9$). The test of this one-factor LRT hypothesis is that when all model expectations are removed, there is no remaining correlation among the observed scores.

## Variations on the One-Factor Concept

Although it is clear that the one-factor model is a strong hypothesis, it is rarely used in this way (see Horn & McArdle, 2007). Part of the reason for this hesitation may be because of the typical problems faced by factor analytic researchers. For example, to obtain the precision required by the LRT, we must have an a priori hypothesis of one common factor for a specific set of variables. Of course, it is not uncommon for researchers to drop participants who either do not meet some sampling requirements (i.e., required language, age > 50, Mini-Mental Status Examination > 24) or whose behavior seems aberrant (i.e., outliers). Although these can be reasonable criteria on practical grounds, any nonrandom sample selection may violate the assumptions of the statistical tests. But perhaps more critically, researchers routinely drop some of variables that are "not working," or rescale some of the variables to remove "odd distributions," and almost any nonrandom variable selection has an impact on the statistical tests. It is not that these are always horrific practices, but it is clear that the standard statistical tests are no longer appropriate after these kinds of changes in the data are made (cf. Brown, 2006).

Let us consider one other CFA extension—the Rasch-type model (see Embretson & Reise, 2000; McDonald, 1999; Wilson, 2005). From a one-factor starting point, if we fix all model loadings to be identical ($\lambda$), we end up with properties that mimic a Rasch scale—that is, the summation of the scores is parallel to the Rasch-type factor score estimates. This model has a pattern that is even more restrictive ($df = 14$; $\sigma_{ij} = \lambda^2$), so it may not fit the data very well, but it is needed to establish the adequacy of a simple Rasch-type summation scale. From this viewpoint, the Rasch model is a highly restricted test of a formal CFA hypothesis. It makes little difference that this Rasch model is more typically used with items than with scales.

The knowledgeable researcher will notice that no effort is made here to evaluate the utility of what are often termed *correlated errors* using the statistical techniques of "modification indices" (MI; e.g., Brown, 2006). There are several reasons why these parameters and this approach to model

fitting are completely ignored from this point here. The first reason is that this approach allows, and even embraces, the estimation of *correlated specifics* (CS). The problem is that almost any CS approach does not match the basic goals of common FA at all (e.g., see Meredith & Horn, 2001; cf. McArdle & Nesselroade, 1994). That is, the estimation of any CS in FA typically attempts to isolate part of the data that cannot be fitted by a specific model. A second reason is that the MI approach, which attempts to recursively locate the single parameter that, if estimated in the model, can alter the model fit the most, does not account for dependencies that are multivariate in nature. It is not surprising that this combined CS–MI approach simply does not even work well as an exploratory tool (see MacCallum, Roznowski, & Necowitz, 1992). From a traditional perspective, FA modeling based on this CS–MI approach is viewed as a misunderstanding of the analytic goals of FA.

### Expanding Structural Factor Analysis Models

Continuing with the example at hand, Figure 12.2 extends these FA concepts a bit further by proposing a less restrictive two-factor hypothesis for the data. In this model, the first three variables are thought to load on the first common factor ($f_1$), and the last three variables are thought to load on a second factor ($f_2$). This is a classic example of a "confirmatory factor" model. The model expectations within each set are the same as before ($\sigma_{ij} = \lambda_i \times \lambda_j$), but across sets of variables we now add a parameter ($\sigma_{ij} = \lambda_i \times \rho_{12} \times \lambda_j$), so this model should fit better than the one-factor version. Given all other assumptions, the difference between model of Figures 12.1b and 12.2a is a testable hypothesis (of $\rho_{12} = 1$).

The two-factor model of Figure 12.2b uses the same number of model parameters but places these in a different location so the first factor has four loadings and the second factor has only two loadings. Unfortunately, the number of parameters in each model of Figures 12.2a and 12.2b is the same, so no formal test of the difference is possible. To create such a test, we often create a composite model where both loadings are allowed. Of course, such a model is no longer "simple" in the sense that a variable such as $Y_4$ can load on both common factors. However, we can form reasonable LRTs to try to determine which model is best for our data.

Following a similar logic, we can create a model where we allow as much room to fit as is possible, and the result is Figure 12.2c, where only two of the variables are used as "reference variables" ($Y_1$ and $Y_6$), and the other four are allowed to load on both common factors. Perhaps it is obvious, but all other models here (Figures 12.1a to 12.2b) are formally nested as proper subsets of the model of Figure 12.2c, so all can be fairly compared for fit. Perhaps it is also obvious that our initial choice of the two reference variables was arbitrary. This means that for a model where one or two other

variables ($Y_2$ and $Y_5$) are chosen as reference variables, the misfit would be exactly the same but the parameter values would be different. To wit, there is more than one "exactly identified" two-factor solution that can be fit to solve the same problem. This is a simple example of the problem of "factor rotation"—given a specific number of identifiable parameters (i.e., 10 loadings here), there are many positions that yield the same pattern of expectations and hence the same $df = 4$ and the same misfit.

Using this form of SEM, we can be clear about the exact model that is fit to the data—we can easily present all the summary statistics to be fitted (i.e., correlations) and the exact model (i.e., as a path diagram). This allows others to replicate our model analyses with their own data. Under the assumptions that (a) the data were selected in advance of the analysis and (b) this model was chosen on an a priori basis, then SEM yields statistical tests of the (c) overall fit of the model to the data ($\chi^2$, $\varepsilon_a$, etc.) and (d) individual standard errors for each model parameter ($z = MLE_p/SE_p$). These statistical indices can be used to judge the adequacy of the fit using consensus rules of agreement, but we must be careful about comparing them with a priori distributions if they are not a priori tests.

## Case Study Example

### Cognition Measurement in the Health and Retirement Study

The example presented next comes from our recent work on an FA of cognition measures in the *Health and Retirement Study* (HRS; see Juster & Suzman, 1995). At the start of this analysis, we recognize it is uncommon to ask cognitive questions in large-scale survey research, even though the cognitive status of the respondent is of obvious importance to providing genuine answers (see Schwarz et al., 1999). Indeed, the HRS has a long and reasonable history of using cognitive items for this purpose (see McArdle, Fisher, & Kadlec, 2007).

Following this logic, the specific application presented here uses publicly available data on a small set of cognitive variables ($v = 7$) measured on a large sample of adults (age > 50, $N > 17{,}000$). Table 12.2 is a list of available cognitive variables in the current HRS data. Some incomplete data have been created by the HRS because not all persons were administered all $v = 7$ tests at any sitting, but overall coverage of all variables is reasonable (>80%). Respondent sampling weights will be used here to approximate a sample that is representative of the U.S. population older than age 50 (see Stapleton, 2002). When using the sampling weights, the model must be fitted using alternative estimators (i.e., a variation of MLE allowing weights, termed *MLR*), and the fit can be altered by a constant of kurtosis (*w4*). Although we do report these values here, the appropriate use of weighted

**TABLE 12.2**

A Listing of the Health and Retirement Study Cognitive Measures

1. Immediate word recall (IR; 10 items)
2. Delayed word recall (DR; 10 items)
3. Serial 7s (S7; to assess working memory)
4. Backward counting (BC; starting with 20 and 86)
5. Dates (DA; today's date and day of the week)
6. Names (NA; object naming, president/vice president names)
7. Incapacity (IN; to complete one or more of the basic tests)

And on some occasions …

8. Vocabulary (VO; adapted from WAIS-R for $T > 95$)
9. Similarities (SI; adapted from WAIS-R for $T = 92, 94$)
10. Newly created "adaptive" measures from the WJ-III

WAIS-R, Wechsler Adult Intelligence Scale–Revised; WJ-III, Woodcock-Johnson III.

chi-square tests is not a key issue of this chapter. Weighted summary statistics about these HRS cognitive variables are presented in Table 12.3.

## Considering One Common Factor

The SFA approach used here starts with a sequence of CFAs but ends on a more relaxed set of EFAs. To initiate the CFAs, the models that were first fitted include the zero-factor model (Figure 12.1a) and the one-factor model (Figure 12.1b). The zero-factor model was fitted mainly as a

**TABLE 12.3**

Health and Retirement Study Summary Statistics From Respondent Interviews ($N = 17{,}351$)

(a) Means and Standard Deviations

| | IR[1] | DR[1] | S7[1] | BC[1] | NA[1] | DA[1] | VO[1] |
|---|---|---|---|---|---|---|---|
| | 55.7 | 43.9 | 70.5 | 95.2 | 94.2 | 91.3 | 55.4 |
| | 18.5 | 22.3 | 34.1 | 21.1 | 14.3 | 16.7 | 21.2 |

(b) Correlations

| | IR[1] | DR[1] | S7[1] | BC[1] | NA[1] | DA[1] | VO[1] |
|---|---|---|---|---|---|---|---|
| IR[1] | 1.000 | | | | | | |
| DR[1] | .773 | 1.000 | | | | | |
| S7[1] | .371 | .359 | 1.000 | | | | |
| BC[1] | .189 | .170 | .227 | 1.000 | | | |
| NA[1] | .283 | .280 | .255 | .201 | 1.000 | | |
| DA[1] | .362 | .345 | .381 | .221 | .308 | 1.000 | |
| VO[1] | .385 | .352 | .393 | .185 | .202 | .403 | 1.00 |

Variable abbreviations appear in Table 12.2. Measured at occasion with most cognitive variables; respondent weights used, 36 patterns of incomplete data, coverage >81%; MLE(MAR) using M+; $\chi^2$(diagonal) = 18,521 on $df = 21$; eigenvalues (%) = [42.4, 14.1, 11.9, 11.2, 8.8, 8.2, 6.6, 3.2].

baseline model for comparison, but the second could be defended based on prior cognitive theory going as far back as Spearman (1904; see Horn & McArdle, 1980, 1992, 2007; McArdle, 2007). The goodness of fit of the zero-factor model is very poor ($\chi^2$ = 18,520, *df* = 21, *k* = 1.61, $\varepsilon_a$ = .225), and the one-factor model seems much better ($\chi^2$ = 2,530, *df* = 14, *w4* = 1.58, $\varepsilon_a$ = .102). For illustration, the ML parameter estimates of the one-factor model are presented in Figure 12.3a. Of course, the one-factor results seem to suggest the one common factor model is only adequate for the first two variables, and the other four variables are largely unique.

We next add a test of the Rasch model of one factor with equal loadings, and it seems to fit even worse ($\chi^2$ = 7,033, *df* = 20, *w4* = 1.76, $\varepsilon_a$ = .142). From these initial analyses, we conclude that more than one common factor is likely to be needed to capture all the variation in these cognitive data. The lack of fit of the Rasch model also provides evidence that the HRS cognitive scores should not simply be added together to form an overall score (i.e., see McArdle et al., 2007). It did not matter what variation of the one-factor model is fitted; it is apparent that one factor of the HRS cognitive variables leaves a lot to be desired.

### Considering More Than One Common Factor

The models we have just fit are common using these kinds of cognitive data. The second set of models was decidedly CFA in origin. I (person JJM) asked a more knowledgeable colleague (Dr. John L. Horn, University of Southern California, person JLH) to create a two-factor hypothesis from these data. The conversation follows (from audio tape, 08/24/2002):

JJM: Can you please take a look at this new HRS data set I am now using?

JLH: OK, but I think this is a seriously impoverished data set for any cognitive research.

JJM: Yes, but the sample is very large and representative, over 17,000 people, and the HRS is now using a one-factor model.

JLH: OK, I will show you how bad this is—how about we do an exploratory factor analysis first?

JJM: We could, but that would distort the a priori basis of the chi-square and other statistical tests.

JLH: I agree, but who actually uses those tests anyway? Do I need to remind you that factor analysis is not a statistical problem anyway?

JJM: Are you saying you just can't do it? After 40 years of cognitive research, you don't have any formal a priori hypotheses at all?

JLH: No, I didn't mean that. I can do it. I suggest what you have here is a little factor of short-term acquisition retrieval, and I do mean little, and probably a second common factor based on the rest of them, whatever they are supposed to be.

The results from fitting this kind of what might be said to be a *semi-formal* a priori CFA two-factor model are presented in Figure 12.3b. The fit of the model shows much improvement ($\chi^2 = 222$, $df = 13$, $w3 = 1.50$, $\varepsilon_a = .030$), and a formal test of whether the interfactor correlation is unity ($\rho_{12} = 1$) is indexed by the LRT difference ($\chi^2 = 2{,}308$, $df = 1$). This initially reminds us that when we have $N > 17{,}000$ people we have great power

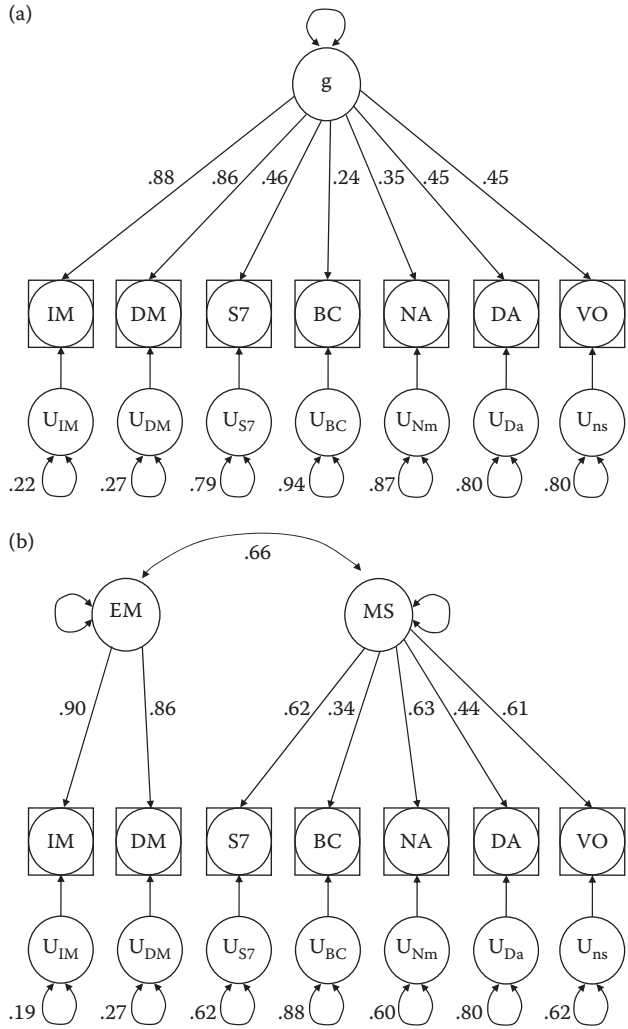



**FIGURE 12.3**
Alternative factor models for the seven HRS cognitive abilities ($N > 17{,}000$). (a) One-factor model results ($\chi^2 = 2{,}530$, $df = 14$, $k = 1.58$, $\varepsilon_a = .102$; standardized MLE listed). (b) Two-factor CFA model results ($\chi^2 = 222$, $df = 13$, $w4 = 1.50$, $\varepsilon_a = .030$). (c) Three-factor CFA model results ($\chi^2 = 214$, $df = 12$, $k = 1.50$, $\varepsilon_a = .030$).

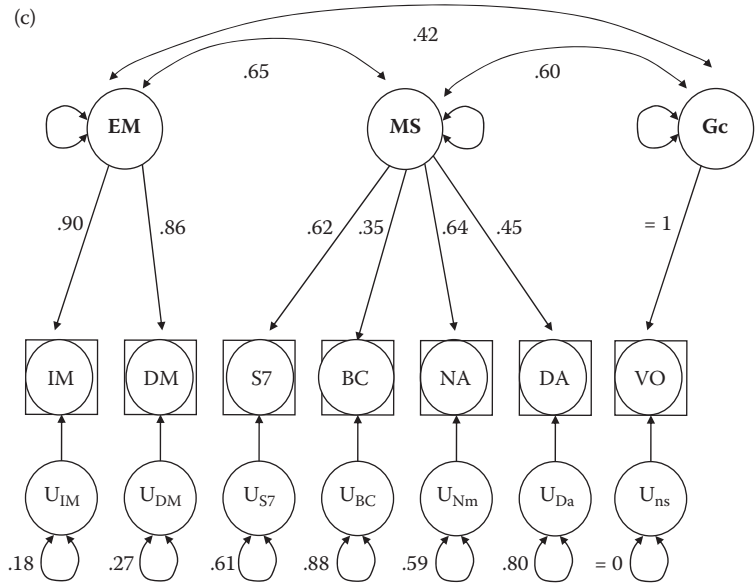



**FIGURE 12.3 (Continued)**
Alternative factor models for the seven HRS cognitive abilities ($N$ > 17,000). (a) One-factor model results ($\chi^2$ = 2,530, $df$ = 14, $k$ = 1.58, $\varepsilon_a$ = .102; standardized MLE listed). (b) Two-factor CFA model results ($\chi^2$ = 222, $df$ = 13, $w4$ = 1.50, $\varepsilon_a$ = .030). (c) Three-factor CFA model results ($\chi^2$ = 214, $df$ = 12, $w4$ = 1.50, $\varepsilon_a$ = .030).

to state that $\rho_{12} = 0.66$ is statistically different than $\rho_{12} = 1$. But the other parameter estimates are more revealing. The isolation of the first two variables load onto a first factor we have labeled *episodic memory* (EM). The second factor has highest loadings for S7 and VO, so it may be a general crystallized (Gc) intelligence factor, but because of all the relatively low level of information required (NA, DA, and BC), we have labeled this as *mental status* (MS). Incidentally, a Rasch version of this two-factor hypothesis does not fit the data very well ($\chi^2$ = 1,962, $df$ = 18, $w4$ = 1.62, $\varepsilon_a$ = .079) and does not seem to fit this model very well. Of course, the model fit could probably be improved further by considering the categorical nature of these three variables (i.e., most people get them all correct). However, it is very clear that the model fits well, and the hypothesis of JLH was clearly confirmed. But this seeming success made us go even further (from audio tape, 08/24/2002):

JJM: So is this enough for now? Are we done? Can we fit it any better?

JLH: Yes. It seems to me that the SAR factor based on the first two variables is reasonable; the next four are simply the mental status of the person, and likely to go together. But the vocabulary is

really a better indicator of crystallized intelligence. Too bad, but the lack of other measures makes vocabulary collapse into the second factor. Can we isolate this one variable in any way?

JJM: Maybe. I will try.

JLH: Incidentally, I think the real problem you have here is that there are no measures of fluid intelligence at all.

The results from fitting this semiformal a priori CFA three-factor model are presented in Figure 12.3c. The fit of the model shows much improvement ($\chi^2 = 214$, $df = 12$, $w3 = 1.50$, $\varepsilon_a = .030$), and a formal test of whether the VO is isolated is indexed by the LRT difference ($\chi^2 = 6$, $df = 1$). Note that no estimate of the uniqueness of VO is estimated because this variable is isolated. This mainly reminds us that there is not much difference between the model of Figure 12.3b and 12.3c in this context. From the parameter estimates, we can see the isolation of the first two factors labeled EM and MS, whereas the third is labeled Gc. Another way to achieve a similar goal was to drop the VO from the data set completely and refit the two-factor model. When this was done, the model fit was excellent ($\chi^2 = 76$, $df = 8$, $w4 = 1.50$, $\varepsilon_a = .022$). Nevertheless, we know that model fitting itself does not seem to be a good way to isolate the Gc factor—a far better way would be to add variables that are indicative of the broader Gc concept (i.e., knowledge tests) and then to test this isolation with these multiple outcomes.

### An Exploratory Factor Analysis

To see what happens when an exploratory approach is taken, the same matrices were input into an EFA algorithm, where a succession of common factors are extracted, and where multiple factors were defined by factor rotation procedures (Browne, 2001). The EFA results presented in Table 12.4 include the misfit indices, including the error of approximation and its confidence interval (see Browne & Cudeck, 1993). The results listed here clearly show the progression from zero to three common factors improves the fit at every step—one factor is far better than zero; two factors seem far better than one; and three factors seem even better than two. The index of misfit that first achieves one of the standard criteria of "good fit" (where $\varepsilon_a < .05$) is the two-factor model.

The two-factor model fitted as an EFA does not explicitly state where the salient loadings are located. To understand this model, we need to apply some techniques of factor rotation (see Browne, 2001). Of course, this is not a standard solution, so we may not be interested in "simple structure"–based rotations. One useful possibility here was defined by Yates (1987) in terms of minimizing the geometric mean of the squared loadings—the so-called Geomin criterion. Additional research on this Geomin criterion

**TABLE 12.4**

Results for a Consecutive Sequence of Four Exactly Identified Factor Models

| Statistic | $k = 0$ | $k = 1$ | $k = 2$ | $k = 3$ |
|---|---|---|---|---|
| $\chi^2$ | 18,520 | 2,530 | 188 | 24 |
| *df* | 21 | 14 | 8 | 3 |
| $\Delta\chi^2$ | — | 15,990 | 2,162 | 414 |
| $\Delta df$ | — | 7 | 6 | 5 |
| $\varepsilon_a$ | .225 | .102 | .036 | .020 |
| $-95\%(\varepsilon_a)$ | .223 | .098 | .032 | .013 |
| $+95\%(\varepsilon_a)$ | .228 | .105 | .041 | .028 |
| *w4* | 1.61 | 1.58 | 1.37 | 1.02 |

has added standard errors for the rotated loadings (Jennrich, 2007). When we carry out these calculations we obtain the results listed in Table 12.5—the first factor is indicated by the IR and DR and can be termed EM, and the second factor is indicated by the last five variables and can be labeled MS. In other words, the EFA gave nearly identical results to our previous CFA of the model of Figure 12.3b. However, this two-factor CFA was not the only possible EFA rotation of the two factors, and this EFA result shows a remarkable consistency with the CFA model of Figure 12.3b. In this way, this EFA approach gives more credibility to the CFA model of Figure 12.3b.

## Beyond the Initial Structural Factor Model

One of the main reasons we want to isolate a reasonable common factor structure is that we can use this model in further forms of data analyses, such as in models of external validity (McArdle & Prescott, 1992). Two examples from our work on HRS cognition measures are presented here.

**TABLE 12.5**

Results for the Two Common Factor Model

| **Measure** | **Factor $\lambda_1$** | **Factor $\lambda_2$** | **Unique $\psi^2$** |
|---|---|---|---|
| IR[1] | **.83** | .07 | .24 |
| DR[1] | **.90** | **–.01** | .21 |
| S7[1] | .10 | **.60** | .63 |
| BC[1] | –.06 | **.39** | .87 |
| NA[1] | –.06 | **.69** | .57 |
| DA[1] | .05 | **.40** | .81 |
| VO[1] | –.02 | **.60** | .63 |

Maximum likelihood estimation (MLE) with Geomin; $\rho = .65$, $\varepsilon_a = .036$; parameters with MLE/SE = $t > 4$ are listed in bold.

In the *latent variable path analysis* (LVP) approach, we can bring additional variables into the same SEM (McArdle & Prescott, 1992). One benefit of this approach is that we can evaluate the regression model with variables that are purified from measurement error. For example, the LVP of Figure 12.4 (for a full description, see McArdle et al., 2007) shows the three-factor CFA, where the three latent variables of EM, MS, and VO are predicted from only six demographic variables (age, education, gender, cohort, dyad status, and mode of testing). For example, the results show strong negative effects of age on EM (–0.7), and this is a larger effect than the impact of age on any observed variable. The independent impacts of education are positive on all factors (+0.5, +0.6, +0.5). The effects of gender are seen only on EM (females greater by +0.5). The independent effects of cohorts are negative on EM and VO, even though the scores are increasing over successive cohorts (i.e., possibly education effects are responsible). Being in a dyad is somewhat positive, and the mode of testing (telephone or face to face) makes only a little difference in latent test scores.

Another kind of SEM analysis that is now possible is based on longitudinal SEM (see McArdle, 2007, 2009). The longitudinal nature of the HRS data collection is very practical—at the initial testing all persons are measured in a face-to-face setting, but at the second testing about 2 years later, the same people are interviewed over the telephone. Presumably, because the same cognitive questions are asked, the tests used measure the same constructs. Figure 12.5 is a display of this concept about measurement of the same latent variables over time. It is now fairly well known that the

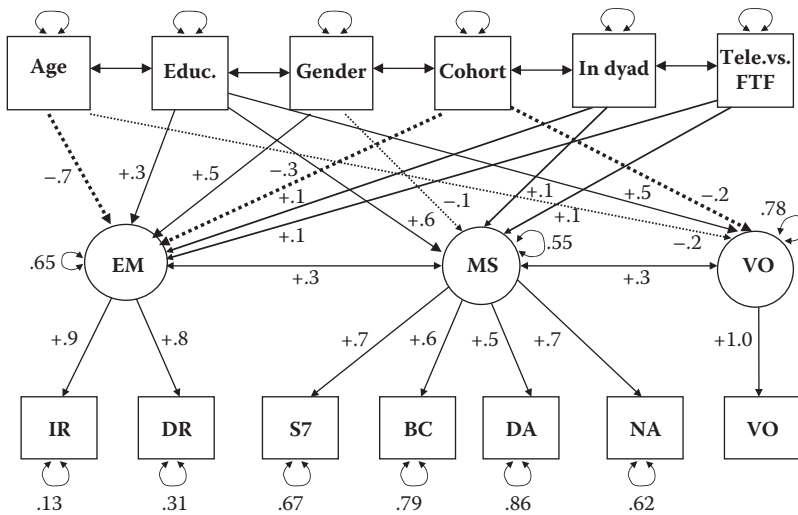



**FIGURE 12.4**
Three common factors related to other HRS demographic indices ($N > 17{,}000$).

general idea of measurement invariance is a testable SEM hypothesis—we force the factor loadings to be identical (or invariant) at both occasions so we can evaluate the loss of fit. If such a model with this kind of "metric invariance" can be said to fit, then we can easily examine other features about the latent variables—means, deviations, cross-regressions, etc. In fact, the need for some form of measurement invariance is so compelling it is hard not to make this the object of the analysis—that is, why not simply use these SEM techniques to isolate the measured variables that seem to have this useful LV property (McArdle, 2007, 2009). This approach, of course, uses CFA software to carry out an EFA analysis (also see Albert et al., 2007; Bowles et al., 2005; McArdle & Cattell, 1994).

To pursue these longitudinal analyses, a new data set based on cognitive variables was constructed from the available archives of the HRS consisting of the first face-to-face (FTF) and first telephone (TEL) testing. To retain the large and representative sample size ($N > 17{,}000$), the VO variable was no longer considered (i.e., it was not measured twice in most cases). The analytic results for the remaining ($v = 6$) variables are presented in Table 12.6. The first three rows (6a) are a list of the model fits for the one-factor model, first as metrically invariant over time, then as configurally invariant (i.e., same nonzero loadings, but not exact values), and then with one-to-one specific longitudinal covariances (as in McArdle & Nesselroade, 1994). The first model does not fit well; the second fits better; and the third is best so far.

The second set of rows presents the fit of the same three models using a two-factor CFA (much like Figure 12.2b), and the fits are uniformly better. The first model is much better; the second is not much different; and the third model, with metric invariance and longitudinal specific factor covariances, is nearly perfect ($\chi^2 = 423$, $df = 57$, $\varepsilon_a = .023$). Thus,

**TABLE 12.6**

Fit Indices for One and Two Common Factors Based on Six Measures at Two Longitudinal Occasions

| 6a: $k = 1$ Models | $\chi^2$ | $df$ | $\Delta\chi^2/\Delta df$ | $\varepsilon_a$ |
|---|---|---|---|---|
| Invariant $\Lambda$, $\Psi^2$ | 8,600 | 69 | — | .087 |
| Configural $\Lambda$ | 8,579 | 64 | 21/5 | .090 |
| MI + specifics covariance | 4,534 | 63 | 4,056/6 | .066 |
| **6b: $k = 2$SS Models** | **$\chi^2$** | ***df*** | **$\Delta\chi^2/\Delta df$** | **$\varepsilon_a$** |
| Invariant $\Lambda$, $\Psi^2$ | 2,579 | 63 | — | .050 |
| Configural $\Lambda$ | 2,578 | 59 | 1/4 | .051 |
| MI + specifics covariance | 423 | 57 | 2,156/6 | .023 |

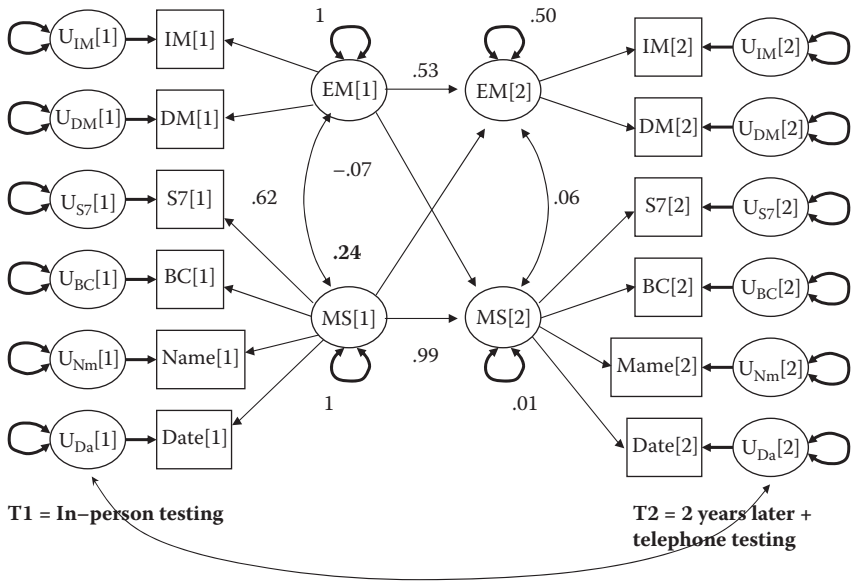



**FIGURE 12.5**
The HRS cognitive measures with factorial invariance over time and mode of testing ($N > 17{,}000$).

whereas we were unsure about one factor, these results suggest the two factors, EM and MS, can be measured using the same six tests in either FTF or TEL modalities without any measurement biases. The results for the latent variable cross-lagged regressions are given in Figure 12.5, and these suggest that the MS[t] is highly stable and most predictive of EM[t + 1]. More analytic work is now being done on these dynamic relationships.

## Prior Work

Let us return to the ethical issues in FA. Ethical issues about the practices in FA have been raised by many others, and the same messages are found in the history of other statistical procedures, such as ANOVA and item response theory. For example, clear recognition of these issues can be found in the classic debates about the "significance of the significance test" (e.g., Harlow, Mulaik, & Steiger, 1997; Lecoutre, Lecoutre, & Poitevineau, 2001). In one compelling resolution, Cattell (1966) rejected the use of the principles of experiment-wise error and suggested the use of what he termed the

*inductive-hypothetico-deductive spiral* (also see Tatsuoka & Tiedeman, 1954). Basically Cattell, among others, was suggesting we consider a continuum with CFA at one end point and EFA at the other (see Table 12.1).

In using CFA, we assume there is more theory, more restrictions (hence more *df*s), overidentified parameter estimates, and ample statistical tests with corresponding good fits. For these reasons, the topographic presentations of CFA seem very strong and useful in research where there has been a lot of reliable work. On the other hand, the EFA end of the continuum is based on less theory, fewer restrictions (lower *df*s), exactly identified parameter estimates, and fewer statistical tests with less good fit. Thus, the EFA seems weak compared with the CFA, can take advantage of chance occurrences in data, and possibly can be misleading. But perhaps the most important aspect of this continuum is that there is a lot of room for many types of FA between the CFA and EFA extremes. There are many FA approaches that are not extremely simple but are not extremely complex either. There are FA models that have some overidentified parameters but also some exactly identified parameters (most, in fact; see McArdle, 1991; McArdle & Cattell, 1994).

Given the favorable advances of CFA, it was somewhat instructive that the more experienced researcher among us (JLH) wanted to first look at the EFA to form a reasonable hypothesis about the data. This was partly indicative of the meta-theory that a specific factor structure fits the data no matter what models are tried (see Horn, 1972). There was also no intention to obscure the fact that statistical tests were not part of the original psychometric history of FA, and there was some resistance in using statistical tests at all (see Kaiser, 1976). These are due partly to what to many seem like absurd assumptions that need to be made for the resulting probabilities to be accurate (i.e., normality of uniquenesses, etc.). But partly this preference for EFA must also be due to years of training on EFA without the new flexibility of CFA.

For reasons defined by the sequence of Figures 12.1, 12.2, and 12.3, the explicit contrast between CFA and EFA is never really clear. In contrast to the newly developed approaches of CFA, the traditions of EFA are much older and were developed at a time when it was difficult to pose a rigorous pattern on factor loadings (Figure 12.2c), even if one was actually known. This is an obvious and clear benefit of CFA. In the past, the EFA was carried out in a sequence to (a) search for the most reasonable number of common factors, and (b) assuming more than one common factor, rotate the factor loadings to a position that seems most interpretable. The first step can use a generic LRT based on a limited number of degrees of freedom, but the second step usually relies on more substantive information—often when we use factor rotation we say we are trying to find a set of loadings that are "most reasonable" for these variables. For these reasons, many scientists now seem to find factor rotation as more artwork than science.

On the other hand, some researchers tend to think the one-factor model is identical in the CFA and EFA framework, and this ignores several key model possibilities. In a CFA, we have control over all the parameters; thus, as a prime example, we can fix the factor loadings at some known a priori values from another study. Indeed, fixed loadings would be an excellent example of a true confirmatory analysis, but a fixed loading is hardly ever part of any contemporary CFA application. The previous description of the Rasch model makes it seem like an ultrastrong CFA, but this also is a naive way to think the Rasch model is typically used. Instead of a strong CFA approach, a good fit of the Rasch model is simply the required goal of the analysis. That is, because a one-factor model with equal loadings is needed for the purposes of further measurement, this strategy implies that items should be eliminated until this goal is reached, and any statistical tests are merely an indicator of when to stop eliminating variables (Embretson & Reise, 2000). Obviously, any difference between CFA and EFA is muddled again.

Thus, the ethical problems with this newer CFA approach are at least twofold. First, as stated above, the use of the term *confirmatory* is a bit odd when we use this only to refer to the pattern hypothesis and we do not place an a priori value on the parameters. People reading about CFA for the first time may view this as a truly confirmatory procedure when, in fact, confirmation is used in a limited way. Second, the test of this CFA model is only exact when we specify the exact pattern in advance of the data—a priori. Unfortunately, the probabilistic basis of the LRT does not normally hold when there are attempts at a "refinement" or a "trimming" of the model using standard data analysis procedures. That is, it is hard to defend an approach where we simply drop variables and/or add arbitrary parameters until our model fits and then claim we can use the chi-square distribution to defend this model fit (cf. Brown, 2006). When we do not have an a priori hypothesis, we do not know whether the resulting probability is an index of any a priori sampling distribution.

A true CFA requires lots of effort at good measurement design and is not typical at all in the current SEM literature. It follows that a true CFA is rarely the case, and we much more typically need to make serious reorganizations and refinements of the model loadings using the data at hand. This standard "model fitting" approach to FA seems to make all CFAs move toward the EFAs, and there is nothing wrong with this. The main ethical problems emerge when we try to hide behind the CFA approach when in fact we are closer to doing EFA. If we do this, in essence, we are *lying with statistics* so we can tell a good story and get our work published. If this minor deception works once, it will probably work again and again; others will follow our lead, and inappropriate practices will become simply the way we do business.

## Conclusion

To its great credit, the American Psychological Association (APA) is a leader in the recognition of ethics problems. Consider the book-length treatments of the *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002) and the earlier book-length commentary of Canter, Bennett, Jones, & Nagy (1994). It is hard to find another group more interested and active in ethical practices than the APA. Unfortunately, when it comes to data analysis, arguably the only component common to all areas of behavioral science, APA guidelines seems to demand little. The APA guidelines present the outdated inclusion of extremely odd practices in presenting probability (multiple asterisks for different $p$ levels), but these guidelines focus more on making tables for APA publications. The sensible suggestions of Wilkinson and the Task Force on Statistical Inference (1999) need to be taken more seriously. But, in reality, we must take the lead on this ourselves and express rules of good behavior using statistics. This chapter concludes with five suggested rules that are designed to lead to good practice in factor analyses.

1. *When reporting results, be honest*. The first principle of ethical FA is that we do not need PURITY of statistical rules and assumptions, but we do need HONESTY. Try to tell us exactly (as briefly as possible) how you selected the people, variables, and occasions, even if it is complicated. Consider missing data, outliers, and transformations, but please report their impacts on the results. Try to tell us exactly how you found the models used, especially if they were not a priori and if they emerged as part of the analysis. Tell us ALL relevant results, not just the BEST ones.
2. *The FA goal is replication*. Clarity is essential in any FA, and the key criterion in any experiment or analysis is *replication* (Lykken, 1968). Remember that confusion can be created by brevity, so we should not simply blame the reviewers. Reviewers want to make sure the work is fact, not fiction. What you are doing might not be clear enough to be replicated, and in this case you must clarify it. If the reviewers suggest you have broken the rules of "purity" (i.e., overall experimentwise error rate $\alpha > .05$), then you need to fight against this illogic directly and with vigor. Possibly you will need to change your favorite journal or funding agency, but at least you will be doing the right thing.
3. *Change the FA terminology*. The statistical terminology is often initially defined for one situation but found to be useful in another. Therefore, we should not simply use the same classical words when they mean something entirely different. For example, we

should immediately change theory or hypothesis → idea; test → examine; prove → demonstrate; data revealed → we noticed; significance → accuracy; predicted → connected; and controlled → adjusted. In the SFA context, we should substitute correlated errors → correlated specifics; confirmatory → overidentified, not rotatable; exploratory → exactly identified, rotatable; and a factor in FA is a thing → a factor in FA is evidence for the existence of a thing (Cattell, 1978). And if we do not know the basis of the probability statements we wish to make, we should drop them from our language and our analyses entirely.

4. *Primary analyses should use existing data*. There are very few barriers to the analysis of existing data, and this will allow most anyone to learn how to carry out analyses and demonstrate that we know how to analyze complex data problems. The analysis of existing data should be a formal requirement before anyone collects any new data on any individual. Of course, the APA publication system and the National Institutes of Health and National Science Foundation federal grant systems need to be willing to recognize this as valid research, too. One helpful hint: We can almost always think of the question in advance of the data selection—"*We cannot analyze a database, but we can analyze a question using a database!*"
5. *Any study should confirm, THEN explore*. In phase 1, confirm. Try to come into an analysis with a plan about the set of ideas you are going to examine and the data you are going to use to do so. This will permit a full and appropriate use of statistical probability tests and other indices of fit. Remember that we do not want the "best" model; we want the "set" of models that fit well separated from those that are "average" and "poor." In a subsequent phase 2, explore. Whether or not your favorite model fits the data on hand, try to improve the fit using any aspect of the data on hand. Do this completely so you can find something better than you had in phase 1. Who knows, maybe new results will then be able to be replicated by others.

In the merger of CFA and EFA into SFA, we are in the awkward position of trying to compromise two different statistical traditions: one old and one new. As long as nothing is lost, the newer techniques (CFA) offer improvements and should be favored over the older techniques (EFA). A key point here is that a lot can be lost in the blind application of CFA in situations where EFA might tell us a lot more, or at least the same thing (the HRS example here). In the classical and rigid approach to confirmation via hypothesis testing, we are taught to disdain the use of separate *t* tests in favor of the more rigorous one-way ANOVA (Scheffe, 1959). In an

exploratory mode, we are asked to wonder whether we missed anything important in the data we have collected (Tukey, 1962, 1977). Obviously, these are all valid points in an extended conversation about data analysis. However, we can all agree that it is wise to know exactly what we are doing ourselves, what boundary guidelines we need to follow, and to make sure we follow these rules ourselves.

Ethical guidelines in the area of FA can be as clear as in any other area of science. The main requirement is to report the sequence of analyses carried out so the reader can repeat or improve on these steps. Odd behaviors can emerge when the scientists forget to report a crucial step in the procedure, but this becomes an ethical problem when we do so on purpose or when we use a statistical test with known violations. This is as much an ethical violation as omitting a relevant reference because we do not like the author (i.e., we do not want to add to his or her h-index!). Unfortunately, there is little way to know when this is going on in these cases, so we must rely on the ethical behavior of the individual scientist. Of course, anyone who observes these behaviors—our students, our colleagues, our children—know what we are doing, and this alone may provide some needed ethical corrections.

*In the SFA approach advocated here, we start with a strict CFA and move toward a more relaxed EFA—this is exactly what we typically need to do, and there is nothing unethical about it!* This approach turns unethical when the sequence of procedures we use is not reported, perhaps in the hope that we can retain the illusion of the precision and power of the newest CFA-based statistical tests. As I have tried to show here, pretending to use CFA when we are really doing a form of EFA is fool-hardy at best—and devious at worst.[1]

## References

Abelson, R. (1995). *Statistics as principled argument*. Mahwah, NJ: Erlbaum.

Albert M., Blacker D., Moss M. B., Tanzi R., & McArdle J. J. (2007). Longitudinal change in cognitive performance among individuals with mild cognitive impairment. *Neuropsychology, 21,* 158–169.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct* (5th ed.). Washington, DC: APA Press.

[1] Author note: Thanks to Drs. A. T. Panter and Sonya K. Sterba for creating this opportunity, to Drs. Daniel and Lynda King for their insightful comments, and to Dr. John L. Horn for his classic advice in dealing with these complex technical and ethical issues: "People often underestimate the dangers of overplanning" (08/24/1990). The work reported here was initially presented at the APA symposium of the same title, Boston, August 2008. This work has been supported by National Institutes of Health Grant AG-007137.

Bowles, R. P., Grimm, K. J., & McArdle, J. J. (2005). A structural factor analysis of vocabulary knowledge and relations to age. *Gerontology: Psychological Sciences, 60B*, 234–241.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150.

Canter, M. B., Bennett, B. E., Jones, S. E., & Nagy, T. F. (1994). *Ethics for psychologists: A commentary on the APA ethics code.* Washington, DC: American Psychological Association.

Cattell, R. B. (1966). Psychological theory and scientific method. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 1–18). Chicago: Rand McNally & Co.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fisher, R. A. (1925). *Statistical methods for research workers*. (14th ed., 1973). New York: Hafner.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Horn, J. L. (1972). State, trait, and change dimensions of intelligence. *The British Journal of Mathematical and Statistical Psychology, 42,* 159–185.

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical and statistical model building (MASMOB) in research on aging. In L. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association.

Horn, J. L., & McArdle, J. J. (1992). A practical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100 years* (pp. 205–247). Mahwah, NJ: Erlbaum.

Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In R. C. MacCallum & R. Cudeck (Eds.), *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31,* 165.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183.

Jöreskog, K. G. (1977). Factor analysis by least-squares and maximum-likelihood methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 125–153). New York: Wiley.

Juster, F. T., & Suzman, R. (1995). *The Health and Retirement Study: An overview.* HRS Working Papers Series 94-1001. *Journal of Human Resources*, *30*, S7–S56.

Kaiser, H. (1976). [Review of the book *Factor analysis as a statistical method*]. *Educational and Psychological Measurement, 36,* 586–589.

Lawley, D. N., & Maxwell, A.E. (1971). *Factor analysis as a statistical method.* New York: Macmillan.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical, 69,* 399–417.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

McArdle, J. J. (1991). Principles versus principals of structural factor analysis. *Multivariate Behavioral Research, 25,* 81–87.

McArdle, J. J. (1994a). Factor analysis. In R. J. Sternberg (Ed.), *The encyclopedia of intelligence* (pp. 422–430). New York: Macmillian.

McArdle, J. J. (1994b). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science, 5,* 11–18.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. MacCallum & R. Cudeck (Eds.), *Factor analysis at 100 years* (pp. 99–130). Mahwah, NJ: Erlbaum.

McArdle, J. J. (2009). Latent variable modeling of longitudinal data. *Annual Review of Psychology, 60,* 577–605.

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research*, *29*(1), 63–113.

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analysis of age trends in tests of cognitive ability in the elderly U.S. population, 1993–2004. *Psychology and Aging, 22,* 525–545.

McArdle, J. J., & Lehman, R.S. (1992). A functionalist view of factor analysis. In D. F. Owens & M. Wagner (Eds.), *Progress in modern psychology: The contributions of functionalism to modern psychology* (pp. 167–187). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: methodological innovations* (pp. 223–267). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Prescott, C. A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18,* 87–115.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meredith, W., & Horn, J. L. (2001). The role of factorial invariance in measuring growth and change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 201–240). Washington, DC: American Psychological Association.

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). New York: Chapman & Hall.

Scheffe, H. (1959). *The analysis of variance*. New York: Wiley.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (Eds.). (1999). *Cognition, aging, and self- reports*. Philadelphia: Psychology Press.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 475–502.

Tatsuoka, M. M., & Tiedeman, D. V. (1954). Discriminant analysis. *Review of Educational Research*, Washington, DC: AERA Press.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Tukey, J. W. (1962) The future of data analysis. *Annals of Mathematical Statistics, 33,* 1–67.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis.* Albany, NY: State University of New York Press.

# 13

# *Ethical Aspects of Multilevel Modeling*

**Harvey Goldstein**
*University of Bristol*

All professional ethical codes stress the importance of honesty and personal integrity, the striving for objectivity, and the avoidance of any attempt to mislead by virtue of professional status. Professional associations for those working in quantitative disciplines, such as the American Psychological Association (APA), the Royal Statistical Society (RSS), the International Statistical Institute (ISI), and the American Statistical Association (ASA), additionally stress the ethical imperative to make use of appropriate and generally accepted technical standards when collecting and analyzing data (APA, 2002; ASA, 1999; ISI, 1985; RSS, 1993). The ASA, for example, makes some specific technical points, such as the recognition that any frequentist statistical test has a nonzero probability of producing a "significant" result when the effect being tested is not present and warning against deliberately selecting the one "significant" result from a large number of tests. The ASA (1999) is also clear that statisticians and those carrying out statistical analyses should "remain current in terms of statistical methodology: yesterday's preferred methods may be barely acceptable today."

I shall lean heavily on this notion that advances in knowledge can not only make previous technologies or methodology less efficient, but also that new knowledge can expose the hitherto hidden distortions and biases inherent in such previous technologies. Thus, new knowledge can make unethical what may previously have been considered acceptable procedure. This impact of knowledge is clear in areas such as medicine, where, for example, the practice of patient bleeding may have been mainstream orthodoxy in the 18th century but would be considered highly unethical if used instead of treatments known to be scientifically effective in the 21st century. It is perhaps less obvious in areas such as statistical analysis, but nevertheless, in principle the same kind of arguments can be made.

The present chapter will look at one particular set of evolving methodologies, those generally termed *multilevel models*. The following sections

will seek to explain what these are and how they extend existing methodologies, how they can produce novel inferences, and how they can extend the range of questions that may be addressed. In particular, I would argue that this general methodology has now reached a stage of maturity, as witnessed by its routine use and its incorporation into major statistical packages, which implies there is an ethical obligation to use it where appropriate. In other words, this methodology is indeed one that has made a large number of yesterday's preferred methods "barely acceptable."

A key point that will be exemplified in the examples I use is that multilevel models are the appropriate tools for addressing certain kinds of research questions. In addition, the existence of such tools allows us to ask certain kinds of research questions that were either difficult or even impossible to address previously, and I will give examples.

## A Brief Introduction to Multilevel Models

Interesting real-life data rarely conform to traditional textbook assumptions about data structures. These assumptions are about observations that can be modeled with independently and identically distributed "error" terms. More often than not, however, the populations that generate data samples have complex structures where measurements on data units are not mutually independent, but rather depend on each other through complex structural relationships. For example, a household survey of voting preferences will typically show variation among households and voting constituencies (constituencies and households differ on average in their political preferences). This implies that the responses from individual respondents within a household or constituency will be more alike than responses from individuals in the population at large. Another example of such "hierarchically structured data" would be measurements on students in different schools (level 2 units), where, for example, schools differ in terms of the average attainments of their students (level 1 units). In epidemiology, we would expect to find differences in such things as fertility and disease rates across geographical and administrative areas. Designed experiments with repeated measures on individual subjects generate a two-level structure where measurement occasions are nested within individuals and where multilevel models provide an appropriate analysis framework. A good introduction to multilevel models is Hox (2002), and a more advanced text is Goldstein (2003).

To formalize the idea of a multilevel model, consider the simple case of a regression model where we assume normality:

$$y_{ij} = \alpha + \beta x_{ij} + u_j + e_{ij}, \quad e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2), \tag{13.1}$$

applied to a sample, say, of school students where $i$ indexes students, $j$ schools, the response $y$ is an attainment measure, and $x$ is a predictor such as a previous test score. I shall use this educational example to discuss the technique because the application will be familiar to a large number of readers. The assumption of identically independently distributed residuals is no longer tenable because the (random) residual term is $u_j + e_{ij}$ rather than $e_{ij}$ as in a traditional linear regression model. Two students from school $j$ now share a common value $u_j$ so that their attainments, in this case adjusted for the predictor $x$, will be more alike than two students chosen at random from different schools. Put another way, the assumed linear relationship between $y$, $x$ has a different intercept term ($\beta_0 + u_j$) for each school. In the school effectiveness literature, this term would be interpreted as each school providing a separate effect on attainment after adjusting for previous achievement—a simple example of the so-called value-added model.

Eq. 13.1 is often known as a "random intercept" or "variance components" model and also sometimes a "mixed model" or "hierarchical linear model." Note that we have chosen to model the school effect as a random variable depending on a single parameter, the variance $\sigma_u^2$, and an alternative would be to fit school as a "fixed effect," using, for example, a set of $m - 1$ dummy variables where $m$ is the number of schools. In some special circumstances this approach may be preferred, but more usually we would wish to consider the set of schools (or geographical areas or households) as a randomly chosen sample from a population of schools about which we wish to make inferences. This is the key issue that distinguishes these models from traditional ones.

If we have hierarchically structured data, and there are few real-life situations where we do not, and if we ignore the structure when modeling, then two consequences follow. First, our inferences will be incorrect. Standard errors will tend to be too small, significance tests too optimistic, and confidence intervals too short. The size of such biases will depend on the strength of the structure, but in general there is little (ethical) justification for ignoring it.

The second issue is that if we do not model the structure, then we can say nothing about it. If one aim of an analysis is to report on the variation between school performance, or the difference in mortality rates between intensive care units, then such units must be included explicitly in our models. Furthermore, they need to be included within a proper multilevel model, rather than, for example, a model that operates, say, just in terms

of school means. For example, we could compute school mean test scores and carry out an analysis where the school was the unit of analysis and the mean score was regressed on other school-level characteristics. If we did this, we are likely to commit the so-called "ecological fallacy." This fallacy has been known since at least the 1950s (Robinson, 1951) and occurs when an analysis is carried out at one level of a data hierarchy (e.g., based on school means), whereas we require inferences about relationships at a different level (e.g., on students). One of the considerations in any analysis is to determine the extent of the variation among higher-level units. In the extreme case where this is very small, we may be able to ignore the multilevel structure and use a single-level model.

The basic model (Equation 13.1) can be extended in a number of ways. The coefficient β can be given a subscript *j* so that the "slope" may vary across school as well as the intercept, further predictors can be added, and generalized linear models such as those for binary responses with a logistic link function can be formulated. Furthermore, we can further structure any of the variances (e.g., between students) as functions of further variables, and this allows for the detailed study of variation in a way that traditional models have been unable to cope with (see Goldstein & Noden, 2003, for an application to school social segregation). Most of these models are in widespread use and available in the major general-purpose statistical packages and in certain more specialized ones (Goldstein, 2003).

In practice, however, data structures are often more complicated than the kind of simple hierarchical structures modeled in Equation 13.1. Consider an educational example where students are followed through both their primary and secondary education with the response being attainment at the end of secondary school. For any given primary school, students will generally move to different secondary schools, and any given secondary school will draw students from a number of primary schools. Therefore, we have a *cross-classification* of primary by secondary schools where each cell of the classification will be populated by students (some may be empty). When we model such a structure, we have a contribution to the response that is the sum of an effect from the primary and an effect from the secondary school attended by a student. If such a cross-classification exists and we ignore it, for example, by fitting a purely nested model using secondary school, we may bias estimates. Goldstein and Sammons (1997), for example, show that in this case adding the primary school as a cross-classification substantially changes the size and interpretation of the secondary school variation.

Pursuing this example further, we know that students do not all remain in the same secondary or primary school. Thus, a student may attend two or three primary schools so that the "effect" of primary school on the response is the average effect of all the primary schools

attended. These models are referred to as *multiple membership models* because a lower-level unit can belong to more than one higher-level unit. Such models are also useful for studying spatial correlation structures where individuals can be viewed as belonging simultaneously to several areas with appropriate weights. Although not always found in all the general purpose packages, such models have found their way into publications in many disciplines (e.g. geography and epidemiology). As with cross-classifications, ignoring a multiple membership structure can lead to biased estimates. Thus, Goldstein, Burgess, and McConnell (2007) demonstrate the use of these models in the study of student mobility across schools and show that fitting a purely hierarchical model leads to underestimation of the between-school variation.

In the next section we look at some of the ethical considerations that should be addressed when designing research involving different types of nested data structures.

## Designing Studies in the Presence of Hierarchical Structures

If we are interested in making inferences about units such as schools, hospitals, or electoral constituencies that belong to a hierarchical structure, then our study design will generally involve a sampling of such units, and we will wish to ensure that we have sufficient precision for the appropriate estimates. Power calculations are now recognized in many areas of application, especially medicine (see Fidler, Chapter 17, this volume), as essential components of good design. Current practice in terms of hierarchical data structures and multilevel modeling typically does not take note of such calculations, in part because the relevant software is scarce. This situation is changing, however, and some flexible, open-source software is becoming available (see, e.g., Browne & Lahi, 2009). One ethical aspect is that if a study is underpowered, for example, too few intensive care units are sampled, we will tend to obtain "nonsignificant" results, and this may either result in the study being ignored or, more importantly, presented as good evidence for a lack of relationship or lack of variation between units. This situation is, of course, the multilevel counterpart of a lack of power in traditional single-level analyses (see Maxwell & Kelley, Chapter 6, this volume) but less often may be recognized as such.

An important example of this in educational research was the early school effectiveness study in inner London schools, *Fifteen Thousand Hours* (Rutter, Maughan, Mortimore, Ouston, & Smith, 1980). This study obtained

information from 2,000 children in 12 secondary schools. The study made comparisons between school types, for example, boys' and girls' schools, found nonsignificant differences, and concluded that such differences are of "negligible importance" (Goldstein, 1980). Yet, with a sample size of only 12 schools, it is hardly surprising that almost all comparisons will be nonsignificant.[1] The authors failed to appreciate this design problem and also made the common error of equating "nonsignificance" with "nonexistence." Although this issue, often referred to as the "units of analysis problem," was fairly well understood at that time and had been discussed in the methodological literature, it might be argued that this should be regarded as merely incompetent rather than unethical behavior. Yet, in their response to this point (Rutter, Maughan, Mortimore, Ouston, & Smith, 1980), the authors refused to accept the strictures. Because that study turned out to be influential and was, in fact, heavily promoted by the publisher of the report as well as by the authors, a refusal to concede that there may have been a serious flaw could be considered by many to constitute a case where ethical norms were breached. This would not be in terms of deliberately providing a misleading description but rather in terms of a failure to ensure that, as researchers, they were prepared properly to acknowledge current good professional practice.

All of this was unfortunate because the lessons for study design were obscured, and the importance of sampling adequate numbers of higher-level units was not made clear to many researchers in this field.

## The Importance of Clustered Designs for Substantive Issues

The example Rutter et al.'s (1980) *Fifteen Thousand Hours* study was intended to demonstrate the importance of sampling to obtain adequate precision for statistical inferences. There are, however, other reasons for taking note of a hierarchical structure.

Sample survey analysis has long recognized the usefulness of clustering respondents to achieve low cost. The "clusters" themselves are sampled with known probability. Because the clusters will often vary randomly in terms of the response variables of interest, this "level 2" variation needs to be accounted for in the analysis. Traditionally, this has been done by fitting single-level models, for example, regressions, and then working

[1] A later analysis of a similar population, but fitting a multilevel model to a large sample of schools, showed clear differences between boys', girls', and mixed schools (Goldstein et al., 1993).

out the "corrected" standard errors for the parameters, for example, using jackknife estimators (Efron & Gong, 1983). By contrast, if we carry out a full multilevel model that recognizes the clustering in the model itself, for example, that each area is a level 2 unit, then this approach achieves the same end with greater statistical efficiency. Moreover, it also directly allows us to include analysis of variables measured at the cluster level. Such variables could be aspects of, for example, neighborhood, and our focus of interest might be how much of the between-area variance such variables could explain.

Therefore, the multilevel approach helps to shift the focus from the clustering simply being a convenient procedure to obtain a sample to a positive attempt to bring in ecological variables that are defined at the cluster level. Thus, in a recent study for the design of a large-scale birth cohort study in the United Kingdom, the think tank Longview (2008) argued for a sample that consists of a nationally representative component together with a small number of tightly clustered samples in local areas or clustered around local institutions. The area samples would include all the births over a period of, say, 1 year so that the characteristics of each child's peer group could be measured, for example, when they attend preschool facilities. The sample would obtain nationally representative data, and the existence of a common set of variables across the sample would allow the various subsamples to be linked. This linking can be done formally within the modeling framework, "borrowing strength" across the subsamples. In other contexts such designs are often known as matrix designs or rotation designs, and they have many advantages in terms of efficiency and being able to combine local and national data (see, e.g., Goldstein, 2003, Chapter 6). In social research this is important because it begins to address potential criticisms of large-scale empirical research on populations: they ignore contextually relevant factors. The ability to combine large representative sample data with more intensive local data that is sensitive to local issues also begins to provide a way of drawing together large-scale data sets and small-scale studies such as those that collect detailed ethnographic data. Thus, the design possibilities for such studies become extended, and this knowledge, as it becomes widely accepted, will exert an ethical pressure to consider these possibilities.

In education there is some considerable interest in peer group effects, often known as *compositional effects*, in which aggregated achievements and behaviors of peers are modeled as influences on a student's own performance or behavior. For example, we might conjecture that the average socioeconomic status or an average of previous test scores for the other students in a school or classroom is related to a student's performance, over and above that student's own characteristics. To do this satisfactorily, however, requires data on the student's peer group, and

this generally implies obtaining data from complete, or near-complete, year groups or institutions. To collect such data is often difficult and unrealistic, not least because over time individuals move and clusters become diluted. The existence of large-scale comprehensive databases, often collected for administrative purposes, has recently allowed some advances to be made, and a discussion and example are given by Leckie and Goldstein (2009). In that study, the authors used a longitudinal database for all the state school pupils in England (the National Pupil Database) that has records on every pupil, recording some basic demographic data and where they are at school. Because it has data on every student, this database allows compositional effects to be studied efficiently. In some circumstances, data from a sample of peer group students may be adequate, but the analysis will then need to recognize that any aggregated variable derived from such a sample is an *estimate* of the desired compositional variable and should be treated as measured with error. Goldstein (2003, Chapter 13) has a discussion of this issue. I shall return to the National Pupil Database and say more about school effects in a later section.

## The Role of the Data Analyst

I have already discussed some of the issues that arise in the course of data analysis and design involving nested data structures. Here I shall elaborate on the specific ethical responsibilities that fall to the data analyst in helping to design a study and in undertaking principal responsibility for data analysis and interpretation for multilevel data. Ideally, any design should be informed by the kind of analysis that will follow. If a multilevel analysis is envisaged, then there needs to be sufficient power to carry this out efficiently, and data relevant to identifying and characterizing higher-level units have to be collected (i.e., unit and cluster identifiers such as school IDs and student IDs). I assume that in general the data analyst is also involved in design, although that will not always be the case, for example, in secondary data analysis. Nevertheless, it will always be desirable that somebody with experience of data analysis is involved with the initial research design. So for practical purposes we can consider this to be the same person.

The point has already been made that real-life data generally have a complex structure that is hierarchical and may also include cross-classifications, etc. It is ethically responsible for the data analyst to be aware of this and also be concerned to make collaborators sensitive to this issue when a study is being designed so that there is sufficient power

for required comparisons, especially those that involve higher-level units. The data analyst will also have a role in formulating questions based on what he or she knows about the possibilities for data modeling. Thus, for example, the ability of multilevel models to model variation, as in the study of segregation, may not be immediately apparent to many researchers. Structuring a study to separate sources of variation may also be important for efficiency and understanding. Thus, O'Muircheartaigh and Campanelli (1999) cross-classified survey interviewers by survey areas and were able to separate the between-interviewer variance from the between-area variance for various responses. Among other things this analysis allowed the "effects" of different interviewers to be estimated and can inform more efficient survey design.

When it comes to modeling, the data analyst again has an ethical responsibility not only to seek the appropriate tools but also to involve collaborators in understanding how they are being used and how results are to be interpreted. This is especially important because of the relative novelty of multilevel models and the novel perspectives that they can provide. Likewise, the data analyst should be involved in the preparation of papers and reports that present results so that appropriate interpretations are communicated.

In some cases, it may be that data analysts are required to familiarize themselves with new software, especially where there is considerable complexity of modeling. There is some guidance available in this respect; see especially the UCLA MLwiN Portal (http://statcomp.ats.ucla.edu/mlm) and University of Bristol Multilevel Modeling Software Reviews (http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software) websites. These sites also give links to other resources, and the latter website has introductory training materials.

Finally, as in all statistical modeling, the analyst needs to be sensitive to the assumptions that are being made. Techniques for checking distributional assumptions using, for example, outlier analysis are available (see, e.g., Goldstein, 2003, Chapter 3). Sensitivity analyses can also be carried out where assumptions are systematically varied to view the effect on estimates. For example, where assumptions are not tenable, a distribution cannot be assumed to be Gaussian, then, as in traditional modeling, transformations or alternative model formulations may be possible.

## A Case History: School League Tables

This section will draw together a number of the ethical concerns already mentioned through discussing the topic of school performance indicators

or "league tables" where multilevel models have been used—and sometimes abused.

Starting in the 1980s, many educational systems, especially in the United States and the United Kingdom, began to experiment with the publication of examination results and test scores for schools and colleges. Visscher (2001) gives a history of international developments and a review of the debate, principally in Europe, and Dorn (1998) provides a detailed account of developments in the United States. These league tables were designed for two principal purposes. The first was to monitor the performance of individual institutions so that "poorly performing" ones could be identified for further attention. At one extreme, this involved their "formative" use as part of a "school improvement" program where results were not published but used to inform individual schools of their possible strengths and weaknesses (Yang, Goldstein, Rath, & Hill, 1999). At the other extreme, they have been used directly in the determination of school funding and teacher remuneration (Dorn, 1998).

The second main purpose was to provide parents and students with information to guide school choice. In the United Kingdom this was explicitly stated in the so-called "parents charter" issued by the John Major Government (Department for Education and Science, 1991), which encouraged parents to make use of the relative positions of (secondary) schools in tables of examination results. The implication was that those schools with higher average performance were educationally more effective.

These early uses of league tables were strongly criticized, especially by teacher unions and academics, on the grounds that average performance was strongly associated with achievement when students started school, and because schools were generally differentiated in terms of these initial achievements, the final outcomes were in large part simply reflecting intake. It was argued that "value-added" or "adjusted" performance was more appropriate, where account was taken of initial differences. To do this, models were constructed that were essentially two-level, with students nested within schools, and typically some form of multilevel analysis was carried out (see Goldstein & Spiegelhalter, 1996, for a technical discussion). To some extent, policy makers took note of this criticism, so that adjusted league tables were introduced, and in England from 1995, it became official Government policy to move toward a "value-added" system.[2]

By 2003, value-added tables for both primary and secondary stages of education were being published in England alongside the unadjusted

[2] In the United Kingdom, the four constituent countries have separate jurisdiction over education. Thus, by 2010, only England still published school league tables, whereas Scotland had never instituted its publication.

ones. Unfortunately, the media in general, although giving great prominence to the unadjusted or "raw" tables, virtually ignore the value-added ones, and the Government appears to be relatively unconcerned with this, leaving itself open to criticisms of complacency and even hypocrisy. The consequences for individual schools of being ranked low on such tables are fairly clear in any system where parents are encouraged to use such rankings to choose schools. Yet, in all this debate, the provisional nature of statistical modeling has largely been overlooked, and the potential "unfairness" to individual schools has largely been ignored. It is certainly the case that adjusted performance comparisons provide a "fairer" way to compare institutions, but they themselves are only as good as the data used to provide them and suffer from numerous drawbacks, some of which we discuss below. Yet, many proponents of adjusted tables have either ignored or downplayed the limitations of the statistical models. Indeed, FitzGibbons and Tymms (2002), who carried out the pilot work for the English value-added tables, defend their use of "simple" methodology by stating that "The multi-level analysis, requiring special software and a postgraduate course in statistical analysis, was in contrast to the ordinary least squares analysis that could be taught in primary schools" and that "value added scores for departments or schools, correlated at worst 0.93, and more usually higher, up to 0.99 on the two (multilevel vs. ordinary least squares) analyses" (p. 10).

In fact, the high correlations quoted result from the fact that only variance component models were fitted by these authors, so that schools varied solely in terms of their intercept terms. In fact, schools are known to be differentially effective (see, e.g., Yang et al., 1999), their "value-added" scores varying according to the intake achievement, gender, and other student-level factors. To understand the role of such factors, it is essential to fit more complex multilevel models that include both an intercept and slope terms to reflect differential school effects. If this is done, the misleading claims made by the above authors do not stand up to careful examination (Yang et al., 1999). This case is an illustration of the ethical failure to understand the true complexity of the system being studied so that oversimple models are used that do not reflect important aspects of the data. The above quotations also reflect a rather worrying antagonism that some researchers exhibit toward the use of complex models on the grounds that "simple models will do the same job." In fact, as I have attempted to illustrate, simple models often do not "do the same job." This kind of intellectual philistinism toward sophisticated quantitative modeling is often found in educational research and is as ethically reprehensible as it is scientifically blinkered. I am not, of course, advocating model complexity for the sake of it, but I am arguing in favor of modeling at a level of complexity that seeks to match the complexity of the real-life data being analyzed.

In fact, the results of statistical models are nearly always provisional. Value-added scores are subject to the adequacy of the model in terms of the variables included and the sampling variability or uncertainty, typically expressed in confidence intervals. In criticizing "raw" league tables and arguing for value-added ones, researchers have sometimes failed to stress that the latter must also be regarded as imperfect. Thus, for example, Yang et al. (1999) suggest that tables of institutional rankings should not be published, but they can be used within an educational system as "screening instruments" to alert schools and administrators where there could be problems that require further investigation. In other words, they are not definitive judgments that in medicine would typically be referred to as *diagnoses*, but rather indicators of where problems may be occurring. In political climates where education is viewed from a market perspective and performance targets are imposed, such a position is difficult to maintain. Yet researchers, who are aware of the limitations of statistical analyses, do need to maintain an ethical position that requires them to stress those limitations.

Finally, on the issue of the use of (value-added) league tables for choosing schools, Leckie and Goldstein (2009) have pointed out that what parents require is a *prediction* of school performance several years ahead for when their children will take their examinations or graduation tests. The additional uncertainty associated with such predictions greatly increases the uncertainty or confidence intervals associated with rankings, to the extent that few institutions can be separated statistically. In other words, they have little use for the purpose of school choice. For governments to continue to promote this purpose in the light of such evidence is clearly unethical, but that is more a question of political than research ethics.

## Conclusion

In this final section I shall attempt to formulate some general guidelines for analysis, design, interpretation, and reporting involving multilevel data, which were drawn from the above discussion. In particular I will emphasize those relevant to the use of complex models.

Using mathematical or statistical models to describe complex systems has always been a kind of catch-up process. As our methodological tools and data collection facilities become more sophisticated, they can uncover more of the complexity that lies within natural or social systems. Unfortunately, all too often researchers confuse a perceived (and often justifiable) need to present findings in an accessible, as simple as possible,

form to nonexperts, with the need to carry out research using complex techniques that are *only* accessible to experts. The challenge for the experts is not to simplify their techniques but to simplify their explanations of those techniques. In the introduction I suggested that multilevel modeling had reached a stage of development and accessibility that should mandate its routine use for modeling complex hierarchical structures, and the above examples have been presented to show how an understanding of multilevel modeling can improve our understandings and generally advance research.

One implication is that not only researchers but also those who train researchers, largely in universities, should incorporate such modeling techniques as routine. It is interesting that there is little emphasis in existing ethical codes of, for example, the APA, ASA, RSS, and ISI organizations on the role of methodological educators. This is unfortunate because these individuals and the materials they produce will have a large influence on the conduct of research and scholarship.

Another implication is that those carrying out research have a responsibility to remain abreast of developments in both methodology and its applications. I would argue that, given access to the Internet, there are now adequate opportunities for this to happen, using the web resources that have been mentioned. Professional societies also play an important role in providing continuing professional development activities, in the form of materials, workshops, and meetings.

Finally, although this chapter has focused on multilevel modeling, much of what it contains is relevant to other methodologies that involve some elements of complexity. Indeed, there is no simple boundary that exists between methodologies. Often, similar topics are studied under different names, and sometimes a methodology appears that unifies a number of previously disparate ones. Thus, multilevel modeling itself is a provisional activity, continually evolving, linking with other methodologies and perhaps eventually becoming incorporated in more general frameworks. Nevertheless, the issues I have discussed, in new forms, will remain, and it is hoped that we shall continue to worry about the ethics of what we are doing.[3]

## References

American Statistical Association. (1999). *Ethical guidelines for statistical practice*. Retrieved from http://www.amstat.org/about/ethicalguidelines.cfm

[3] Author note: I am very grateful to John Bynner and to the editors of this volume for helpful comments on early drafts.

Browne, W. J., & Lahi, M. G. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim Software Package*. Retrieved from http://www.cmm.bristol.ac.uk/learning-training/multilevel-models/samples.shtml#mlpowsim

Department for Education and Science. (1991). *The parent's charter: You and your child's education.* London: Department for Education and Science.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6,* 1–33.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *The American Statistician, 37,* 36–48.

FitzGibbon, C. T., & Tymms, P. (2002). Technical and ethical issues in indicator systems: doing things right and doing things wrong. *Education Policy Analysis Archives, 10,* 1–26.

Goldstein, H. (1980). Critical notice of fifteen thousand hours. *Journal of Child Psychology & Psychiatry, 21,* 363–369.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society, Series A, 170,* 941–954.

Goldstein, H., & Noden, P. (2003). Modelling social segregation. *Oxford Review of Education, 29,* 225–237.

Goldstein, H., Rasbash J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19,* 425–433.

Goldstein, H., & Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: A cross-classified multilevel analysis. *School Effectiveness and School Improvement, 8,* 219–230.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, 159,* 385–443.

Hox, J. (2002). *Multilevel analysis, techniques and applications.* Mahwah, NJ: Erlbaum.

International Statistical Institute. (1985). *Declaration on professional ethics*. Retrieved from http://isi.cbs.nl/ethics0index.htm

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society, A,* 172.

Longview. (2008). Scientific case for a new cohort study. Retrieved from http://www.longviewuk.com/pages/reportsnew.shtml

O'Muircheartaigh, C., & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, 162,* 437–446.

Robinson, W. S. (1951). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Royal Statistical Society. (1993). *Code of conduct.* Retrieved from http://www.rss.org.uk/main.asp?page=1875

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1980). Critical notice of fifteen thousand hours. *Journal of Child Psychology & Psychiatry, 21,* 363–369.

Visscher, A. (2001). Public school performance indicators: problems and recommendations. *Studies in Educational Evaluation, 27,* 199–214.

Yang, M., Goldstein, H., Rath, T., & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education, 25,* 469–483.

# 14

# *The Impact of Missing Data on the Ethical Quality of a Research Study*

**Craig K. Enders**
*Arizona State University*

**Amanda C. Gottschall**
*Arizona State University*

The purpose of this chapter is to explore the impact of missing data on the ethical quality of a research study. In doing so, we borrow heavily from the work of Rosenthal (1994) and Rosenthal and Rosnow (1984). The overarching principle of Rosenthal's (1994) work is that ethics is closely linked with the quality of a research study, such that high-quality studies are more ethically defensible than low-quality studies. Missing data pose an obvious threat to quality at the analysis stage of a study (e.g., when a researcher uses a missing data handling technique that is prone to bias), but ethical issues arise throughout the entire research process. Accordingly, we explore the linkage between quality and ethics at the design and data collection phase, the analysis phase, and the reporting phase. In doing so, we also apply Rosenthal and Rosnow's (1984) *cost–utility model* to certain missing data issues (see also Rosnow & Rosenthal, Chapter 3, this volume). In this framework, the costs associated with a study (e.g., potential harm to participants, time, money, resources) are weighed against its utility (e.g., potential benefits to participants, science, or society). As it relates to ethics, a study is more defensible when its benefits exceed its costs.

## Missing Data Mechanisms

To fully appreciate the impact that missing data can have on the quality (and thus the ethics) of a research study, it is necessary to understand missing data theory. Rubin and colleagues (Little & Rubin, 2002; Rubin, 1976)

developed a classification system for missing data problems that is firmly entrenched in the methodological literature. These so-called missing data mechanisms describe how the propensity for missing data is related to measured variables, if at all. From a practical perspective, missing data mechanisms serve as assumptions that dictate the performance of different analytic approaches. This section gives a brief conceptual description of Rubin's missing data mechanisms, and more detailed accounts are available elsewhere in the literature (Enders, 2010; Little & Rubin, 2002; Rubin, 1976; Schafer & Graham, 2002).

To begin, a *missing completely at random* (MCAR) mechanism occurs when the probability of missing data on a variable is unrelated to other measured variables and to the values of the variable itself. When these conditions hold, the observed data are a simple random sample of the hypothetically complete data set. To illustrate, suppose that an industrial organizational psychologist is studying psychological well-being in the workplace and finds that some of the well-being scores are missing for purely haphazard reasons (e.g., an employee left the study because she went on maternity leave, an employee quit because his spouse accepted a job in another state, or an employee was on vacation when the surveys were administered). An MCAR mechanism would describe this scenario if the reasons for missingness are uncorrelated with well-being scores and with other measured variables.

Under a *missing at random* (MAR) mechanism, the probability of missing data on a variable is related to the observed values of other variables in the analysis model, but not to the unobserved values of the variable itself. As an example, consider a health study where researchers restrict the administration of a sensitive sexual behavior questionnaire to participants that are above the age of 15. Provided that age is included in any analysis that involves the sexual behavior variable, this example satisfies the MAR mechanism because missingness is unrelated to sexual behavior. Said differently, there is no residual relationship between the propensity for missing data and sexual behavior after controlling for age.

Finally, a *missing not at random* (MNAR) mechanism occurs when the probability of missingness is directly related to the scores on the incomplete variable itself. For example, suppose that a psychologist is studying quality of life in a clinical trial for a new cancer medication and finds that a number of patients become so ill (i.e., their quality of life becomes so poor) that they can no longer participate in the study. This example is consistent with an MNAR mechanism because the probability of missing data on the quality of life measure is directly related to a participant's quality of life. As an aside, an MNAR mechanism can also result when a cause or correlate of missingness is omitted from an analysis (e.g., the health researchers from the MAR example analyze the sexual behavior data without incorporating age into the analysis).

From a practical perspective, Rubin's mechanisms are vitally important because they serve as assumptions that dictate the performance of different *missing data handling techniques*. For example, an analysis method that assumes an MCAR mechanism will produce accurate parameter estimates under a more limited set of circumstances than a technique that requires MAR data because MCAR is a more restrictive condition than MAR. Based on theory alone, it is possible to reject MCAR-based methods in favor of approaches that assume MAR or MNAR. The inherent difficulty with missing data problems is that there is no way to determine which mechanism is at play in a given analysis. Although the observed data may provide evidence against an MCAR mechanism, there is no way to empirically differentiate MAR from MNAR (establishing that there is or is not a relationship between an incomplete variable and the probability of missingness on that variable requires knowledge of the missing values). Consequently, the statistical and ethical quality of a missing data analysis ultimately relies on the credibility of one or more untestable assumptions, and the onus is on the researcher to choose and defend a particular set of assumptions. We explore the ethical ramifications of different analytic choices in more detail later in the chapter.

## Missing Data Techniques

A brief description of common analytic approaches is necessary before addressing ethical issues. Space limitations preclude a comprehensive overview of missing data handling options, so the subsequent sections describe techniques that are used with some regularity in the social and the behavioral sciences. Throughout the chapter, we make the argument that the ethical quality of a particular analysis is linked to the credibility of its assumptions, so this section organizes the techniques according to their assumptions about the missing data mechanism. The following descriptions are necessarily brief, but a number of resources are available to readers who want additional details (Allison, 2002; Enders, 2010; Graham, 2009; Little & Rubin, 2002; Schafer, 1997; Schafer & Graham, 2002).

### Atheoretical Methods

The group of atheoretical missing data handling procedures includes methods that are known to produce biases under any missing data mechanism or do not have a theoretical foundation that dictates their expected performance. This category includes many of the ad hoc solutions that have appeared in the literature over the past several decades, at least three

of which have enjoyed widespread use: mean imputation, last observation carried forward, and averaging the available items. Mean imputation replaces missing values on a variable with the arithmetic average of the complete observations. This method is among the worst approaches available because it severely distorts estimates of variation and association under any missing data mechanism. Last observation carried forward is an imputation procedure for longitudinal data that replaces missing repeated measures variables with the observation that immediately precedes dropout. This is one of the most widely used imputation techniques in medical studies and in clinical trials (Wood, White, & Thompson, 2004), despite the fact that the procedure is prone to bias, even under an MCAR mechanism (Molenberghs et al., 2004). Finally, in the context of item-level missing data on questionnaires, researchers in the social and behavioral sciences routinely compute scale scores by averaging the available item responses. For example, if a respondent answered 13 of 15 items on a one-dimensional scale, the scale score would be the average of the 13 complete items. We include this procedure in the atheoretical category because the methodological literature has yet to establish the conditions under which the procedure may or may not work. The lack of empirical support for this procedure is troubling, given that it is a widely used method for handling item-level missing data (Schafer & Graham, 2002).

### MCAR-Based Methods

The category of MCAR-based analyses includes three common approaches: listwise deletion, pairwise deletion, and regression imputation. Listwise deletion removes cases with missing data from consideration, whereas pairwise deletion discards cases on an analysis-by-analysis basis. Regression imputation takes the different tack of filling in missing values with predicted scores from a regression equation (this method is unbiased under MCAR, but only after applying adjustments to variance and covariance terms). Listwise and pairwise deletion are perhaps the most widely used missing data handling techniques in the social and behavioral sciences (Peugh & Enders, 2004), most likely because of their widespread implementation in computer software packages. Considered as a whole, MCAR-based methods are virtually never better than MAR-based approaches, even when the MCAR mechanism is plausible (e.g., because they lack power). For this reason, we argue that these techniques detract from the ethical quality of a study.

### MAR-Based Methods

*Maximum likelihood estimation* and *multiple imputation* are the principal MAR-based missing data handling procedures. Maximum likelihood

estimation (also referred to as *direct maximum likelihood* and *full information maximum likelihood*) uses an iterative algorithm to audition different combinations of population parameter values until it identifies the particular set of values that produce the best fit to the data (i.e., the highest log likelihood value). The estimation process is largely the same with or without missing data, except that missing data estimation does not require individuals to contribute a full set of scores. Rather, the estimation algorithm uses all the available data to identify the most probable population parameters. Importantly, the estimation process does not impute missing values during this process. On the other hand, multiple imputation is a three-step process that consists of an imputation phase, an analysis phase, and a pooling phase. The purpose of the imputation phase is to create multiple copies of the data set, each of which contains different estimates of the missing values. The imputation phase is essentially a regression-based procedure where the complete variables predict the incomplete variables. In the analysis phase, the researcher performs each analysis $m$ times, once for each imputed data set. The analysis phase yields $m$ sets of parameter estimates and standard errors that are subsequently combined into a single set of results in the pooling phase. Relative to MCAR-based analysis methods, maximum likelihood and multiple imputation are desirable because they yield unbiased estimates under either an MCAR or MAR mechanism. Even when the MCAR mechanism is plausible, MAR-based analyses are still superior because they maximize statistical power.

### MNAR-Based Methods

MNAR-based analyses simultaneously incorporate a model for the data (i.e., the analysis that would have been performed had the data been complete) and a model for the propensity for missing values. The *selection model* and the *pattern mixture model* are the two well-established frameworks for performing MNAR-based analyses. The selection model is a two-part model that combines the substantive analysis with an additional set of regression equations that predict the response probabilities for each incomplete outcome variable. For example, in a linear growth curve analysis, the selection part of the model consists of a set of logistic regressions that describe the probability of response at each measurement occasion. In the logistic model, each incomplete outcome variable has a binary missing data indicator, and the probability of response at wave $t$ depends on the outcome variable at time $t$ and the outcome variable from the previous data collection wave. Simultaneously estimating the two parts of the model adjusts the substantive model parameters to account for the MNAR mechanism.

The basic idea behind a pattern mixture model is to form subgroups of cases that share the same missing data pattern and to estimate the

substantive model within each pattern. Doing so yields pattern-specific estimates of the model parameters, and computing the weighted average of these estimates yields a single set of results that appropriately adjusts for an MNAR mechanism. For example, to apply the pattern mixture model to a linear growth curve analysis, the growth model parameters are first estimated separately for each missing data pattern, and the pattern-specific parameter values are subsequently averaged into a single set of estimates. Conceptually, the pattern mixture model is akin to a multiple group model, where the missing data patterns define the subgroups (e.g., in a longitudinal study, cases with two waves of data form a group, cases with three waves of data form a group, etc.).

Despite their intuitive appeal, MNAR models require potentially tenuous assumptions that go beyond the missing data mechanism. Among other things, the selection model relies heavily on multivariate normality, and even modest departures from normality can severely distort the estimates from the substantive portion of the model. In the case of a pattern mixture model, certain pattern-specific estimates are usually inestimable. For example, in a linear growth curve model, it is impossible to estimate a growth trajectory for a pattern of cases with a single observation, and it is impossible to estimate certain variance components for a pattern with two observations. Consequently, the researcher must specify values for all inestimable parameters in the model. Again, the final estimates are prone to substantial bias if these user-specified values are incorrect. It is worth noting that MNAR-based analysis models have received considerable attention in the methodological literature. Methodologists have proposed methods for addressing the practical limitations of these models (e.g., approaches for generating values for the inestimable parameters in a pattern mixture model), but these methods have not been thoroughly studied. Until further research accumulates, MNAR-based analysis techniques should be viewed with some caution.

## Ethical Issues Related to Design and Data Collection

Having established some necessary background information, this section explores ethical considerations that arise during the design and data collection phases of a study. When faced with the prospect of missing data, it may seem that a researcher's primary goal is to do damage control by minimizing the negative consequences to the study. This is largely true in situations where missingness is beyond the researcher's control, and attending to missing data issues before and during data collection can mitigate damage. Although it may seem counterintuitive to do so,

researchers can also incorporate intentional missing data into the data collection design. These so-called planning missingness designs can bolster the ethical quality of a study by reducing its costs and maintaining its utility. The subsequent sections describe these issues in more detail.

## An Ounce of Prevention

Obviously, the best approach to dealing with missing data is to avoid the problem altogether. For the purposes of this chapter, it is more interesting to explore the ethical issues that arise from missing data, so we chose not to focus on prevention strategies (a detailed discussion of this topic could easily fill an entire chapter by itself). Nevertheless, it is important to note that researchers have developed a variety of techniques for reducing attrition, and there is substantial literature available on this topic. Some of these retention strategies are specific to particular disciplines (e.g., Bernhard et al., 1998), whereas others are quite general. For a detailed review of retention and tracking strategies, we recommend that readers consult Ribisl et al. (1996) and the references contained therein. For a general discussion of design-based strategies for preventing missing data, interested readers can also refer to Chapter 4 in McKnight, McKnight, Sidani, and Figueredo (2007).

## The Role of Auxiliary Variables

When performing missing data analyses, methodologists frequently recommend a so-called *inclusive analysis strategy* that incorporates a number of auxiliary variables into the analysis (Collins, Schafer, & Kam, 2001; Enders, 2010; Graham, 2003, 2009; Rubin, 1996; Schafer & Graham, 2002). An *auxiliary variable* is a variable that would not have appeared in the analysis model had the data been complete but is included in the analysis because it is a potential correlate of missingness or a correlate of an incomplete variable. Auxiliary variables are beneficial because they can reduce bias (e.g., by making the MAR assumption more plausible) and can improve power (e.g., by recapturing some of the lost information). For these reasons, incorporating auxiliary variables into an MAR-based analysis is a useful strategy for mitigating the negative impact of missing data.

To illustrate the idea behind an inclusive analysis strategy, consider a health study where researchers restrict the administration of a sensitive sexual behavior questionnaire to participants who are above the age of 15. An analysis of the sexual behavior variable would only satisfy the MAR assumption if age (the cause of missingness) is part of the statistical model. Omitting age from the model would likely distort the estimates, whereas incorporating age into the analysis as an auxiliary variable would

completely eliminate nonresponse bias (assuming that age was the only determinant of missingness). Correlates of the incomplete analysis model variables are also useful auxiliary variables, regardless of whether they are also related to missingness. For example, a survey question that asks teenagers to report whether they have a steady boyfriend or girlfriend is a useful auxiliary variable because it is correlated with the incomplete sexual activity scores. Introducing correlates of the incomplete variables as auxiliary variables may or may not reduce bias, but doing so can improve power by recapturing some of the missing information.

The previous health study is straightforward because the researchers control the missing data mechanism and the cause of missingness is a measured variable. In most realistic situations, the missing data mechanism is unknown, and the true causes of missingness are unlikely to be measured variables. Consequently, implementing an effective inclusive analysis strategy requires proactive planning to ensure that the data include potential correlates of missingness and correlates of the variables that are likely to have missing data. Identifying correlates of incomplete variables is relatively straightforward (e.g., via a literature review), but selecting correlates of missingness usually involves educated guesswork (see Enders, 2010, for additional details). When all else fails, asking participants about their intentions to complete the study is also a possibility. For example, in the context of a longitudinal design, Schafer and Graham (2002) recommend using a survey question that asks respondents to rate their likelihood of dropping out of the study before the next wave of data collection. These authors argue that incorporating auxiliary variables such as this into the analysis "may effectively convert an MNAR situation to MAR" (p. 173). Given the potential pitfalls associated with MNAR models, taking proactive steps to satisfy the MAR assumption may be the best way to maximize the quality of the analysis.

Despite the potential benefits of doing so, collecting a set of auxiliary variables raises ethical concerns. In particular, adding variables to a study increases respondent burden, requires participants to devote more time to the study, and generally increases the potential for unintended negative consequences. The impact on participants is an important consideration in and of itself, but collecting additional variables may also affect the overall integrity of the data by inducing fatigue or reducing motivation. Of course, increasing respondent burden can increase the probability of attrition, which defeats the purpose of collecting auxiliary variables. The conventional advice from the missing data literature is to incorporate a comprehensive set of auxiliary variables (Rubin, 1996), but there is clearly a need to balance statistical issues with practical and ethical concerns. Establishing definitive guidelines for the size of the auxiliary variable set is difficult because the costs and benefits associated with these additional variables will likely vary across studies. Nevertheless, the most useful

auxiliary variables are those that have strong correlations with the incomplete analysis variables. For example, incorporating a single pretest score as an auxiliary variable is often more beneficial than using several variables with weak to moderate correlations. Consequently, identifying a small set of auxiliary variables that are likely to maximize the squared multiple correlation with the incomplete variables is often a good strategy.

### Documenting the Reasons for Missing Data

Researchers often view missing data as an analytic problem that they can address after the data are collected. However, documenting the reasons for attrition during the data collection phase is an important activity that can bolster the ethical quality of a study by making the subsequent analytic choices more defensible. Later in the chapter we propose an ethical continuum that differentiates missing data handling techniques according to the quality of the estimates that they produce. These classifications are inherently subjective because the data provide no mechanism for choosing between MAR and MNAR analysis models. Ultimately, researchers have to weigh the credibility of different untestable assumptions when choosing among missing data handling techniques. Defending analytic choices is difficult without knowing why the data are missing, so devoting resources to tracking the causes of attrition is important. Documenting and reporting the causes of missingness is also important for planning future studies because the information can help guide the selection of effective auxiliary variables (e.g., if a school-based study reports that student mobility is a common cause of attrition, then future studies might include a survey question that asks parents how likely they are to move during the course of the school year). Ultimately, this may improve the overall quality of scientific research by converting some MNAR analyses to MAR. Of course, it is usually impossible to fully document the real world causes of missing data, but this information is still highly valuable, even if it is incomplete or partially speculative. Interested readers can consult Enders, Dietz, Montague, and Dixon (2006) and Graham, Hofer, Donaldson, MacKinnon, and Schafer (1997) for examples of longitudinal studies that tracked and reported the sources of attrition.

### Planned Missing Data Designs

Much of this chapter is concerned with ethical issues related to unintentional missing data. The development of MAR-based analysis techniques has made planned missing data research designs a possibility. The idea of intentional missing data may sound absurd, but planned missingness designs solve important practical problems, and they do so without compromising a study's integrity. In particular, these designs can

cut research costs (e.g., money, time, resources) and can reduce respondent burden. Rosenthal and Rosnow (1984) and Rosenthal (1994) argue that research studies that minimize costs (e.g., those that require fewer resources and reduce respondent burden) are more ethically defensible than studies with high costs, so the use of intentional missing data can actually improve the ethical quality of a study (see also Rosnow & Rosenthal, Chapter 3, this volume). This section provides a brief description of planned missing data designs, and interested readers can consult Graham, Taylor, Olchowski, and Cumsille (2006) for additional details.

To illustrate the issue of reducing costs and resources, consider a schizophrenia study where a researcher wants to use magnetic resonance imaging (MRI) to collect neuroimaging data. In this scenario, data collection is both expensive and potentially inconvenient for study participants (e.g., because the researcher may only have access to the MRI facility during off-peak hours). To reduce costs, the researcher could obtain less expensive computed tomography (CT) scan data from all subjects and could restrict the MRI data to a random subsample of participants. As a second example, consider an obesity prevention study that uses body composition as an outcome variable. Using calipers to take skinfold measurements is a widely used and inexpensive approach for measuring body fat. However, caliper measurements are often unreliable, so a better approach is to use hydrostatic (i.e., underwater) weighing or air displacement. Like the MRI, hydrostatic weighing and air displacement measures are expensive, and the equipment is difficult to access. In a planned missing data design, the prevention researchers could use calipers to collect body composition data from the entire sample and could restrict the more expensive measures to a subset of participants. Importantly, MAR-based analysis methods allow the researchers from the previous examples to use the entire sample to estimate the associations between the expensive measures and other study variables, even though a subset of the sample has missing data. If the researchers simultaneously incorporate the inexpensive measures (e.g., the CT scan data and the caliper measurements) into the analysis as auxiliary variables, the reduction in power resulting from missing data may be minimal.

Next, consider the issue of respondent burden. In Rosenthal and Rosnow's (1984) cost–utility framework, respondent burden would be one of the costs associated with doing research. Consequently, studies that minimize respondent burden are more ethically defensible than studies that impose a high burden. The previous scenarios illustrate how planned missing data can reduce respondent burden (e.g., by reducing the number of subjects who need to undergo an MRI during off-peak hours), but there are other important examples. For instance, researchers in the social and behavioral sciences routinely use multiple-item questionnaires to measure constructs (e.g., psychologists use several

questionnaire items to measure depression, each of which taps into a different depressive symptom). Using multiple-item scales to measure even a small number of constructs can introduce a substantial time burden. Obviously, reducing the number of questionnaires or reducing the number of items on each questionnaire can mitigate the problem, but these strategies may be undesirable because they can limit a study's scope and can reduce the content validity of the resulting scale scores. Planned missing data designs are an excellent alternative. In the context of questionnaire research, planned missingness designs distribute questionnaire items (or entire questionnaires) across different forms, such that any single participant responds to a subset of the items. Again, it is important to note that MAR-based methods allow researchers to perform their analyses as though the data were complete. Consequently, these designs reduce respondent burden without limiting the scope of the study or the content validity of the scale scores.

Respondent burden is also a serious problem in longitudinal studies. Graham, Taylor, and Cumsille (2001) describe a number of planned missingness designs for longitudinal studies. The basic idea behind these designs is to divide the sample into a number of random subgroups, each of which has a different missing data pattern. For example, in a study with six data collection waves, one subgroup may have intentional missing data at the third wave; another subgroup may be missing at the fifth wave; and yet another subgroup may have missing values at the second and the fourth waves. Interestingly, Graham et al. show that longitudinal planned missingness designs can achieve higher power than complete-data designs that use the same number of data points. This finding has important implications for maximizing the resources in a longitudinal study. For example, in a five-wave study with a budget that allows for 1,000 total assessments, collecting incomplete data from a sample of 230 respondents can produce higher power than collecting complete data from a sample of 200 respondents.

Researchers are sometimes skeptical of planned missing data designs, presumably because they hold the belief that missing data are harmful and something to avoid. It is important to emphasize that planned missingness designs produce MCAR data, so the intentional missing values that result from these designs are completely benign and are incapable of introducing bias. The primary downside of these designs is a reduction in statistical power. However, empirical studies suggest that the loss in power may be rather small (Enders, 2010; Graham et al., 2001, 2006), and researchers can mitigate this problem by carefully deciding which variables to make missing (preliminary computer simulations are particularly useful in this regard). Given their potential benefits, planned missing data designs may be an ethical imperative, particularly for high-cost studies.

## Ethical Issues Related to Data Analysis

During the analysis phase, researchers have to make a number of important decisions, the most obvious being the choice of analytic technique. Later in this section, we explore quality differences among missing data handling techniques and propose an ethical continuum that ranks analytic methods according to the quality of the estimates that they produce. This section also explores a number of other analytic issues that can impact the quality of a research study.

### How Much Missing Data Is Too Much?

One question that often arises with missing data is, "How much is too much?" A recent report by the American Psychological Association (APA) speculated that publishing missing data rates in journal articles will prompt researchers to "begin considering more concretely what acceptable levels of attrition are" (APA Publications and Communications Board Work Group on Journal Article Reporting Standards, 2008, p. 849). Establishing useful cutoffs for an acceptable level of attrition is difficult because it is the missing data mechanism that largely dictates the performance of an analytic method, not the percentage of missing data. In truth, the missing data rate may not be that important, provided that underlying assumptions are satisfied. As an example, some planned missing data designs (e.g., the three-form design) produce a 67% missing data rate for certain pairs of variables. This seemingly alarming amount of missing data causes no problems because the data are MCAR, by definition. Using MAR-based methods to analyze the data from such a design can produce unbiased parameter estimates with surprisingly little loss in power (Enders, 2010; Graham et al., 2006).

To be fair, high missing data rates (or even small to moderate missing data rates, for that matter) can be detrimental when the missing data mechanism is beyond the researcher's control, as it typically is. For example, most researchers would be uncomfortable with 67% attrition in a longitudinal clinical trial, and rightfully so. Here again, the missing data mechanism is the problem, not the attrition rate per se. If the reasons for missingness are largely unrelated to the outcome variable after controlling for other variables in the analysis (i.e., the mechanism is MAR), then the resulting parameter estimates should be accurate, albeit somewhat noisy. However, if missingness is systematically related to the outcome variable (i.e., the mechanism is MNAR), then the parameter estimates may be distorted. Unfortunately, when the reasons for missingness are beyond the researcher's control, it is impossible to use the observed data

to differentiate these two scenarios, so there is usually no way to gauge the impact of the missing data rate on the validity of a study's results.

Given that the missing data mechanism is usually unknown, determining what is and is not an acceptable level of attrition becomes a bit of an arbitrary exercise. Nevertheless, journal editors and reviewers do impose their own subjective criteria when evaluating manuscripts. As an example, a former student recently contacted me for advice on dealing with a manuscript revision where 90% of the sample had dropped out by the third and final wave of data collection. Not surprisingly, the journal editor and the reviewers voiced legitimate concerns about missing data. In this situation, assuaging the reviewers that the missing data pose no problem is impossible because the missing data mechanism is unknown. This scenario raises an interesting ethical question: In light of extreme attrition, is it better to report a potentially flawed set of results, or is it better to discard the data altogether? The word "potentially" is important because a high missing data rate does not necessarily invalidate or bias the analysis results.

Rosenthal and Rosnow's (1984) cost–utility framework is useful for considering the ethical ramifications of abandoning analyses that suffer from serious attrition problems. The basic premise of the cost–utility model is that the decision to conduct a study depends on the cost–utility ratio of doing the research (e.g., conducting a study with high costs and low utility is indefensible) and the cost–utility ratio of *not* doing the research (e.g., failing to conduct a study that may produce positive outcomes may be unethical). When considering the ethics of conducting a study, Rosenthal and Rosnow argue that "The failure to conduct a study that could be conducted is as much an act to be evaluated on ethical grounds as is the conducting of a study" (p. 562). Applying this idea to missing data, it is reasonable to argue that the failure to report the results from a study with high attrition is as much an act to be evaluated on ethical grounds as is the reporting of such results. We suspect that editors and reviewers generally consider only one side of the ethical issue, the costs of reporting results that are potentially distorted by missing data. However, the costs of discarding the data are not necessarily trivial and are also important to consider. Among other things, these costs include (a) the loss of all potential benefits (e.g., new knowledge, positive outcomes) from the study, (b) the waste of time, resources, and money that accrued from conducting the study, and (c) the fact that any negative impact that the study might have had on participants was for naught.

As an aside, researchers sometimes believe that MAR-based analyses work as long as the missing data rate falls below some critical threshold. The thought is, if the proportion of missing data exceeds this threshold, MAR methods become untrustworthy and ad hoc missing data handling approaches (e.g., MCAR-based methods) provide more accurate results. To be clear, this view is not supported by the methodological literature.

Simulation studies have repeatedly shown that the advantages of using MAR-based approaches over MCAR-based methods increase as the missing data rate increases. With small amounts of missing data, the differences between competing methods tend to be relatively small, but the relative benefits of MAR methods increase as the proportion of missing data increases. Consequently, there is no support for the notion that high missing data rates are a prescription for avoiding MAR-based methods in favor of more "conservative" traditional approaches.

### Imputation Is Not Just Making Up Data

Researchers sometimes object to imputation, presumably because they equate it to the unethical practice of fabricating data. For example, in the decision letter to my former student, the journal editor stated, "I have never been a fan of imputation." This type of cynicism is largely valid for single imputation (e.g., mean imputation, regression imputation, last observation carried forward, averaging the available items) techniques because filling in the data with a single set of values and treating those values as though they are real data produces standard errors that are inappropriately small. Of course, the other problem with most single imputation procedures is that they tend to produce biased parameter estimates, irrespective of their standard errors.

Importantly, MAR-based multiple imputation does not suffer from these problems because it (a) has a strong theoretical foundation, (b) produces accurate estimates under an MCAR and MAR mechanism, and (c) incorporates a correction factor that appropriately adjusts standard errors to compensate for the uncertainty associated with the imputed values. From a mathematical perspective, it is important to realize that multiple imputation and maximum likelihood estimation are asymptotically (i.e., in large samples) equivalent procedures. Maximum likelihood estimation produces estimates that effectively average over an infinite number of imputed data sets, although it does so without filling in the values. Multiple imputation uses a simulation-based approach that repeatedly fills in the missing data to accomplish the same goal. Some of the objections to imputation may stem from the fact that researchers place undue emphasis on the filled-in data values without considering the fact that the data set is just a means to a more important end, which is to estimate the population parameters. In truth, multiple imputation is nothing more than a mathematical tool for achieving this end goal. In that sense, it is the final parameter estimates that matter, not the imputed values themselves.

### Revisiting an Inclusive Analysis Strategy

Earlier in the chapter we described an inclusive analysis strategy that incorporates auxiliary variables (correlates of missingness or correlates of

the incomplete analysis variables) into the statistical model. In line with the premise that research quality and ethics are linked, we believe that MAR-based analyses that incorporate auxiliary variables are more ethically defensible than analyses that do not. For one, an inclusive analysis strategy is more likely to satisfy the MAR assumption, thereby reducing the potential for bias. Auxiliary variables can also mitigate the power loss resulting from missing data, thereby maximizing resources and reducing costs. As an example, Baraldi and Enders (2010) used data from the Longitudinal Study of American Youth to illustrate the impact of auxiliary variables. In their analysis, including three useful auxiliary variables in a regression model reduced standard errors by an amount that was commensurate with increasing the total sample size by 12% to 18% (the magnitude of the reduction in standard errors varied across regression coefficients). From a cost–utility perspective, there is no question that an inclusive analysis strategy is desirable because it maximizes existing resources, whereas collecting more data requires additional costs (e.g., time, money, risks to participants). The standard error reduction in the Baraldi and Enders (2010) study is probably close to the upper limit of what would be expected in practice, but even a modest improvement (e.g., a reduction in standard errors that is commensurate with a 5% increase in the sample size) supports our argument.

As an aside, it is possible for the same analysis to produce different estimates with and without the auxiliary variables. When this happens, there is no way of knowing which set of estimates is more accurate, but two points are worth remembering: The auxiliary variable analysis has the most defensible set of assumptions (i.e., it is more likely to satisfy MAR), and methodological studies have yet to identify detrimental effects of an inclusive strategy (e.g., including a large set of useless auxiliary variables does not appear to negatively impact the resulting estimates and standard errors; Collins et al., 2001). These two factors clearly favor the estimates from the auxiliary variable model, but ethical issues can arise if researchers are tempted to choose the analysis results that best align with their substantive hypotheses. To avoid this ethical pitfall, researchers should disclose the fact that the two analyses produced conflicting results, perhaps reporting the estimates from the alternate analysis in a footnote or in a supplementary appendix in the electronic version of the manuscript.

### An Ethical Continuum of Analysis Options

Researchers have a variety of options for analyzing data sets with missing values. Earlier in the chapter, we described four categories of missing data handling procedures: atheoretical methods, MCAR-based methods, MAR-based methods, and MNAR-based methods. In this section, we

propose an ethical continuum that differentiates missing data handling techniques according to the quality of the estimates that they produce. On one end, the continuum is anchored by low-quality approaches that are difficult to defend on ethical grounds, whereas the other end of the continuum is defined by defensible approaches that have a strong theoretical foundation. When comparing certain categories of methods, there are distinct and consistent quality differences that are difficult to dispute (e.g., there is little question that MAR-based procedures are more defensible than MCAR-based analyses). However, differentiating among techniques that rely on one or more untestable assumptions is a subjective exercise. Consequently, some readers will disagree with certain aspects of our proposed continuum, and rightfully so. In proposing the ethical continuum, it is not our intent to form rigid distinctions that cast a negative light on certain analytic choices. Quite the contrary, choosing among the theory-based approaches at the high end of the quality continuum requires researchers to judge the credibility of different sets of assumptions. The veracity of these assumptions will vary across situations, so the ordering of certain procedures is fluid.

Figure 14.1 shows a graphic of our proposed continuum. The low-quality end of the continuum is anchored by the collection of atheoretical analysis techniques. This group of procedures includes missing data handling procedures that (a) are known to produce biases under any missing data mechanism, (b) do not have a theoretical framework that dictates their expected performance, or (c) lack empirical evidence supporting their widespread use. It is worth noting that the low-quality endpoint includes at least three procedures that enjoy widespread use (mean imputation, last observation carried forward, and averaging the available items). As seen in the figure, MCAR-based approaches provide an improvement in quality. MCAR methods require a rather strict assumption about the cause of missing data (i.e., the propensity for missing data is unrelated to all study variables), but the situations where these techniques produce accurate parameter estimates are well established. However, even if the MCAR mechanism is plausible, MAR-based analyses generally increase

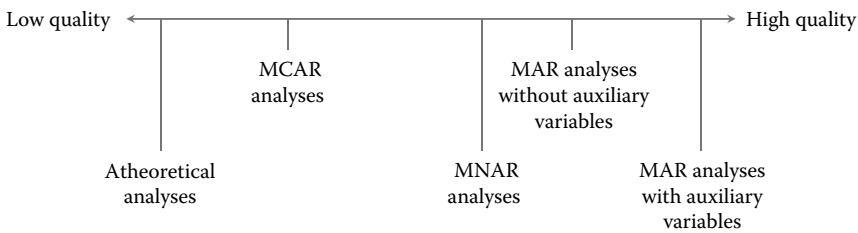



**FIGURE 14.1**
An ethical continuum of missing data handling techniques.

statistical power, thus making better use of available resources. This alone provides a strong ethical justification for abandoning MCAR-based missing data handling methods.

Assessing the relative quality of MAR- and MNAR-based analysis techniques is less clear-cut because the accuracy of these procedures relies on one or more untestable assumptions. To be clear, all the procedures at the high-quality end of the continuum are capable of producing unbiased parameter estimates when their requisite assumptions hold. Although some readers will likely disagree, we believe that the range of conditions that satisfies the assumptions of an MNAR-based analysis will generally be narrower than the range of conditions that satisfies the assumptions of an MAR-based analysis. Consequently, we assigned a slightly higher quality ranking to MAR-based analyses (i.e., multiple imputation and maximum likelihood estimation). In part, our rationale was based on the fact that MNAR models require assumptions that go beyond the missing data mechanism. For example, Enders (2010) gives an analysis example where a modest departure from normality causes a selection model to produce estimates that are less accurate than those of maximum likelihood estimation, despite the fact that the selection model perfectly explains the MNAR missing data mechanism. Pattern mixture models rely on equally tenuous assumptions because they require the analyst to specify values for one or more unknown parameters.

The ethical continuum assigns the highest-quality rating to MAR-based analyses that incorporate auxiliary variables. Although this choice will not be met with universal agreement, we believe that a well-executed MAR analysis generally has the most defensible assumptions, even when there is reason to believe that dropout is systematically related to the incomplete outcome variable. Other methodologists have voiced a similar opinion. For example, Schafer (2003, p. 30) discussed the tradeoffs between MAR and MNAR analysis models, stating that "Rather than rely heavily on poorly estimated MNAR models, I would prefer to examine auxiliary variables that may be related to missingness … and include them in a richer imputation model under assumption of MAR." Similarly, Demirtas and Schafer (2003, p. 2573) stated that "The best way to handle drop-out is to make it ignorable [i.e., consistent with an MAR mechanism]." They go on to recommend that researchers should collect data on variables that predict attrition and incorporate these variables into their analyses.

Again, it is important to reiterate that some of the classifications in Figure 14.1 are subjective and open to debate. Ultimately, the data provide no mechanism for choosing between MAR and MNAR analyses, so researchers have to weigh the credibility of different sets of untestable assumptions. Missing data techniques are only as good as the veracity of

their assumptions, so adopting a defensible analysis that minimizes the risk of violating key assumptions will maximize the ethical quality of a research study. The need to defend analytic choices has important implications for data collection (e.g., documenting the reasons for missingness, planning for attrition by collecting data on auxiliary variables) and for reporting the results from a missing data analysis. We address the latter topic in a subsequent section.

### Sensitivity Analyses

The purpose of a sensitivity analysis is to explore the variability of a parameter estimate across models that apply different assumptions. For example, in longitudinal studies, methodologists often recommend fitting MAR- and MNAR-based growth models to the same data. This strategy seems eminently sensible given the difficulty of defending a set of untestable assumptions. If the key parameter estimates are stable across different missing data models, then the choice of analytic procedure makes very little difference. Unfortunately, it is relatively common for *sensitivity analyses* to produce discrepant sets of estimates. For example, Enders (2010) used an artificial data set to illustrate a sensitivity analysis for a linear growth model with a binary treatment status variable as a predictor. None of the five analysis models accurately reproduced the true parameter estimates, and the estimates varied dramatically across models (e.g., the MAR-based growth model and the MNAR-based selection model underestimated the true effect size, whereas MNAR-based pattern mixture models overestimated the true effect). Other methodologists have reported similar discrepancies from sensitivity analyses (Demirtas & Schafer, 2003; Foster & Fang, 2004).

In many situations, sensitivity analyses add no clarity to analytic choices, and researchers are left to decide among potentially disparate sets of results. Unfortunately, the data provide no mechanism for choosing among competing analyses, and two models that produce different estimates can produce comparable fit. Ideally, researchers should choose the estimates from the model with the most defensible set of assumptions, but it may be tempting to adopt the estimates that best align with the substantive hypotheses. Sensitivity analyses with longitudinal data seem particularly prone to this ethical dilemma because estimates can vary so dramatically from one model to the next. To avoid this ethical dilemma, researchers should present an argument that supports their assumptions and should disclose the fact that estimates differed across analysis models. Ideally, the estimates from alternate analyses could appear in a footnote or in a supplementary appendix in the electronic version of the manuscript.

## Ethical Issues Related to Reporting

A final set of ethical concerns arises when reporting the results from a missing data analysis. Rosenthal (1994) describes a number of ethical issues related to research reporting, most of which involve misrepresentation of research findings (e.g., inappropriately generalizing, making claims that are not supported by the data, failing to report findings that contradict expectations). In the context of a missing data analysis, two additional forms of misrepresentation are problematic: providing an insufficient level of detail about the missing data and the treatment of missing data, and overstating the benefits of a missing data handling technique. This section explores these two issues in detail.

### Reporting Standards for Missing Data Analyses

A 1999 report by the APA Task Force on Statistical Inference encouraged authors to report unanticipated complications that arise during the course of a study, including "missing data, attrition, and nonresponse" (Wilkinson & Task Force on Statistical Inference, 1999, p. 597). At the time of the Task Force report, missing data reporting practices were abysmal, and many published research studies failed to report any information about missing data. In a comprehensive methodological review, Peugh and Enders (2004) examined hundreds of published articles in the 1999 and 2003 volumes of several education and psychology journals. In the 1999 volumes, approximately one third of the articles with detectable missing data (e.g., studies where the degrees of freedom values unexpectedly changed across a set of analyses of variance) explicitly acknowledged the problem. Whether it was a result of the Task Force report or a general increase in awareness of missing data issues, reporting practices improved in the 2003 volumes, such that three quarters of the studies with detectable missing data disclosed the problem. Obviously, failing to report any information about missing data is a gross misrepresentation, regardless of intent (to be fair, the researchers that authored the papers in the review were probably unaware that missing data pose a problem). Missing data reporting practices have arguably progressed since 1999, but there is still room for improvement.

Recently, several organizations have published detailed guidelines aimed at improving reporting practices in scientific journals. In the social sciences, the American Educational Research Association (2006) published the *Standards for Reporting on Empirical Social Science Research in AERA Publications*, and APA published *Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be?* (APA Publications and Communications Board Work Group on Journal Article Reporting

Standards, 2008). Similar reports have appeared in the medical and clinical trials literature (Altman et al., 2001; Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004; Moher, Schulz, & Altman, 2001). Although these reports have a general focus, they do provide specific recommendations for dealing with missing data. The APA Journal Article Reporting Standards (JARS) report provides the most comprehensive recommendations concerning missing data, so we briefly summarize its main points here.

The JARS report recommends that researchers describe (a) the percentage of missing data, (b) empirical evidence or theoretical arguments in support of a particular missing data mechanism, (c) the missing data handling technique that was used for the analyses, and (d) the number and characteristics of any cases that were deleted from the analyses. Following guidelines from the clinical trials literature (the Consolidated Standards of Reporting Trials, or CONSORT statement; Moher et al., 2001), the JARS report recommends a diagrammatic flowchart that, among other things, describes the amount of and the reasons for missing data at each stage of the research process (see p. 846). We believe that the JARS recommendations are adequate for most studies, but some analyses may require additional details (e.g., planned missing data designs).

Given the rather abysmal state of missing data reporting practices, it is hard to argue against the need for more detailed research reports. Nevertheless, devoting additional space to missing data issues decreases the amount of journal space that is available for reporting substantive results. In some situations, satisfying the JARS recommendations requires relatively little journal space, whereas other situations are more demanding. For example, a thorough description of a multiple imputation procedure may be somewhat lengthy because it involves many nuances and subjective choices. Similarly, planned missing data designs often require preliminary computer simulations to assess the power of different missing data patterns, and describing these preliminary analyses may require an excessive amount of journal space. As a compromise, researchers may want to rely more heavily on electronic resources to convey the procedural details of their missing data analyses. In situations where the missing data handling procedure is very involved, the printed version of the manuscript could include a brief description of the core analytic details, and the electronic version could include an appendix that documents the analytic steps in more elaborate detail.

### Overstating the Benefits of an Analytic Technique

MAR and MNAR analyses are sometimes met with skepticism because they are relatively new to many disciplines. When faced with the prospect of "selling" an unfamiliar missing data technique, a natural inclination is to provide a detailed description of the analysis along with empirical

evidence that supports its use (e.g., references to computer simulation studies that demonstrate the procedure's efficacy). However, many researchers are unfamiliar with Rubin's missing data mechanisms, so describing the benefits of an analysis without also describing its assumptions can mislead the reader into believing that the procedure is an analytic panacea for missing data. A similar type of misrepresentation can occur when a manuscript provides insufficient details about the missing data handling procedure. Consequently, it is important for authors to provide a thorough but balanced description of the missing data handling procedure that addresses the benefits and the assumptions of their analytic choices.

The recommendation to use honest and balanced reporting practices is unlikely to be met with objections. However, the pressure to publish in top-tier journals creates situations that are at odds with this practice. As an example, I previously described an interaction with a former student who contacted me for advice on dealing with a manuscript revision that involved a substantial amount of missing data. In the decision letter, the journal editor responded to the use of maximum likelihood estimation by saying that "I have never been a big fan of imputation." The editor's response is misguided because it incorrectly characterizes maximum likelihood estimation as an imputation technique and because it implies that imputation is an inherently flawed procedure (presumably, the editor's opinion stems from the misconception that imputation is "making up data"). Ignoring the problems associated with the high attrition rate, the editor's objection is easily addressed by describing the benefits of maximum likelihood estimation and bolstering this description with relevant citations from the methodological literature. The ethical concern is that the revised manuscript could overstate the benefits of maximum likelihood while downplaying (or completely omitting any discussion of) the untestable MAR assumption. This type of unbalanced reporting can potentially mislead readers who are unfamiliar with the intricacies of missing data analyses.

Misrepresentation can also occur when authors fail to describe their missing data handling procedures in sufficient detail. As an example, consider a hypothetical passage from a manuscript that reads, "We used maximum likelihood estimation, a missing data handling technique that the methodological literature characterizes as state of the art." Although it is true that methodologists have described MAR-based methods in this way (Schafer & Graham, 2002), this passage is misleading because it implies that maximum likelihood estimation is a cure-all for missing data problems. As a second example, consider a hypothetical passage that states, "Because there is reason to believe that the data are missing not at random (i.e., attrition is systematically related to the outcome variable), we used a selection model to correct for attrition-related bias." The lack of detail in the preceding passage is potentially misleading because it

fails to inform the reader that the accuracy of the selection model heavily depends on the multivariate normality assumption—so much so that an MAR-based analysis may yield better estimates in many situations. The fact that many researchers are unfamiliar with the nuances of MAR- and MNAR-based analysis techniques magnifies ethical concerns related to lack of detail in reporting because the factors that affect the performance of a particular technique may not be widely understood.

Fortunately, ethical issues related to misrepresentation are easily avoided by following recommendations from the JARS report. In particular, the report states that manuscripts should include "evidence and/or theoretical arguments for the causes of data that are missing, for example, missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR)" (p. 843). Because the missingness mechanism largely dictates the performance of most missing data techniques, discussing the plausibility of a purported mechanism can delineate the range of conditions under which an analytic method is likely to produce accurate estimates. Further describing the importance of other assumptions (e.g., normality) reduces the chances of leaving the reader with an overly optimistic impression of the analysis.

## Conclusion

The purpose of this chapter is to explore the impact of missing data on the ethical quality of a research study. Consistent with Rosenthal and colleagues (Rosenthal, 1994; Rosenthal & Rosnow, 1984), we operate from the premise that ethics is closely linked to the quality of a research study, such that high-quality studies are more ethically defensible than low-quality studies. Missing data are obviously important at the analysis phase, but ethical issues arise throughout the entire research process. Accordingly, we explore the linkage between quality and ethics at the design and data collection phase, the analysis phase, and the reporting phase.

During the design and the data collection phase, researchers should proactively plan for missing data to minimize negative consequences to the study. In particular, collecting data on auxiliary variables can make the MAR assumption more plausible, and documenting the reasons for attrition can help build an argument that supports subsequent analytic choices. Although it may seem counterintuitive to do so, researchers can also incorporate intentional missing data into the data collection design. These so-called planning missingness designs can bolster the ethical quality of a study by reducing costs and respondent burden. Given their potential benefits, planned missing data designs may be an ethical imperative, particularly for high-cost studies.

During the analysis phase, researchers have to make a number of important decisions, the most obvious being the choice of analytic technique. We propose an ethical continuum that differentiates missing data handling methods according to the quality of the estimates that they produce. MCAR-based analysis techniques are rarely justified, so the choice is usually between MAR and MNAR models. Both sets of procedures are capable of producing accurate parameter estimates when their requisite assumptions hold, but they are also prone to bias when the assumptions are violated. Because MNAR models require strict assumptions that go beyond the missing data mechanism (e.g., in the case of a selection model, multivariate normality), we argue that the range of conditions that satisfies the assumptions of an MNAR-based analysis will generally be narrower than the range of conditions that satisfies the assumptions of an MAR-based analysis. In our view, an MAR-based analysis that incorporates auxiliary variables is often the most defensible procedure, even when there is reason to believe that dropout is systematically related to the incomplete outcome variable.

Finally, we explored ethical issues related to reporting the results from a missing data analysis. Recently, several organizations have published detailed guidelines aimed at improving reporting practices in scientific journals, and these reporting guidelines generally include recommendations regarding missing data. The APA JARS report is particularly detailed and recommends that researchers describe (a) the percentage of missing data, (b) empirical evidence or theoretical arguments in support of a particular missing data mechanism, (c) the missing data handling technique that was used for the analyses, and (d) the number and characteristics of any cases that were deleted from the analyses.

In summary, maximizing the ethical quality of a study requires researchers to attend to missing data throughout the entire research process. We believe that a good MAR analysis will often lead to better estimates than an MNAR analysis. Ultimately, the data provide no mechanism for choosing between MAR and MNAR analyses, so researchers have to weigh the credibility of different sets of untestable assumptions when making this choice. Adopting a defensible analysis that minimizes the risk of violating key assumptions maximizes the ethical quality of a research study, and achieving this goal is only possible with careful planning during the design and data collection phase.

## References

Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., … Lang, T. (2001). The revised CONSORT statement for reporting

randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663–694.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35,* 33–40.

APA Publications and Communications Board Work Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5–37.

Bernhard, J., Cella, D. F., Coates, A. S., Fallowfield, L., Ganz, P. A., Moinpour, C. M., … Hürny, C. (1998). Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Statistics in Medicine, 17,* 517–532.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330–351.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine, 22,* 2553–2575.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *American Journal of Public Health, 94,* 361–366.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K., Dietz, S., Montague, M., & Dixon, J. (2006). Modern alternatives for dealing with missing data in special education research. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disorders* (Vol. 19, pp. 101–130). New York: Elsevier.

Foster, E. M., & Fang, G.Y. (2004). Alternative methods for handling attrition: An illustration using data from the Fast Track evaluation. *Evaluation Review, 28,* 434–464.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10,* 80–100.

Graham, J.W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K .J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325–366). Washington, DC: American Psychological Association.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in the analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of Internal Medicine, 134,* 657–662.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics, 5,* 445–464.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525–556.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5,* 127–134.

Rosenthal, R., & Rosnow, R. L. (1984). Applying Hamlet's question to the ethical conduct of research: A conceptual addendum. *American Psychologist, 39,* 561–563.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473–489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica, 57,* 19–35.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials, 1,* 368–376.

# 15

# *The Science and Ethics of Causal Modeling*

**Judea Pearl**
*University of California, Los Angeles*

The research questions that motivate most quantitative studies in the health, social, and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given treatment or program in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual in a specific incident? These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone.

Solving causal problems mathematically requires certain extensions in the standard mathematical language of statistics, and these extensions are not generally emphasized in the mainstream literature and education. As a result, a profound tension exists between the scientific questions that a researcher wishes to ask and the type of questions traditional analysis can accommodate, let alone answer. Bluntly, scientists speak causation, and statistics delivers correlation. This tension has resulted in several ethical issues concerning the statement of a problem, the implementation of a study, and the reporting of finding. This chapter describes a simple causal extension to the language of statistics, shows how it leads to a coherent methodology that avoids the ethical problems mentioned, and permits researchers to benefit from the many results that causal analysis has produced in the past 2 decades.

Following an introductory section that defines the demarcation line between associational and causal analysis, the rest of the chapter will deal with the estimation of three types of causal queries: (a) queries about the effect of potential interventions, (b) queries about *counterfactuals* (e.g., whether event $x$ would occur had event $y$ been different), and (c) queries about the *direct* and *indirect effects*.

## From Associational to Causal Analysis: Distinctions and Barriers

### The Basic Distinction: Coping With Change

The aim of standard statistical analysis, typified by regression, estimation, and hypothesis testing techniques, is to assess parameters of a distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate probabilities of past and future events, and update probabilities of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only probabilities of events under static conditions but also the dynamics of events under *changing conditions*, for example, changes induced by treatments or external interventions.

This distinction implies that causal and associational concepts do not mix. There is nothing in the joint distribution of symptoms and diseases to tell us whether curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change—say from observational to experimental setup—because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by causal assumptions that identify those relationships that remain invariant when external conditions change.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: One cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies.[1]

### Formulating the Basic Distinction

A formal demarcation line that makes the distinction between associational and causal concepts crisp and easy to apply can be formulated as follows. An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. Examples of associational concepts are correlation, regression,

[1] The methodology of "causal discovery" (Pearl, 2000b, Chapter 2; Spirtes, Glymour, & Scheines, 2000) is likewise based on the causal assumption of "faithfulness" or "stability" but will not be discussed in this chapter.

dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odds ratio, marginalization, Granger causality, conditionalization, "controlling for," and so on. Examples of causal concepts are randomization, influence, effect, confounding, "holding constant," disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, mediation, and attribution. The former can, whereas the latter cannot, be defined in terms of distribution functions.

This demarcation line is extremely useful in tracing the assumptions that are needed for substantiating various types of scientific claims. Every claim invoking causal concepts must rely on some premises that invoke such concepts; it cannot be inferred from, or even defined in, terms of statistical associations alone.

### Ramifications of the Basic Distinction

This principle has far-reaching consequences that are not generally recognized in the standard statistical literature. Many researchers, for example, are still convinced that confounding is solidly founded in standard, frequentist statistics, and that it can be given an associational definition saying (roughly): "$U$ is a potential confounder for examining the effect of treatment $X$ on outcome $Y$ when both $U$ and $X$ and $U$ and $Y$ are not independent." That this definition and all its many variants must fail (Pearl, 2009a, Section 6.2)[2] is obvious from the demarcation line above; if confounding were definable in terms of statistical associations, we would have been able to identify confounders from features of nonexperimental data, adjust for those confounders, and obtain unbiased estimates of causal effects. This would have violated our golden rule: Behind any causal conclusion there must be some causal assumption, untested in observational studies. Hence the definition must be false. Therefore, to the bitter disappointment of generations of epidemiologists and social science researchers, confounding bias cannot be detected or corrected by statistical methods alone; one must make some judgmental assumptions regarding causal relationships in the problem before an adjustment (e.g., by stratification) can safely correct for confounding bias.

This distinction implies that causal relations cannot be expressed in the language of probability and hence that any mathematical approach to causal analysis must acquire new notation for expressing causal relations—probability calculus is insufficient. To illustrate, the syntax of probability calculus does not permit us to express the simple fact that "symptoms do not cause diseases," let alone draw mathematical conclusions from such facts. All

[2] Any intermediate variable $U$ on a causal path from $X$ to $Y$ satisfies this definition, without confounding the effect of $X$ on $Y$.

we can say is that two events are dependent—meaning that if we find one, we can expect to encounter the other, but we cannot distinguish statistical dependence, quantified by the conditional probability $P(disease \mid symptom)$ from causal dependence, for which we have no expression in standard probability calculus. Therefore, scientists seeking to express causal relationships must supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the relation "symptoms cause disease" is distinct from the symbolic representation of "symptoms are associated with disease."

## Two Mental Barriers: Untested Assumptions and New Notation

The preceding two requirements: (a) to commence causal analysis with untested,[3] theoretically or judgmentally based assumptions, and (b) to extend the syntax of probability calculus to articulate such assumptions, constitute the two main sources of confusion in the ethics of formulating, conducting, and reporting empirical studies.

Associational assumptions, even untested, are testable in principle, given a sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference stands out in Bayesian analysis. Although the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to prior causal assumptions, say that treatment does not change gender, remains substantial regardless of sample size.

This makes it doubly important that the notation we use for expressing causal assumptions be cognitively meaningful and unambiguous so that one can clearly judge the plausibility or inevitability of the assumptions articulated. Statisticians can no longer ignore the mental representation in which scientists store experiential knowledge because it is this representation and the language used to access that representation that determine the reliability of the judgments on which the analysis so crucially depends.

How does one recognize causal expressions in the statistical literature? Those versed in the potential–outcome notation (Holland, 1988; Neyman, 1923; Rubin, 1974) can recognize such expressions through the subscripts that are attached to counterfactual events and variables, for example, $Y_x(u)$ or $Z_{xy}$. Some authors use parenthetical expressions, for example, $Y(0)$, $Y(1)$, $Y(x, u)$, or $Z(x, y)$. The expression $Y_x(u)$, for example, may stand for the value that outcome $Y$ would take in individual $u$, had treatment $X$ been at level $x$. If $u$ is chosen at random, $Y_x$ is a random variable, and one can talk about the probability that $Y_x$ would attain a value $y$ in the population,

[3] By "untested" I mean untested using frequency data in nonexperimental studies.

written $P(Y_x = y)$. Alternatively, Pearl (1995) used expressions of the form $P(Y = y \mid set(X = x))$ or $P(Y = y \mid do(X = x))$ to denote the probability (or frequency) that event $(Y = y)$ would occur if treatment condition $(X = x)$ were enforced uniformly over the population.[4] Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality.[5]

However, few have taken seriously the textbook requirement that any introduction of new notation must entail a systematic definition of the syntax and semantics that govern the notation. Moreover, in the bulk of the statistical literature before 2000, causal claims rarely appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain statistical parameters (e.g., regression coefficients), and in the verbal description with which investigators justify assumptions. For example, the assumption that a covariate not be affected by a treatment—a necessary assumption for the control of confounding (Cox, 1958, p. 48)—is expressed in plain English, not in a mathematical expression.

The next section provides a conceptualization that overcomes these mental barriers; it offers both a friendly mathematical machinery for cause–effect analysis and a formal foundation for counterfactual analysis.

## Structural Causal Models, Diagrams, Causal Effects, and Counterfactuals

### Structural Equations as Oracles for Causes and Counterfactuals

How can one express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such a relationship mathematically was made in the 1920s by the geneticist Sewall Wright (1921), who used a combination of equations and graphs. For example, if $X$ stands for a disease variable and $Y$ stands for a certain symptom of the disease, Wright would write a linear equation:

$$y = \beta x + u \tag{15.1}$$

[4] Clearly, $P(Y = y \mid do(X = x))$ is equivalent to $P(Y_x = y)$. This is what we normally assess in a controlled experiment, with $X$ randomized, in which the distribution of $Y$ is estimated for each level $x$ of $X$.

[5] These notational clues should be useful for detecting inadequate definitions of causal concepts; any definition of confounding, randomization, or instrumental variables that is cast in standard probability expressions, void of graphs, counterfactual subscripts, or $do(*)$ operators, can safely be discarded as inadequate.

where $x$ stands for the level (or severity) of the disease, $y$ stands for the level (or severity) of the symptom, and $u$ stands for all factors, other than the disease in question, that could possibly affect $Y$. In interpreting this equation one should think of a physical process whereby Nature *examines* the values of $x$ and $u$ and, accordingly, *assigns* variable $Y$ the value $y = \beta x + u$.

To express the directionality inherent in this assignment process, Wright augmented the equation with a diagram, later called "path diagram," in which arrows are drawn from (perceived) causes to their (perceived) effects, and, more importantly, the absence of an arrow makes the empirical claim that the value Nature assigns to one variable is indifferent to that taken by another (see Figure 15.1).

The variables $V$ and $U$ are called "exogenous"; they represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model.

If correlation is judged possible between two exogenous variables, $U$ and $V$, it is customary to connect them by a dashed double arrow, as shown in Figure 15.1b.

To summarize, path diagrams encode causal assumptions via missing arrows, representing claims of zero influence, and missing double arrows (e.g., between $V$ and $U$), representing the (causal) assumption $Cov(U, V) = 0$.

The generalization to nonlinear system of equations is straightforward. For example, the nonparametric interpretation of the diagram of Figure 15.2a corresponds to a set of three functions, each corresponding to one of the observed variables:

$$\begin{aligned} z &= f_Z(w) \\ x &= f_X(z, v) \\ y &= f_Y(x, u) \end{aligned} \tag{15.2}$$

where in this particular example, $W$, $V$, and $U$ are assumed to be jointly independent but, otherwise, arbitrarily distributed.

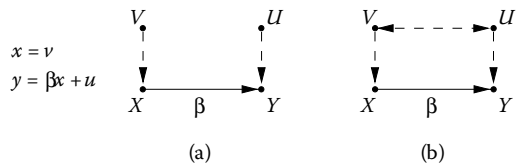



**FIGURE 15.1**
A simple structural equation model and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

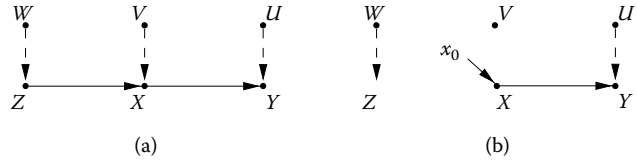



**FIGURE 15.2**
(a) The diagram associated with the structural model of Equation 15.2. (b) The diagram associated with the modified model, $M_{x_0}$, of Equation 15.3, representing the intervention $do(X = x_0)$.

Remarkably, unknown to most economists and philosophers,[6] *structural equation models* provide a formal interpretation and symbolic machinery for analyzing counterfactual relationships of the type: "$Y$ would be $y$ had $X$ been $x$ in situation $\boldsymbol{U} = \boldsymbol{u}$," denoted $Y_x(u) = y$. Here $\boldsymbol{U}$ represents the vector of all exogenous variables.[7]

The key idea is to interpret the phrase "had $X$ been $x_0$" as an instruction to modify the original model and replace the equation for $X$ by a constant $x_0$, yielding the submodel, $M_{x_0}$,

$$\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x, u) \end{aligned} \tag{15.3}$$

the graphical description of which is shown in Figure 15.2b.

This replacement permits the constant $x_0$ to differ from the actual value of $X$, namely, $f_X(z, v)$, without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multistage models, where the dependent variable in one equation may be an independent variable in another (Balke & Pearl, 1994; Pearl, 2000). In general, we can formally define the postintervention distribution by the equation:

$$P_M(y \mid do(x)) \stackrel{\Delta}{=} P_{M_x}(y). \tag{15.4}$$

In words: In the framework of model $M$, the postintervention distribution of outcome $Y$ is defined as the probability that model $M_x$ assigns to each outcome level $Y = y$.

[6] Connections between structural equations and a restricted class of counterfactuals were recognized by Simon and Rescher (1966). These were later generalized by Balke and Pearl (1995), who used modified models to permit counterfactual conditioning on dependent variables.

[7] Because $U = u$ may contain detailed information about a situation or an individual, $Y_x(u)$ is related to what philosophers called "token causation," whereas $P(Y_x = y|Z = z)$ characterizes "Type causation," that is, the tendency of $X$ to influence $Y$ in a subpopulation characterized by $Z = z$.

From this distribution, one is able to assess treatment efficacy by comparing aspects of this distribution at different levels of $x_0$. A common measure of treatment efficacy is the difference

$$E(Y \mid do(x_0')) - E(Y \mid do(x_0)) \tag{15.5}$$

where $x_0'$ and $x_0$ are two levels (or types) of treatment selected for comparison. For example, to compute $E(Y_{x_0})$, the expected effect of *setting* $X$ to $x_0$ (also called the average causal effect of $X$ on $Y$, denoted $E(Y \mid do(x_0))$ or, generically, $E(Y \mid do(x_0))$), we solve Equation 15.3 for $Y$ in terms of the exogenous variables, yielding $Y_{x_0} = f_Y(x_0, u)$, and average over $U$ and $V$. It is easy to show that in this simple system, the answer can be obtained without knowing the form of the function $f_Y(x, u)$ or the distribution $P(u)$. The answer is given by:

$$E(Y_{x_0}) = E(Y \mid do(X = x_0)) = E(Y \mid x_0)$$

which is estimable from the observed distribution $P(x, y, z)$. This result hinges on the assumption that $W$, $V$, and $U$ are mutually independent and on the topology of the graph (e.g., that there is no direct arrow from $Z$ to $Y$).

In general, it can be shown (Pearl, 2009a, Chapter 3) that, whenever the graph is Markovian (i.e., acyclic with independent exogenous variables) the postinterventional distribution $P(Y = y \mid do(X = x))$ is given by the following expression:

$$P(Y = y \mid do(X = x)) = \sum_t P(y \mid t, x)P(t) \tag{15.6}$$

where $T$ is the set of direct causes of $X$ (also called "parents") in the graph. Again, we see that all factors on the right side are estimable from the distribution $P$ of observed variables, and hence the counterfactual probability $P(Y_x = y)$ is estimable with mere partial knowledge of the generating process—the topology of the graph and independence of the exogenous variable are all that is needed.

When some variables in the graph (e.g., the parents of $X$) are unobserved, we may not be able to estimate (or "identify" as it is called) the postintervention distribution $P(y \mid do(x))$ by simple conditioning, and more sophisticated methods would be required. Likewise, when the query of interest involves several hypothetical worlds simultaneously, for example, $P(Y_x = y \mid Y_{x'} = y')$, the Markovian assumption may not suffice for identification and additional assumptions, touching on the form of the data-generating functions (e.g., monotonicity) may need to be invoked. These issues will be discussed in the "Confounding and Causal Effect Estimation" and "An Example: Mediation, Direct and Indirect Effects" sections.

This interpretation of counterfactuals, cast as solutions to modified systems of equations, provides the conceptual and formal link between structural equation models, used in economics and social science, and the Neyman–Rubin potential–outcome framework to be discussed in the "Relation to Potential Outcomes and the Demystification of 'Ignorability'" section. But first we discuss two longstanding problems that have been completely resolved in purely graphical terms, without delving into algebraic techniques.

### Confounding and Causal Effect Estimation

Although good statisticians have always known that the elucidation of causal relationships from observational studies must be shaped by assumptions about how the data were generated, the relative roles of assumptions and data, and ways of using those assumptions to eliminate confounding bias, have been a subject of much controversy.[8] The structural framework of the "Structural Equations as Oracles for Causes and Counterfactural" section puts these controversies to rest.

## Covariate Selection: The Back-Door Criterion

Consider an observational study where we wish to find the effect of $X$ on $Y$, for example, treatment on response, and assume that the factors deemed relevant to the problem are structured as in Figure 15.3—some are affecting the response; some are affecting the treatment; and some are affecting both treatment and response.

Some of these factors may be unmeasurable, such as genetic trait or lifestyle; others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment, namely, that if we compare treated versus untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of "deconfounding" factors is called a "sufficient set" or a set "admissible for adjustment." The problem of defining a sufficient set, let alone finding one, has baffled epidemiologists and social scientists for decades (see Greenland, Pearl, & Robins, 1999; Pearl, 1998, 2003, for review).

[8] A recent flare-up of this controversy can be found in Pearl (2009c, 2009d, 2010c) and Rubin (2009), which demonstrate the difficulties statisticians encounter in articulating causal assumptions and typical mistakes that arise from pursuing causal analysis within the structure-less "missing data" paradigm.

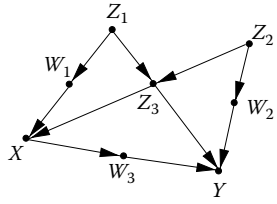



**FIGURE 15.3**
Graphical model illustrating the back-door criterion. Error terms are not shown explicitly.

The following criterion, named "back-door" in Pearl (1993), settles this problem by providing a graphical method of selecting a sufficient set of factors for adjustment. It states that a set $S$ is admissible for adjustment if two conditions hold:

1. No element of $S$ is a descendant of $X$.
2. The elements of $S$ "block" all "back-door" paths from $X$ to $Y$, namely, all paths that end with an arrow pointing to $X$.[9]

Based on this criterion we see, for example, that in the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, and $\{W_2, Z_3\}$, each is sufficient for adjustment because each blocks all back-door paths between $X$ and $Y$. The set $\{Z_3\}$, however, is not sufficient for adjustment because, as explained in footnote 3, it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The implication of finding a sufficient set $S$ is that stratifying on $S$ is guaranteed to remove all confounding bias relative to the causal effect of $X$ on $Y$. In other words, it renders the causal effect of $X$ on $Y$ estimable, via

$$P(Y = y \mid do(X = x)) = \sum_s P(Y = y \mid X = x, S = s)P(S = s). \tag{15.7}$$

Because all factors on the right side of the equation are estimable (e.g., by regression) from the preinterventional data, the causal effect can likewise be estimated from such data without bias.

The back-door criterion allows us to write Equation 15.7 directly, after selecting a sufficient set $S$ from the diagram, without resorting to any algebraic manipulation. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from

[9] In this criterion, a set $S$ of nodes is said to *block* a path $p$ if either (a) $p$ contains at least one arrow-emitting node that is in $S$, or (b) $p$ contains at least one collision node that is outside $S$ and has no descendant in $S$. (See Pearl, 2000b, pp. 16–17, 335–337.)

judging whether "$X$ is conditionally ignorable given $S$," a formidable mental task required in the potential–outcome framework (Rosenbaum & Rubin, 1983). The criterion also enables the analyst to search for an optimal set of covariates—namely, a set $S$ that minimizes measurement cost or sampling variability (Tian, Paz, & Pearl, 1998). A complete identification condition, including models with no sufficient sets (e.g., Figure 15.3, assuming that $X$, $Y$, and $W_3$ are the only measured variables) is given in Shpitser and Pearl (2006).

Another problem that has a simple graphical solution is to determine whether adjustment for two sets of covariates would result in the same confounding bias (Pearl & Paz, 2010). This criterion allows one to assess, before taking any measurement, whether two candidate sets of covariates, differing substantially in dimensionality, measurement error, cost, or sample variability, are equally valuable in their bias-reduction potential.

### Counterfactual Analysis in Structural Models

Not all questions of causal character can be encoded in $P(y \mid do(x))$-type expressions, in much the same way that not all causal questions can be answered from experimental studies. For example, questions of attribution (e.g., I took an aspirin and my headache is gone; was it *due* to the aspirin?) or of susceptibility (e.g., I am a healthy nonsmoker; would I be as healthy had I been a smoker?) cannot be answered from experimental studies, and naturally, these kind of questions cannot be expressed in $P(y) \mid do(x))$ notation.[10] To answer such questions, a probabilistic analysis of counterfactuals is required, one dedicated to the relation "$Y$ would be $y$ had $X$ been $x$ in situation $\boldsymbol{U} = \boldsymbol{u}$," denoted $Y_x(\boldsymbol{u}) = y$.

As noted in the "Structural Equations as Oracles for Causes and Counterfacturals" section, the structural definition of counterfactuals involves modified models, like $M_{x_0}$ of Equation 15.3, formed by the intervention $do(X = x_0)$ (Figure 15.2b). Denote the solution of $Y$ in model $M_x$ by the symbol $Y_{M_x}(u)$; the formal definition of the counterfactual $Y_x(u)$ in a structural causal model is given by (Blake & Pearl, 1994; Pearl, 2009a, p. 98):

$$Y_x(\boldsymbol{u}) \stackrel{\Delta}{=} Y_{M_x}(\boldsymbol{u}). \tag{15.8}$$

[10] The reason for this fundamental limitation is that no death case can be tested twice, with and without treatment. For example, if we measure equal proportions of deaths in the treatment and control groups, we cannot tell how many death cases are attributable to the treatment itself; it is possible that many of those who died under treatment would be alive if untreated and, simultaneously, many of those who survived with treatment would have died if not treated.

The quantity $Y_x(\boldsymbol{u})$ can be given experimental interpretation; it stands for the way an individual with characteristics $(\boldsymbol{u})$ would respond had the treatment been $x$, rather than the treatment $x = f_X(\boldsymbol{u})$ received by that individual. In our example, because $Y$ does not depend on $v$ and $w$, we can write: $Y_{x_0}(\boldsymbol{u}) = f_Y(x_0, u)$. Clearly, the distribution $P(u, v, w)$ induces a well-defined probability on the counterfactual event $Y_{x_0} = y$, as well as on joint counterfactual events, such as "$Y_{x_0} = y$ AND $Y_{x_1} = y'$," which are, in principle, unobservable if $x_0 \neq x_1$. Thus, to answer attributional questions, such as whether $Y$ would be $y_1$ if $X$ were $x_1$, given that in fact $Y$ is $y_0$ and $X$ is $x_0$, we need to compute the conditional probability $P(Y_{x_1} = y_1 \mid Y = y_0, X = x_0)$, which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model. For example, assuming a linear equation for $Y$ (as in Figure 15.1),

$$y = \beta x + u,$$

the conditions $Y = y_0$ and $X = x_0$ yield $V = x_0$ and $U = y_0 - \beta x_0$, and we can conclude that, with probability one, $Y_{x_1}$ must take on the value: $Y_{x_1} = \beta x_1 + U = \beta(x_1 - x_0) + y_0$. In other words, if $X$ were $x_1$ instead of $x_0$, $Y$ would increase by $\beta$ times the difference $(x_1 - x_0)$. In nonlinear systems, the result would also depend on the distribution of $U$, and, for that reason, attributional queries are generally not identifiable in nonparametric models (Pearl, 2009a, Chapter 9).

In general, if $x$ and $x'$ are incompatible, then $Y_x$ and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "$Y$ would be $y$ if $X = x$ and $Y$ would be $y'$ if $X = x'$." Such concerns have been a source of objections to treating counterfactuals as jointly distributed random variables (Dawid, 2000). The definition of $Y_x$ and $Y_{x'}$ in terms of two distinct submodels neutralizes these objections (Pearl, 2009a, p. 206) because the contradictory joint statement is mapped into an ordinary event (among the background variables) that satisfies both statements simultaneously, each in its own distinct submodel; such events have well-defined probabilities.

The structural interpretation of counterfactuals (Equation 15.8) also provides the conceptual and formal basis for the Neyman–Rubin potential–outcome framework, an approach that takes a controlled randomized trial as its starting paradigm, assuming that nothing is known to the experimenter about the science behind the data. This "black box" approach was developed by statisticians who found it difficult to cross the two mental barriers discussed in the "Two Mental Barriers" section. The next section establishes the precise relationship between the structural and potential–outcome paradigms, and outlines how the latter can benefit from the richer representational power of the former.

### Relation to Potential Outcomes and the Demystification of "Ignorability"

The primitive object of analysis in the potential–outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that outcome $Y$ would obtain in experimental unit $u$, had treatment $X$ been $x$" (Neyman, 1923; Rubin, 1974). Here, *unit* may stand for an individual patient, an experimental subject, or an agricultural plot. In the "Counterfactual Analysis in Structural Models" section, we saw (Equation 15.8) that this counterfactual entity has a natural interpretation in structural equations as the solution for $Y$ in a modified system of equation, where *unit* is interpreted as vector $\boldsymbol{u}$ of background factors that characterize an experimental unit. Thus, each structural equation model carries a collection of assumptions about the behavior of hypothetical units, and these assumptions permit us to derive the counterfactual quantities of interest. In the potential–outcome framework, however, no equations are available for guidance, and $Y_x(u)$ is taken as primitive, that is, an undefined quantity in terms of which other quantities are defined—not a quantity that can be derived from some model. In this sense, the structural interpretation of $Y_x(u)$ provides the formal basis for the *potential outcome* approach; the formation of the submodel $M_x$ explicates mathematically how the hypothetical condition "had $X$ been $x$" could be realized and what the logical consequences are of such a condition.

The distinct characteristic of the potential outcome approach is that, although investigators must think and communicate in terms of undefined, hypothetical quantities such as $Y_x(u)$, the analysis itself is conducted almost entirely within the axiomatic framework of probability theory. This is accomplished by treating the new hypothetical entities $Y_x$ as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence.

Naturally, these hypothetical entities are not entirely whimsy. They are assumed to be connected to observed variables via consistency constraints (Robins, 1986), such as

$$X = x \Rightarrow Y_x = Y, \tag{15.9}$$

which states that, for every $u$, if the actual value of $X$ turns out to be $x$, then the value that $Y$ would take on if "$X$ *were* $x$" is equal to the actual value of $Y$. For example, a person who chose treatment $x$ and recovered would also have recovered if given treatment $x$ by design. Whether additional constraints should tie the observables to the unobservables is not a question that can be answered in the potential–outcome framework, which lacks an underlying model.

The main conceptual difference between the two approaches is that whereas the structural approach views the intervention $do(x)$ as an operation that changes the distribution but keeps the variables the same, the potential–outcome approach views the variable $Y$ under $do(x)$ to be a different variable, $Y_x$, loosely connected to $Y$ through relations such as Equation 15.9, but remaining unobserved whenever $X \neq x$. The problem of inferring probabilistic properties of $Y_x$ then becomes one of "missing data," for which estimation techniques have been developed in the statistical literature.

Pearl (2009a, Chapter 7) shows, using the structural interpretation of $Y_x(u)$ (Equation 15.8), that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like Equation 15.9 are automatically satisfied in the structural interpretation, and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \text{ for all } y, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{15.10}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \text{ for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z. \tag{15.11}$$

Equation 15.10 ensures that the intervention $do(Y = y)$ results in the condition $Y = y$, regardless of concurrent interventions, say $do(Z = z)$, that may be applied to variables other than $Y$. Equation 15.11 generalizes Equation 15.9 to cases where $Z$ is held fixed, at $z$.

### Problem Formulation and the Demystification of "Ignorability"

The main drawback of this black box approach surfaces in the phase where a researcher begins to articulate the "science" or "causal assumptions" behind the problem at hand. Such knowledge, as we have seen in the "Two Mental Barriers" section, must be articulated at the onset of every problem in causal analysis—causal conclusions are only as valid as the causal assumptions on which they rest.

To communicate scientific knowledge, the potential–outcome analyst must express causal assumptions in the form of assertions involving counterfactual variables. For example, in our example of Figure 15.2a, to communicate the understanding that $Z$ is randomized (hence independent of $V$ and $U$), the potential–outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$.[11] To further formulate the understanding that $Z$

[11] The notation $Y \perp\!\!\!\perp X \mid Z$ stands for the conditional independence relationship $P(Y = y, X = x \mid Z = z) = P(Y = y \mid Z = z)\, P(X = x \mid Z = z)$ (Dawid, 1979).

does not affect $Y$ directly, except through $X$, the analyst would write a so-called, "exclusion restriction": $Y_{xz} = Y_x$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest; in other cases, only bounds on the solution can be obtained. For example, if one can plausibly assume that a set $Z$ of covariates satisfies the relation:

$$Y_x \perp\!\!\!\perp X \mid Z \tag{15.12}$$

(assumption that was termed *conditional ignorability* by Rosenbaum & Rubin, 1983), then the causal effect $P(Y_x = y)$ can readily be evaluated to yield:

$$\begin{aligned} P(Y_x = y) &= \sum_z P(Y_x = y \mid z)P(z) \\ &= \sum_z P(Y_x = y \mid x, y)P(z) \quad \text{(using (12))} \\ &= \sum_z P(Y_x = y \mid x, z)P(z) \quad \text{(using (9))} \\ &= \sum_z P(y \mid x, z)P(z). \end{aligned} \tag{15.13}$$

The last expression contains no counterfactual quantities and coincides precisely with the standard covariate-adjustment formula of Equation 15.7.

We see that the assumption of conditional ignorability (Equation 15.12) qualifies $Z$ as a sufficient covariate for adjustment; indeed, one can show formally (Pearl, 2009a, pp. 98–102, 341–343) that Equation 15.12 is entailed by the "back-door" criterion of the "Confounding and Causal Effect Estimation" section.

The derivation above may explain why the potential outcome approach appeals to conservative statisticians; instead of constructing new vocabulary (e.g., arrows), new operators ($do(x)$), and new logic for causal analysis, almost all mathematical operations in this framework are conducted within the safe confines of probability calculus. Save for an occasional application of the consistency rule, Equation 15.11 or Equation 15.9, the analyst may forget that $Y_x$ stands for a counterfactual quantity—it is treated as any other random variable, and the entire derivation follows the course of routine probability exercises.

However, this mathematical orthodoxy exacts a high cost at the critical stage where causal assumptions are formulated. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability (Equation 15.12), the key to the derivation of Equation 15.15, holds in any familiar situation, say, in the experimental setup of

Figure 15.2a. This assumption reads: "the value that $Y$ would obtain had $X$ been $x$, is independent of $X$, given $Z$." Even the most experienced potential–outcome expert would be unable to discern whether any subset $Z$ of covariates in Figure 15.3 would satisfy this conditional independence condition.[12] Likewise, to convey the structure of the chain $X \rightarrow W_3 \rightarrow Y$ (Figure 15.3) in the language of potential–outcome, one would need to write the cryptic expression: $W_{3_x} \perp\!\!\!\perp \{Y_{w_3}, X\}$, read: "the value that $W_3$ would obtain had $X$ been $x$ is independent of the value that $Y$ would obtain had $W_3$ been $w_3$ jointly with the value of $X$." Such assumptions are cast in a language so far removed from ordinary understanding of cause and effect that, for all practical purposes, they cannot be comprehended or ascertained by ordinary mortals. As a result, researchers in the graphless potential–outcome camp rarely use "conditional ignorability" (Equation 15.12) to guide the choice of covariates; they view this condition as a hoped-for miracle of nature rather than a target to be achieved by reasoned design.[13]

Having translated "ignorability" into a simple condition (i.e., back door) in a graphical model permits researchers to understand what conditions covariates must fulfill before they eliminate bias, what to watch for and what to think about when covariates are selected, and what experiments we can do to test, at least partially, if we have the knowledge needed for covariate selection.

Aside from offering no guidance in covariate selection, formulating a problem in the potential–outcome language encounters three additional hurdles. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated, whether the judgments articulated are *redundant*, or whether those judgments are *self-consistent*. The need to express, defend, and manage formidable counterfactual relationships of this type explains the slow acceptance of causal analysis among health scientists and statisticians, and why economists and social scientists continue to use structural equation models instead of the potential–outcome alternatives advocated in Angrist, Imbens, and Rubin (1996), Holland (1988), and Sobel (1998).

On the other hand, the algebraic machinery offered by the counterfactual notation, $Y_x(u)$, once a problem is properly formulated, can be

[12] Inquisitive readers are invited to guess whether $X_z \perp\!\!\!\perp Z \mid Y$ holds in Figure 15.2a.

[13] The opaqueness of counterfactual independencies explains why many researchers within the potential–outcome camp are unaware of the fact that adding a covariate to the analysis (e.g., $Z_3$ in Figure 15.3) may increase confounding bias. Paul Rosenbaum, for example, writes: "There is little or no reason to avoid adjustment for a variable describing subjects before treatment" (Rosenbaum, 2002, p. 76). Rubin (2009) goes as far as stating that refraining from conditioning on an available measurement is "nonscientific ad hockery" because it goes against the tenets of Bayesian philosophy. (See Pearl, 2009c, 2009d, 2010c, for a discussion of this fallacy.)

extremely powerful in refining assumptions (Angrist et al., 1996), deriving consistent estimands (Robins, 1986), bounding probabilities of necessary and sufficient causation (Tian & Pearl, 2000), and combining data from experimental and nonexperimental studies (Pearl, 2009a). Pearl (2009a, p. 232) presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams, translating these assumptions into counterfactual notation, performing the mathematics in the algebraic language of counterfactuals (using Equations 15.9, 15.10, and 15.11), and, finally, interpreting the result in plain causal language. The "An Example: Mediation, Direct and Indirect Effects" section illustrates such symbiosis.

## Methodological Dictates and Ethical Considerations

The structural theory described in the previous sections dictates a principled methodology that eliminates the confusion between causal and statistical interpretations of study results, as well as the ethical dilemmas that this confusion tends to spawn. The methodology dictates that every investigation involving causal relationships (and this entails the vast majority of empirical studies in the social and behavioral sciences) should be structured along the following four-step process:[14]

1. **Define:** Express the target quantity $Q$ as a function $Q(M)$ that can be computed from any model $M$, regardless of how realistic it is.
2. **Assume:** Formulate causal assumptions using ordinary scientific language, and represent their structural part in graphical form.
3. **Identify:** Determine whether the target quantity is identifiable (i.e., expressible as distributions).
4. **Estimate:** Estimate the target quantity if it is identifiable, or approximate it if it is not.

### Defining the Target Quantity

The definitional phase is the most neglected step in current practice of quantitative analysis. The structural modeling approach insists on defining the target quantity, be it "causal effect," "program effectiveness," "mediated effect," "effect on the treated," or "probability of causation" before specifying any aspect of the model, without making functional or

[14] Pearl (2010a) identifies five steps, which include model testing.

distributional assumptions, before choosing a method of estimation, and before seeing any data.

The investigator should view this definition as an *algorithm* that receives a model $M$ as an input and delivers the desired quantity $Q(M)$ as the output. Surely, such an algorithm should not be tailored to any aspect of the input $M$; it should be general and ready to accommodate any conceivable model $M$ whatsoever. Moreover, the investigator should imagine that the input $M$ is a completely specified model, with all the functions $f_X, f_Y, \ldots$ and all the $U$ variables (or their associated probabilities) given precisely. This is the hardest step for statistically trained investigators to make; knowing in advance that such model details will never be estimable from the data, the definition of $Q(M)$ appears like a futile exercise in fantasyland—it is not.

For example, the formal definition of the causal effect $P(y \mid do(x))$, as given in Equation 15.4, is universally applicable to all models, parametric and nonparametric, through the formation of a submodel $M_x$. By defining causal effect procedurally, thus divorcing it from its traditional parametric representation, the structural theory avoids the many pitfalls and confusions that have plagued the interpretation of structural and regressional parameters for the past half century.[15]

## Explicating Causal Assumptions

This is the second most neglected step in causal analysis. In the past, the difficulty has been the lack of language suitable for articulating causal assumptions, which, aside from impeding investigators from explicating assumptions, also inhibited them from giving causal interpretations to their findings.

Structural equation models, in their counterfactual reading, have settled this difficulty. Today we understand that the versatility and natural appeal of structural equations stem from the fact that they permit investigators to communicate causal assumptions formally and in the very same vocabulary that scientific knowledge is stored.

Unfortunately, however, this understanding is not shared by all causal analysts; some analysts vehemently resist the resurrection of structural models and insist instead on articulating causal assumptions exclusively

[15] Note that $b$ in Equation 15.1, the incremental causal effect of $X$ on $Y$, is defined procedurally by

$$\beta \stackrel{\Delta}{=} E(Y \mid do(x_0 + 1)) - E(Y \mid do(x_0)) = \frac{\partial}{\partial x} E(Y_x).$$

Naturally, all attempts to give $b$ statistical interpretation have ended in frustrations (Holland, 1988; Wermuth, 1992; Wermuth & Cox, 1993; Whittaker, 1990), some persisting well into the 21st century (Sobel, 2008).

in the unnatural (although formally equivalent) language of potential outcomes, ignorability, treatment assignment, and other metaphors borrowed from clinical trials. This assault on structural modeling is perhaps more dangerous than the causal–associational confusion because it is riding on a halo of exclusive ownership to scientific principles and, while welcoming causation, uproots it away from its natural habitat.

Early birds of this exclusivist attitude have already infiltrated the American Psychological Association's (APA) guidelines (Wilkinson & the Task Force on Statistical Inference, 1999), where we can read passages such as: "The crucial idea is to set up the causal inference problem as one of missing data," (item 72) or "If a problem of causal inference cannot be formulated in this manner (as the comparison of potential outcomes under different treatment assignments), it is not a problem of inference for causal effects, and the use of 'causal' should be avoided," (item 73) or, even more bluntly, "The underlying assumptions needed to justify any causal conclusions should be carefully and explicitly argued, not in terms of technical properties like 'uncorrelated error terms,' but in terms of real world properties, such as how the units received the different treatments" (item 74).

The methodology expounded in this article testifies against such restrictions. It demonstrates a viable and principled formalism based on traditional structural equations paradigm, which stands diametrically opposed to the "missing data" paradigm. It renders the vocabulary of "treatment assignment" stifling and irrelevant (e.g., there is no "treatment assignment" in sex discrimination cases). Most importantly, it strongly prefers the use of "uncorrelated error terms" (or "omitted factors") over its "strong ignorability" alternative, which even experts admit cannot be used (and has not been used) to reason about underlying assumptions.

In short, the APA's guidelines should be vastly more inclusive and borrow strength from multiple approaches. The next section demonstrates the benefit of a symbiotic, graphical–structural–counterfactual approach to deal with the problem of mediation, or effect decomposition.

## An Example: Mediation, Direct and Indirect Effects

### Direct Versus Total Effects

The causal effect we have analyzed so far, $P(y \mid do(x))$, measures the *total* effect of a variable (or a set of variables) $X$ on a response variable $Y$. In many cases, this quantity does not adequately represent the target of

investigation, and attention is focused instead on the direct effect of $X$ on $Y$. The term *direct effect* is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of $Y$ to changes in $X$ while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from $X$ to $Y$ with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring is a target of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

From a policy-making viewpoint, an investigator may be interested in decomposing effects to quantify the extent to which racial salary disparity is the result of educational disparity, or, taking a health care example, the extent to which sensitivity to a given exposure can be reduced by eliminating sensitivity to an intermediate factor, standing between exposure and outcome. Another example concerns the identification of neural pathways in the brain or the structural features of protein-signaling networks in molecular biology (Brent & Lok, 2005). Here, the decomposition of effects into their direct and indirect components carries theoretical scientific importance because it tells us "how nature works" and therefore enables us to predict behavior under a rich variety of conditions.

Yet despite its ubiquity, the analysis of mediation has long been a thorny issue in the social and behavioral sciences (Baron & Kenny, 1986; Judd & Kenny, 1981; MacKinnon, Fairchild, & Fritz, 2007a; Muller, Judd, & Yzerbyt, 2005; Shrout & Bolger, 2002) primarily because structural equation modeling in those sciences was deeply entrenched in linear analysis, where the distinction between causal parameters and their regressional interpretations can easily be conflated. As demands grew to tackle problems involving binary and categorical variables, researchers could no longer define direct and indirect effects in terms of structural or regressional coefficients, and all attempts to extend the linear paradigms of effect decomposition to nonlinear systems produced distorted results (MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007b). These difficulties have accentuated the need to redefine and derive *causal effects* from first principles, uncommitted to distributional assumptions, or a particular parametric form of the equations. The structural methodology presented in this chapter adheres to this philosophy, and it has produced indeed a principled solution to the mediation problem, based on the counterfactual reading of structural equations (Equation 15.8). The following subsections summarize the method and its solution.

### Controlled Direct Effects

A major impediment to progress in mediation analysis has been the lack of notational facility for expressing the key notion of "holding the mediating variables fixed" in the definition of direct effect. Clearly, this notion must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, regression conditioning, matching, or adjustment. For example, consider the simple mediation models of Figure 15.4, where the error terms (not shown explicitly) are assumed to be independent. It will not be sufficient to measure the association between gender ($X$) and hiring ($Y$) for a given level of qualification ($Z$) (see Figure 15.4b) because, by conditioning on the mediator $Z$, we create spurious associations between $X$ and $Y$ through $W_2$, even when there is no direct effect of $X$ on $Y$ (Pearl, 1998).

Using the $do(x)$ notation enables us to correctly express the notion of "holding $Z$ fixed" and formulate a simple definition of the *controlled direct effect* (CDE) of the transition from $X = x$ to $X = x'$:

$$CDE \stackrel{\Delta}{=} E(Y \mid do(x'), do(z)) - E(Y \mid do(x), do(z)).$$

Or, equivalently, using counterfactual notation:

$$CDE \stackrel{\Delta}{=} E(Y_{x'z}) - E(Y_{xz})$$

where $Z$ is the set of all mediating variables. The readers can easily verify that, in linear systems, the controlled direct effect reduces to the path coefficient of the link $X \rightarrow Y$ (see footnote 14) regardless of whether confounders are present (as in Figure 15.4b) and regardless of whether the error terms are correlated.

This separates the task of definition from that of identification, as demanded by the "Defining the Target Quantity" section. The identification

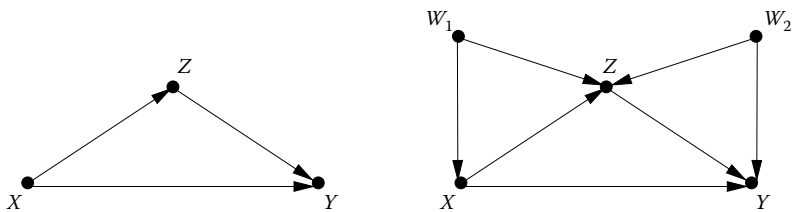



**FIGURE 15.4**
(a) A generic model depicting mediation through $Z$ with no confounders. (b) A mediation model with two confounders, $W_1$ and $W_2$.

of *CDE* would depend, of course, on whether confounders are present and whether they can be neutralized by adjustment, but these do not alter its definition. Graphical identification conditions for expressions of the type $E(Y \mid do(x), do(z_1), do(z_2), \ldots, do(z_k))$ in the presence of unmeasured confounders were derived by Pearl and Robins (1995) (see Pearl, 2009a, Chapter 4) and invoke sequential application of the back-door conditions discussed in the "Confounding and Causal Effect Estimation" section.

### Natural Direct Effects

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from $X$ to $Y$; therefore, the direct effect is independent of the values at which we hold $Z$. In nonlinear systems, those values would, in general, modify the effect of $X$ on $Y$ and thus should be chosen carefully to represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high $z$) and females for low-paying jobs (low $z$).

When the direct effect is sensitive to the levels at which we hold $Z$, it is often more meaningful to define the direct effect relative to some "natural" baseline level that may vary from individual to individual and represents the level of $Z$ just before the change in $X$. Conceptually, we can define the natural direct effect $DE_{x, x'}(Y)$ as the expected change in $Y$ induced by changing $X$ from $x$ to $x'$ while keeping all mediating factors constant at whatever value they *would have obtained* under $do(x)$. This hypothetical change, which Robins and Greenland (1992) conceived and called "pure" and Pearl (2001) formalized and analyzed under the rubric "natural," mirrors what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had been the same" (*Carson v. Bethlehem Steel Corp.*, 1996).

Extending the subscript notation to express nested counterfactuals, Pearl (2001) gave a formal definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{15.14}$$

Here $Y_{x',Z_x}$ represents the value that $Y$ would attain under the operation of setting $X$ to $x'$ and simultaneously setting $Z$ to whatever value it would have obtained under the setting $X = x$. We see that $DE_{x, x'}(Y)$, the natural direct effect of the transition from $x$ to $x'$, involves probabilities of *nested counterfactuals* and cannot be written in terms of the $do(x)$ operator. Therefore, the natural direct effect cannot in general be identified or estimated, even with the help of ideal, controlled experiments (see footnote 10)—a point emphasized in Robins and Greenland (1992). However, aided by Equation

15.8 and the notational power of nested counterfactuals, Pearl (2001) was nevertheless able to show that if certain assumptions of "no confounding" are deemed valid the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y \mid do(x', z)) - E(Y \mid do(x, z))]P(z \mid do(x)). \quad (15.15)$$

The intuition is simple; the natural direct effect is the weighted average of the controlled direct effect, using the causal effect $P(z \mid do(x))$ as a weighting function.

One condition for the validity of Equation 15.17 is that $Z_x \perp\!\!\!\perp Y_{x',z} \mid W$ holds for some set $W$ of measured covariates. This technical condition in itself, like the ignorability condition of Equation 15.12, is close to meaningless for most investigators because it is not phrased in terms of realized variables. The structural interpretation of counterfactuals (Equation 15.8) can be invoked at this point to unveil the graphical interpretation of this condition. It states that $W$ should be admissible (i.e., satisfy the back-door condition) relative the path(s) from $Z$ to $Y$. This condition, satisfied by $W_2$ in Figure 15.4b, is readily comprehended by empirical researchers, and the task of selecting such measurements, $W$, can then be guided by the available scientific knowledge. Additional graphical and counterfactual conditions for identification are derived in Pearl (2001), Petersen, Sinisi, and van der Laan (2006), and Imai, Keele and Yamamoto (2008).

In particular, it was shown (Pearl, 2001) that Equation 15.15 is both valid and identifiable in Markovian models (i.e., no unobserved confounders) where each term on the right can be reduced to a "*do*-free" expression using Equation 15.6 or Equation 15.7 and then estimated by regression.

For example, for the model in Figure 15.4b, Equation 15.15 reads:

$$DE_{x,x'}(Y) = \sum_z \sum_{w_2} P(w_2)[E(Y \mid x', z, w_2)) - E(Y \mid x, z, w_2))] \sum_{w_1} P(z \mid x, w_1)P(w_1). \quad (15.16)$$

However, for the confounding-free model of Figure 15.4a, we have:

$$DE_{x,x'}(Y) = \sum_z [E(Y \mid x', z) - E(Y \mid x, z)]P(z \mid x). \quad (15.17)$$

Both Equations 15.16 and 15.17 can easily be estimated by a two-step regression.

### Natural Indirect Effects

Remarkably, the definition of the natural direct effect (Equation 15.14) can be turned around and provides an operational definition for the *indirect*

*effect*—a concept shrouded in mystery and controversy because it is impossible, using standard intervention, to disable the direct link from $X$ to $Y$ so as to let $X$ influence $Y$ solely via indirect paths.

The *natural indirect effect* (*IE*) of the transition from $x$ to $x'$ is defined as the expected change in $Y$ affected by holding $X$ constant, at $X = x$, and changing $Z$ to whatever value it would have attained had $X$ been set to $X = x'$. Formally, this reads (Pearl, 2001):

$$IE_{x,x'}(Y) \stackrel{\Delta}{=} E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{15.18}$$

which is almost identical to the direct effect (Equation 15.14) save for exchanging $x$ and $x'$ in the first term.

Indeed, it can be shown that, in general, the total effect (*TE*) of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \stackrel{\Delta}{=} E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{15.19}$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula:

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{15.20}$$

Because each term above is based on an independent operational definition, this equality constitutes a formal justification for the additive formula used routinely in linear systems.

Note that, although it cannot be expressed in *do*-notation, the indirect effect has clear policy-making implications. For example, in the hiring discrimination context, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender dependent.

More generally, a policy maker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of $X$ on $Y$ through a selected set of paths (Avin, Shpitser, & Pearl, 2005).

In all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Therefore, Pearl (2001) has suggested that *signal sensing* is more fundamental to the notion of causation than *manipulation*, the latter being but a crude way of stimulating the

former in experimental setup. The mantra "No causation without manipulation" must be rejected (see Pearl, 2009a, Section 11.4.5).

It is remarkable that counterfactual quantities like *DE* and *IE* that could not be expressed in terms of $do(x)$ operators, and therefore appear void of empirical content, can, under certain conditions, be estimated from empirical studies and serve to guide policies. Awareness of this potential should embolden researchers to go through the definitional step of the study and freely articulate the target quantity $Q(M)$ in the language of science, that is, counterfactuals, despite the seemingly speculative nature of each assumption in the model (Pearl, 2000).

### The Mediation Formula: A Simple Solution to a Thorny Problem

This subsection demonstrates how the solution provided in Equations 15.17 and 15.20 can be applied to practical problems of assessing mediation effects in nonlinear models. We will use the simple mediation model of Figure 15.4a, where all error terms (not shown explicitly) are assumed to be mutually independent, with the understanding that adjustment for appropriate sets of covariates $W$ may be necessary to achieve this independence and that integrals should replace summations when dealing with continuous variables (Imai et al., 2008).

Combining Equations 15.17, 15.19, and 15.20, the expression for the indirect effect, *IE*, becomes:

$$IE_{xx'}(Y) = \sum_z E(Y \mid x, z)[P(z \mid x') - P(z \mid x)], \tag{15.21}$$

which provides a general and easy-to-use formula for mediation effects, applicable to any nonlinear system, any distribution (of $U$), and any type of variables. Moreover, the formula is readily estimable by regression, making no assumption whatsoever about the parametric form of the underlying process. Owed to its generality and ubiquity, I have referred to this expression as the "Mediation Formula" (Pearl, 2009b).

The Mediation Formula represents the average increase in the outcome $Y$ that the transition from $X = x$ to $X = x'$ is expected to produce absent any direct effect of $X$ on $Y$. Although based on solid causal principles, it embodies no causal assumption other than the generic mediation structure of Figure 15.4a. When the outcome $Y$ is binary (e.g., recovery, or hiring) the ratio $(1 - IE/TE)$ represents the fraction of responding individuals who owe their response to direct paths, whereas $(1 - IE/TE)$ represents the fraction who owe their response to $Z$-mediated paths.

The Mediation Formula tells us that *IE* depends only on the expectation of the counterfactual $Y_{xz}$, not on its functional form $f_Y(x, z, u_Y)$ or its distribution $P(Y_{xz} = y)$. Therefore, it calls for a two-step regression that, in

principle, can be performed nonparametrically. In the first step, we regress $Y$ on $X$ and $Z$ and obtain the estimate:

$$g(x,z) = E(Y \mid x,z)$$

for every ($x$, $z$) cell. In the second step, we estimate the expectation of $g(x, z)$ conditional on $X = x'$ and $X = x$, respectively, and take the difference:

$$IE_{x,x'}(Y) = E_{Z|x'}(g(x,z)) - E_{Z|x}(g(x,z)).$$

Nonparametric estimation is not always practical. When $Z$ consists of a vector of several mediators, the dimensionality of the problem would prohibit the estimation of $E(Y \mid x, z)$ for every ($x$, $z$) cell, and the need arises to use parametric approximation. We can then choose any convenient parametric form for $E(Y \mid x, z)$ (e.g., linear, logit, probit), estimate the parameters separately (e.g., by regression or maximum likelihood methods), insert the parametric approximation into Equation 15.21, and estimate its two conditional expectations (over $z$) to get the mediated effect (Pearl, 2010b; VanderWeele, 2009).

When applied to linear models, the Mediation Formula yields, of course, the standard product of coefficients. For example, the linear version of Figure 15.4a reads:

$$\begin{aligned} x &= u_X \\ z &= b_x x + u_Z \\ y &= c_x x + c_z z + u_Y. \end{aligned} \tag{15.22}$$

Computing the conditional expectation in Equation 15.21 gives:

$$E(Y \mid x,z) = E(c_x x + c_z z + u_y) = c_x x + c_z z$$

and yields:

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (c_x x + c_z z)[P(z \mid x') - P(z \mid x)] \\ &= c_z[E(Z \mid x') - E(Z \mid x)] && (15.23) \\ &= (x' - x)(c_z b_x) && (15.24) \\ &= (x' - x)(b - c_x) && (15.25) \end{aligned}$$

where $b$ is the total effect coefficient, $b = (E(Y \mid x') - E(Y \mid x)) \,/\, (x' - x) = c_x + c_z b_x$.

Thus, we obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference in two regression coefficients (Equation 15.25) or a product of two regression coefficients (Equation 15.24), with $Y$ regressed on both $X$ and $Z$. However, when extended to nonlinear systems, these two strategies yield conflicting results (MacKinnon & Dwyer, 1993; MacKinnon et al., 2007b), and the question arose as to which strategy should be used in assessing the size of mediated effects (Freedman, Graubard, & Schatzkin, 1992; MacKinnon & Dwyer, 1993; MacKinnon et al., 2007b; Molenberghs et al., 2002). Pearl (2010b) shows that both strategies yield highly distorted results in nonlinear models, even when correct parametric forms are assumed. The reason lies in a violation of step 1 (defining the target quantity) of the "Methodological Dictates and Ethical Considerations" section. Researchers failed to define the causal quantity of interest and were postulating, estimating, and comparing parameters that were related to, yet hardly resembling, *DE* and *IE*. The Mediation Formula captures the correct target quantity and helps researchers cross the nonlinear barrier that has held back the mediation literature for more than half a century. Simple examples using Bernoulli/binary noise, logistic, and probit models are illustrated in Pearl (2010a, 2010b).

In addition to providing causally sound estimates for mediation effects, the Mediation Formula also enables researchers to evaluate analytically the effectiveness of various parametric specifications relative to any assumed model. This type of analytical "sensitivity analysis" has been used extensively in statistics for parameter estimation but could not be applied to mediation analysis because of the absence of an objective target quantity that captures the notion of indirect effect in both linear and nonlinear systems, free of parametric assumptions. The Mediation Formula has removed this barrier (Imai, Keele, & Tingley, 2010; Li, Schneider, & Bennett, 2007).

The derivation of the Mediation Formula (Pearl, 2001) was facilitated by taking seriously the four steps of the structural methodology ("Methodological Dictates and Ethical Considerations" section) together with the graph–counterfactual–structural symbiosis spawned by the structural interpretation of counterfactuals (Equation 15.8). In contrast, when the mediation problem is approached from an exclusivist potential–outcome viewpoint, void of the structural guidance of Equation 15.8, counterintuitive definitions ensue, carrying the label "principal stratification" (Rubin, 2004, 2005), which are at variance with common understanding of direct and indirect effects. For example, the direct effect is definable only in units absent of indirect effects. This means that a grandfather would be deemed to have no direct effect on his grandson's behavior in families where he has had some effect on the father. This precludes from the analysis all typical families, in which a father and a grandfather have simultaneous,

complementary influences on children's upbringing. In linear systems, to take an even sharper example, the direct effect would be undefined whenever indirect paths exist from the cause to its effect. The emergence of such paradoxical conclusions underscores the wisdom, if not necessity, of a symbiotic analysis, in which the counterfactual notation $Y_x(u)$ is governed by its structural definition, Equation 15.8.[16] It also brings into focus the ethical issue of inclusiveness and its role in scientific research and education.

## Conclusion

Statistics is strong in inferring distributional parameters from sample data. Causal inference requires two addition ingredients: a science-friendly language for articulating causal knowledge, and a mathematical machinery for processing that knowledge, combining it with data and drawing new causal conclusions about a phenomenon. This chapter presents nonparametric *structural causal models* as a formal and meaningful language for meeting these challenges, thus easing the ethical tensions that follow the disparity between causal quantities sought by scientists and associational quantities inferred from observational studies. The algebraic component of the structural language coincides with the potential–outcome framework, and its graphical component embraces Wright's method of path diagrams (in its nonparametric version). When unified and synthesized, the two components offer empirical investigators a powerful methodology for causal inference that resolves longstanding problems in the empirical sciences. These include the control of confounding, the evaluation of policies, the analysis of mediation, and the algorithmization of counterfactuals.

In particular, the analysis of mediation demonstrates the benefit of adhering to the methodological principles described. The development of the Mediation Formula (Equations 15.17 and 15.20) has liberated researchers from the blindfolds of parametric thinking and allows them to assess direct and indirect effects for any type of variable, with minimum assumptions regarding the underlying process.[17]

[16] Such symbiosis is now standard in epidemiology research (Hafeman & Schwartz, 2009; Petersen et al., 2006; Robins, 2001; VanderWeele, 2009; VanderWeele & Robins, 2007) and is making its way slowly toward the social and behavioral sciences (e.g., Elwert & Winship, 2010; Morgan & Winship, 2007).

[17] Author note: Portions of this chapter are adapted from Pearl (2009a, 2009b, 2010a). I am grateful to A. T. Panter and Sonya K. Serba for their encouragement and flexibility in the writing of this chapter. This research was supported in part by grants from National Science Foundation (IIS-0535223) and Office of Naval Research (N000-14-09-1-0665).

## References


Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association, 91,* 444–472.

Avin, C., Shpitser, I., & Pearl. J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05* (pp. 357–363). Edinburgh, UK: Morgan-Kaufmann Publishers.

Balke, A., & Pearl, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* (Vol. I, pp. 230–237). Menlo Park, CA: MIT Press.

Balke, A., & Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence 11* (pp. 11–18). San Francisco: Morgan Kaufmann.

Baron, R. M., & Kenny, D. A.(1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Brent, R., & Lok, L. (2005). A fishing buddy for hypothesis generators. *Science, 308,* 523–529.

Cox, D. R. (1958). *The planning of experiments.* New York: John Wiley and Sons.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B, 41,* 1–31.

Dawid, A. P. (2000). Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association, 95,* 407–448.

Elwert, F., & Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 327–336). London: College Publications.

Freedman, L. S., Graubard, B. I., & Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine, 8,* 167–178.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10,* 37–48.

Hafeman, D. M., & Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology, 3,* 838–845.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. In C. Clogg (Ed.), *Sociological Methodology* (pp. 449–484). Washington, DC: American Sociological Association.

Imai, K., Keele, L., & Tingley, D. (2010). *A general approach to causal mediation analysis. Technical report*. Princeton, NJ: Princeton University.

Imai, K., Keele, L., & Yamamoto, T. (2008). *Identification, inference, and sensitivity analysis for causal mediation effects. Technical report*. Princeton, NJ: Princeton University.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.

Li, Y., Schneider, J. A., & Bennett, D. A. (2007). Estimation of the mediation effect with a binary mediator. *Statistics in Medicine, 26,* 3398–3414.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 4,* 144–158.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007a). Mediation analysis. *Annual Review of Psychology, 58,* 593–614.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007b). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials, 4,* 499–513.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., & Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials, 23,* 607–625.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research (analytical methods for social research).* New York: Cambridge University Press.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89,* 852–863.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science, 5,* 465–480.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science, 8,* 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82,* 669–710.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research, 27,* 226–284.

Pearl, J. (2000). Comment on A. P. Dawid's causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 428–431.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco: Morgan Kaufmann.

Pearl, J. (2003). Statistics and causal inference: A review. *Test Journal, 12,* 281–345.

Pearl, J. (2009a). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Pearl, J. (2009b). Causal inference in statistics: An overview. *Statistics Surveys, 3,* 96–146. Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf

Pearl, J. (2009c). Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine, 28,* 1415–1416. Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf

Pearl, J. (2009d). *Myth, confusion, and science in causal analysis. Technical report R-348.* Los Angeles: University of California, Los Angeles. Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf

Pearl, J. (2010a). An introduction to causal inference. *The International Journal of Biostatistics, 6.* doi: 10.2202/1557-4679.1203. Retrieved from http://www.bepress.com/ijb/vol6/iss2/7

Pearl, J. (2010b). *The mediation formula: A guide to learning causal pathways. Technical report TR-363.* Los Angeles: University of California, Los Angeles. Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf

Pearl, J. (2010c). On a class of bias-amoplifying covariates that endanger effect estimates. In P. Grunwald & P. Spirtes (Eds.), *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (pp. 417–432). Corralis, OR: AUAI Press.

Pearl, J., & Paz, A. (2010). Confounding equivalence in observational studies. In P. Grunwald & P. Spirtes (Eds.), *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (pp. 433–441). Corralis, OR: AUAI Press. Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r343.pdf

Pearl, J., & Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence 11* (pp. 444–453). San Francisco: Morgan Kaufmann.

Petersen, M. L., Sinisi, S. E. & van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology, 17,* 276–284.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—applications to control of the healthy workers survivor effect. *Mathematical Modeling, 7,* 1393–1512.

Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology, 12,* 313–320.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3,* 143–155.

Rosenbaum, P., & Rubin, D. (1983).The central role of propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31,* 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Rubin, D. B. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine, 28,* 1420–1423.

Shpitser, I., & Pearl, J. (2006). Identification of conditional interventional distributions. In R. Dechter & T. S. Richardson (Eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (pp. 437–444). Corvallis, OR: AUAI Press.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422–445.

Simon, H. A., & Rescher, N. (1966). Cause and counterfactual. *Philosophy and Science, 33,* 323–340.

Sobel, M. E. (1998). Causal inference in statistical models of the process of socioeconomic achievement. *Sociological Methods & Research, 27,* 318–348.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics, 33,* 230–231.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Tian, J., Paz, A., & Pearl, J. (1998). *Finding minimal separating sets. Technical report R-254.* Los Angeles: University of California, Los Angeles.

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence, 28,* 287–313.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology, 20,* 18–26.

VanderWeele, T. J., & Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology, 18,* 561–568.

Wermuth, N. (1992). On block-recursive regression equations. *Brazilian Journal of Probability and Statistics* (with discussion), *6,* 1–56.

Wermuth, N., & Cox, D. (1993). Linear dependencies represented by chain graphs. *Statistical Science, 8,* 204–218.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* Chichester, UK: John Wiley.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

# Section V

# Ethics and Communicating Findings

# 16

# *Ethical Issues in the Conduct and Reporting of Meta-Analysis*

**Harris Cooper**
*Duke University*

**Amy Dent**
*Duke University*

A research synthesis focuses on empirical studies and attempts to summarize past research by drawing overall conclusions from separate studies that address the same or related hypotheses. The research synthesist's goal is "to present the state of knowledge concerning the relation(s) of interest and to highlight important issues that research has left unresolved" (Cooper, 2010, p. 4). Meta-analysis is a type of research synthesis. It involves the statistical integration of data from separate but similar studies typically using the summary statistics presented in research reports. Meta-analysts (a) systematically collect as many published and unpublished reports addressing a topic as possible, (b) extract effect sizes from the reports, (c) statistically combine the effect sizes to obtain an estimate of the average effect size and the associated confidence interval, and (d) examine sample and study features that might influence study outcomes.

When it comes to ethical considerations, research synthesists and meta-analysts have it easy. Unlike primary researchers, they face no issues regarding the treatment of the humans or animals who participate in their work. There are no *institutional review boards* to convince that the benefits of their work outweigh the risks. Because public documents are the object of study, informed consent and confidentiality are not an issue; public documents cannot be deceived or mistreated.

Still, conducting a research synthesis or meta-analysis is not without ethical considerations. Meta-analysts face the same ethical issues faced by quantitative methodologists discussed in the other chapters in this volume but in a different context. Some of these ethical considerations relate to the process of reporting and publishing research results of any kind. For example, one treatment of ethical obligations in reporting research

can be found in the *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association [APA], 2002). Here, researchers, whether reporting a new data collection or meta-analysis, are obligated not to fabricate data, to correct errors when they are found, not to plagiarize the work of others or publish data more than once, to allocate authorship credit appropriately, and to share their data with others for purposes of verification. These ethical obligations are reproduced verbatim from the APA *Principles* in Table 16.1.

In the context of discussing the use and misuse of quantitative methods more generally, Brown and Hedges (2009) provide one of the few previous treatments of meta-analysis in an ethical context. They begin by stating one premise that informs all the chapters in this book:

> Methodological rigor is closely related to ethical vigilance: When research, statistical calculations, and data presentation can be done better and more accurately, they should be. That is, there is an ethical imperative to demand and use the highest standards of research and data presentation. (Brown & Hedges, 2009, p. 375)

With regard to meta-analysis in particular, Brown and Hedges identify three points at which a lack of methodological rigor raises ethical issues. First, they point out that meta-analysis can involve collecting, summarizing, and integrating massive amounts of data. Performing these tasks improperly—whether purposely or inadvertently—can lead to erroneous conclusions. Certainly when meta-analyses are conducted improperly on purpose the ethical violation is clear. In the inadvertent case, a lack of vigilance or the carrying out of analyses that are beyond an investigator's expertise also can suggest an ethical breach. Second, Brown and Hedges point out that the decision to include or exclude a study from a meta-analysis can raise ethical issues unless the criteria for study inclusion and exclusion have been made transparent and uniformly applied to studies. Finally, Brown and Hedges assert that it is an ethical obligation of meta-analysts to consider the possibility that publication bias may influence their results.

Thus, it appears these authors suggest three ethical dicta to be followed in conducting and reporting a meta-analysis: (a) extract and analyze your data accurately; (b) make your inclusion and exclusion criteria explicit and apply them consistently; and (c) test for publication bias. Is that it? Perhaps not. Brown and Hedges astutely point out that "what starts out as an identification of best practices can evolve into ethical expectations" (p. 378). We would add to this the suggestion that "best practice" becomes an ethical consideration when the aspect of methodology under consideration is one for which the conclusions of research are heavily dependent. How much does it influence findings if you do

**TABLE 16.1**

Entries in the *Ethical Principles of Psychologists and Code of Conduct* Relating to Reporting Research Results and Publication

**8.10 Reporting Research Results**

(a) Psychologists do not fabricate data. (See also Standard 5.01a, Avoidance of False or Deceptive Statements.)

(b) If psychologists discover significant errors in their published data, they take reasonable steps to correct such errors in a correction, retraction, erratum, or other appropriate publication means.

**8.11 Plagiarism**

Psychologists do not present portions of another's work or data as their own, even if the other work or data source is cited occasionally.

**8.12 Publication Credit**

(a) Psychologists take responsibility and credit, including authorship credit, only for work they have actually performed or to which they have substantially contributed. (See also Standard 8.12b, Publication Credit.)

(b) Principal authorship and other publication credits accurately reflect the relative scientific or professional contributions of the individuals involved, regardless of their relative status. Mere possession of an institutional position, such as department chair, does not justify authorship credit. Minor contributions to the research or to the writing for publications are acknowledged appropriately, such as in footnotes or in an introductory statement.

(c) Except under exceptional circumstances, a student is listed as principal author on any multiple-authored article that is substantially based on the student's doctoral dissertation. Faculty advisors discuss publication credit with students as early as feasible and throughout the research and publication process as appropriate. (See also Standard 8.12b, Publication Credit.)

**8.13 Duplicate Publication of Data**

Psychologists do not publish, as original data, data that have been previously published. This does not preclude republishing data when they are accompanied by proper acknowledgment.

**8.14 Sharing Research Data for Verification**

(a) After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release. This does not preclude psychologists from requiring that such individuals or groups be responsible for costs associated with the provision of such information.

(b) Psychologists who request data from other psychologists to verify the substantive claims through reanalysis may use shared data only for the declared purpose. Requesting psychologists obtain prior written agreement for all other uses of the data.

*Source:* From APA, *Ethical Principles of Psychologists and Code of Conduct*, APA, New York, 2002. With permission.

it right or wrong? How easy is it to "manipulate" findings (to arrive at a predetermined or biased outcome) by doing it one way or the other? Because the techniques used in meta-analysis are relatively new and still evolving, we can anticipate that standards of best practice and ethical expectations are also rapidly evolving.

Examining how the standards surrounding meta-analysis are evolving is the objective of this chapter. To begin, we will present and provide a brief background concerning a set of guidelines for the reporting of meta-analyses recently developed by the APA, called *meta-analysis reporting standards* (MARS; APA Publication and Communication Board Working Group on Journal Article Reporting Standards,[1] 2008). Then, we will describe the results of a survey conducted using the members of the Society for Research Synthesis Methodology as respondents. Participants were asked about what aspects of a meta-analysis were and were not important to report. Those that rose to a level that suggested omitting these aspects of meta-analysis might be considered a breach of ethics will be discussed in some detail. Finally, we will conclude with some other ethics-related issues that have emerged for meta-analysts, specifically, the use of *auxiliary websites*, the use of individual participant data in meta-analysis, and the identification of duplicate publications.

## APA's Meta-Analysis Reporting Standards

In developing its meta-analysis reporting standards, the APA Working Group distinguished between three levels of prescription: "recommendations," "standards," and "requirements." Using Merriam Webster's Online Dictionary (2007) as its source of definitions, *to recommend* was defined as "to present as worthy of acceptance or trial … to endorse as fit, worthy, or competent …"; a *standard* was defined as "… something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality …"; and a *requirement* was defined as something that was asked for "by right and authority … to call for as suitable or appropriate … to demand as necessary or essential …." From an ethical perspective, failing to meet requirements certainly could be considered problematic, but failing to meet recommendations or standards would be less troubling, depending on the circumstance. The APA Working Group decided that its proposals

[1] The first author of this chapter served as chair of this working group.

should "… be viewed as standards or, at least, a beginning effort at developing standards" (p. 847).

MARS was developed by integrating four efforts by other groups of researchers and editors knowledgeable about meta-analysis: the *QUOROM Statement* (*Quality of Reporting of Meta-analysis*; Moher et al., 1999) and its revision, *PRISMA* (*Preferred Reporting Items for Systematic Reviews and Meta-analyses*; Moher, Tetzlaff, Liberati, Altman, & the PRISMA Group, 2009), *MOOSE* (*Meta-analysis of Observational Studies in Epidemiology*; Stroup et al., 2000), and the Potsdam consultation on meta-analysis (Cook, Sackett, & Spitzer, 1995). The APA Working Group combined the nonredundant elements contained in these previous documents, rewrote some items for an audience of psychologists (and others who might use APA's *Publication Manual*, 2010), and added a few suggestions of its own. Then, the APA Working Group asked for additional suggestions from a subgroup of members of the Society for Research Synthesis Methodology known to have interest in psychology and from the members of APA's Publications & Communications Board. The approved final version (along with the Journal Article Reporting Standards [JARS]) appeared in the *American Psychologist* in December 2008 and was reproduced in the sixth edition of the APA *Publication Manual* (2010). The MARS is reproduced in Table 16.2.

## A Survey of the Members of the Society for Research Synthesis Methodology

The MARS calls on meta-analysts to address more than 70 different aspects of methodology in their reports. Certainly, these are not of equal import, and their importance would vary as a function of the topic under consideration. Perhaps then, another level of guidance is needed: Which aspects of meta-analysis reporting are optional, context dependent, and required for the authors to have met their obligations as researchers?

To answer this question, we conducted an online survey of the 74 members of the Society for Research Synthesis Methodology. Forty-two (57%) of the Society's members responded to the survey 2 weeks after receiving the invitation (and 1 week after a reminder was sent). In addition to responding to the survey, we also asked participants several questions about their background. In one, they categorized their broad area of interest into one of five categories. Of the 42 respondents, 18 chose medicine and health; 21 chose psychology, education, or public and social policy; and 3 chose "other." Also, we asked how many years the respondents

**TABLE 16.2**

Meta-Analysis Reporting Standards

| Paper Section and Topic | Description |
|---|---|
| *Title* | |
| | • Make it clear that the report describes a research synthesis and include "meta-analysis," if applicable<br>• Footnote funding source(s) |
| *Abstract* | |
| | • The problem or relation(s) under investigation<br>• Study eligibility criteria<br>• Type(s) of participants included in primary studies<br>• Meta-analysis methods (indicating whether a fixed or random model was used)<br>• Main results (including the more important effect sizes and any important moderators of these effect sizes)<br>• Conclusions (including limitations)<br>• Implications for theory, policy, and/or practice |
| *Introduction* | |
| | • Clear statement of the question or relation(s) under investigation<br>  • Historical background<br>  • Theoretical, policy, and/or practical issues related to the question or relation(s) of interest<br>  • Rationale for the selection and coding of potential moderators and mediators of results<br>  • Types of study designs used in the primary research, their strengths and weaknesses<br>  • Types of predictor and outcome measures used, their psychometric characteristics<br>  • Populations to which the question or relation is relevant<br>  • Hypotheses, if any |
| *Method* | |
| Inclusion and exclusion criteria | • Operational characteristics of independent (predictor) and dependent (outcome) variable(s)<br>• Eligible participant populations<br>• Eligible research design features (e.g., random assignment only, minimal sample size) |

| | |
|---|---|
| | • Time period in which studies needed to be conducted<br>• Geographical and/or cultural restrictions |
| Moderator and mediator analyses | • Definition of all coding categories used to test moderators or mediators of the relation(s) of interest |
| Search strategies | • Reference and citation databases searched<br>• Registries (including prospective registries) searched<br>  • Keywords used to enter databases and registries<br>  • Search software used and version (e.g., Ovid)<br>• Time period in which studies needed to be conducted, if applicable<br>• Other efforts to retrieve all available studies, e.g.,<br>  • Listservs queried<br>  • Contacts made with authors (and how authors were chosen)<br>  • Reference lists of reports examined<br>• Method of addressing reports in languages other than English<br>• Process for determining study eligibility<br>  • Aspects of reports were examined (i.e., title, abstract, and/or full text)<br>  • Number and qualifications of relevance judges<br>  • Indication of agreement<br>  – How disagreements were resolved<br>• Treatment of unpublished studies |
| Coding procedures | • Number and qualifications of coders (e.g., level of expertise in the area, training)<br>• Intercoder reliability or agreement<br>• Whether each report was coded by more than one coder and, if so, how disagreements were resolved<br>• Assessment of study quality<br>  • If a quality scale was employed, a description of criteria and the procedures for application<br>  • If study design features were coded, what these were<br>• How missing data were handled |

(*Continued*)

**TABLE 16.2 (Continued)**

Meta-Analysis Reporting Standards

| Paper Section and Topic | Description |
|---|---|
| Statistical methods | • Effect size metric(s)<br>  • Effect sizes calculating formulas (e.g., means and SDs, use of univariate F to r transform, etc.)<br>  • Corrections made to effect sizes (e.g., small sample bias, correction for unequal ns, etc.)<br>• Effect size averaging and/or weighting method(s)<br>• How effect size confidence intervals (or standard errors) were calculated<br>• How effect size credibility intervals were calculated, if used<br>• How studies with more than one effect size were handled<br>• Whether fixed and/or random effects models were used and the model choice justification<br>• How heterogeneity in effect sizes was assessed or estimated<br>• Means and SDs for measurement artifacts, if construct-level relationships were the focus<br>• Tests and any adjustments for data censoring (e.g., publication bias, selective reporting)<br>• Tests for statistical outliers<br>• Statistical power of the meta-analysis<br>• Statistical programs or software packages used to conduct statistical analyses |
| *Results* | |
| | • Number of citations examined for relevance<br>• List of citations included in the synthesis<br>• Number of citations relevant on many but not all inclusion criteria excluded from the meta-analysis<br>  • Number of exclusions for each exclusion criteria (e.g., effect size could not be calculated), with examples<br>• Table giving descriptive information for each included study, including effect size and sample size<br>• Assessment of study quality, if any<br>• Tables and/or graphic summaries<br>  • Overall characteristics of the database (e.g., number of studies with different research designs)<br>  • Overall effect size estimates, including measures of uncertainty (e.g., confidence and/or credibility intervals) |

- Results of moderator and mediator analyses (analyses of subsets of studies)
  - Number of studies and total sample sizes for each moderator analysis
  - Assessment of interrelations among variables used for moderator and mediator analyses
- Assessment of bias including possible data censoring

*Discussion*

- Statement of major findings
- Consideration of alternative explanations for observed results
  - Impact of data censoring
- Generalizability of conclusions, e.g.,
  - Relevant populations
  - Treatment variations
  - Dependent (outcome) variables
  - Research designs, etc.
- General limitations (including assessment of the quality of studies included)
- Implications and interpretation for theory, policy, or practice
- Guidelines for future research

*Source:* From Journal Article Reporting Standards Working Group, *Am. Psychol.,* 63, 839, 2008. With permission.

had been working with research synthesis methodologies. Twenty of the respondents said they had 15 or more years of experience.[2]

Each participant was told that he or she would be presented with about 70 different aspects of conducting a research synthesis or meta-analysis and would respond to the question "Generally speaking, how important is it that each be described in the report of a synthesis?"[3] The response scale was:

10 = Generally, it would be considered UNETHICAL in my field NOT TO INCLUDE this information in the report (10 on the scale was labeled "generally must include").

5 = Generally, researchers in my field MIGHT or MIGHT NOT report this information depending on characteristics of the specific literature (5 read "depends on the specific literature").

1 = Generally, it is UNNECESSARY in my field for researchers to report this information (1 read "generally unnecessary")[4]

N/A = Generally, this aspect of research synthesis is NOT APPLICABLE to my area of interest.

A comment box was provided after each request for a rating.

Table 16.3 presents the results of the survey. Note that nearly all of the elements of the MARS were included on the survey with a few exceptions we deemed to be trivial (e.g., we deemed it unnecessary to ask the Society members what the elements of a title or abstract ought to be for a synthesis report).

[2] We used these two background questions to examine whether responses to the survey were related to the participant's background. However, of 280 statistical tests we conducted (comparing the area and experience of members on the (a) frequency of scores of 10 and (b) mean response for each of the 70 items), we found two that reached statistical significance ($p < .05$). Given that this was less than the expected number of significant findings if chance were operating, we concluded that responses did not differ as a function of substantive area or experience.

[3] It might have been more interesting to poll the experts on what they considered best practice in conducting a meta-analysis rather than on reporting standards. However, issues regarding best practice in meta-analysis are more complex and in some instances still contentious. This would make the relevance of such a survey to ethical issues more difficult to discern. For example, meta-analysis experts disagree about whether it is ever appropriate to use a fixed-effect, rather than a random-effect, model for estimating error; we will see that most agree that reporting which model was used and why is a necessary part of reporting.

[4] Two respondents pointed out that the survey left it ambiguous whether answers should be based on their personal opinion or on the norms that prevailed in their field. We should have made it clearer that we were interested in norms of the field. Regardless, for the relatively broad purposes to which the survey is put, we think it is not a severe problem that responses include a mix of perceived norms and the wishes of experts.

**TABLE 16.3**

Results of the Survey of the Society for Research Synthesis Methodology Regarding Reporting Standards for Meta-Analysis

| Aspect of Meta-Analysis Reporting | Number of 10s[a] | Number Below 5[b] | Mean |
|---|---|---|---|
| *Introduction* | | | |
| **1. Clear statement of the research question** | **34** | **0** | **9.74** |
| 2. Narrative account of the development of the research question | 8 | 2 | 6.73 |
| 3. Theoretical, policy, and/or practical issues related to the research question | 9 | 4 | 7.36 |
| 4. Rationale for the selection and coding of potential moderators and mediators of results | 14 | 2 | 7.62 |
| 5. Types of study designs … their strengths and weaknesses | 14 | 3 | 7.80 |
| **6. Independent (predictor) and dependent (outcome) variables of primary interest** | **25** | **1** | **9.07** |
| **7. Populations to which the question is relevant** | **23** | **2** | **8.45** |
| 8. Hypotheses, if any | 14 | 4 | 7.46 |
| *Methods* | | | |
| **1. Operational definitions of independent (predictor) and dependent (outcome) variable(s)** | **23** | **2** | **8.51** |
| **2. Eligible participant populations** | **24** | **1** | **8.88** |
| **3. Eligible research design features …** | **28** | **2** | **8.90** |
| 4. Time period in which studies needed to be conducted | 20 | 3 | 8.17 |
| 5. Geographical and/or cultural restrictions | 10 | 4 | 6.69 |
| **6. Whether unpublished studies were included or excluded** | **24** | **2** | **8.67** |
| **7. Reference and citation databases searched** | **30** | **2** | **9.19** |
| 8. Registries (including prospective registries) searched | 17 | 7 | 7.30 |
| 9. Keywords used to enter databases and registries | 17 | 4 | 8.05 |
| 10. Search software used to enter electronic databases (e.g., Ovid) | 9 | 18 | 5.57 |
| 11. Conference proceedings searched | 7 | 11 | 6.23 |

(*Continued*)

**TABLE 16.3 (Continued)**

Results of the Survey of the Society for Research Synthesis Methodology Regarding Reporting Standards for Meta-Analysis

| Aspect of Meta-Analysis Reporting | Number of 10s[a] | Number Below 5[b] | Mean |
|---|---|---|---|
| 12. Listservs queried | 7 | 10 | 5.89 |
| 13. Contacts made with researchers in the field and how these researchers were chosen | 5 | 9 | 6.12 |
| 14. Whether reference lists of reports were examined | 12 | 4 | 7.62 |
| 15. Method of addressing reports in languages other than English | 14 | 5 | 7.14 |
| 16. Aspects of reports used to determine relevance (i.e., title, abstract, and/or full text) | 16 | 7 | 7.57 |
| 17. Number and qualifications of relevance judges | 13 | 10 | 6.81 |
| 18. Indications of judge agreement if more than one judge examined each report | 15 | 7 | 7.31 |
| 19. How judge disagreements were resolved | 17 | 4 | 7.90 |
| 20. Number and qualifications of coders (e.g., level of expertise in the area, training) | 10 | 10 | 6.59 |
| 21. Intercoder reliability or agreement | 15 | 7 | 7.43 |
| 22. Whether each report was coded by more than one coder … how disagreements … resolved | 19 | 3 | 8.02 |
| 23. How missing data were handled | 17 | 2 | 8.12 |
| 24. Definitions of ALL coding categories … | 13 | 3 | 7.60 |
| 25. Criteria of the quality scale and procedure for application | 20 | 6 | 7.80 |
| **26. Study design features that were coded** | **23** | **1** | **8.74** |
| **27. Effect size metric(s)** | **34** | **0** | **9.68** |
| 28. Effect sizes calculating formulas … | 8 | 11 | 6.24 |
| 29. Corrections made to effect sizes … | 20 | 1 | 8.34 |
| **30. Effect size averaging and weighting method(s)** | **31** | **0** | **9.00** |
| 31. How effect size confidence intervals (or standard errors) were calculated | 17 | 3 | 8.13 |
| 32. How effect size credibility intervals were calculated | 10 | 4 | 6.89 |
| **33. How studies with more than one effect size were handled** | **27** | **1** | **8.93** |

| | | | |
|---|---|---|---|
| **34. Whether fixed and/or random effects models were used** | **32** | **2** | **9.31** |
| **35. The justification for the choice of the error model (fixed, random)** | **22** | **7** | **8.00** |
| **36. How heterogeneity in effect sizes was assessed or estimated** | **23** | **1** | **8.93** |
| 37. Means and SDs for measurement artifacts | 2 | 7 | 5.24 |
| 38. Tests and any adjustments for data censoring (e.g., publication bias, selective reporting) | 14 | 3 | 7.83 |
| 39. Tests for statistical outliers | 9 | 8 | 6.14 |
| 40. Statistical power of the meta-analysis | 3 | 23 | 4.07 |
| 41. Statistical programs or software packages used to conduct statistical analyses | 16 | 9 | 7.36 |

[a] "Number of 10s" is the number of respondents out of 42 who said, "Generally, it would be considered UNETHICAL in my field NOT TO INCLUDE this information in the report" (10 on the scale was labeled "generally must include").

[b] "Number below 5" is the number of respondents who gave this reporting aspect a score less than 5.

[c] Questions with " … " have been shortened in the table by removing examples or clarifying information. Precise wording of MARS questions can be found in Table 16.2.

To begin interpreting Table 16.3, it is interesting to look at the responses regarding Brown and Hedges' (2009) three dicta for conducting a meta-analysis. The first of the dicta—extract and analyze your data accurately—is hard to attach to any particular question or questions because it is a broad prescription. The second dicta—make your inclusion and exclusion criteria explicit and apply them consistently—relates to several questions on the survey. These included the operational definitions of independent and dependent variables, eligible participant populations, eligible research designs, and any time period, geographic, or cultural restrictions. For these questions, between 10 and 28 respondents answered that "it would be considered UNETHICAL in my field NOT TO INCLUDE this information in the report." The third dicta—test for publication bias—relates to two questions. Twenty-four respondents believed it would be unethical not to report whether unpublished research was included or excluded from the synthesis, and 14 believed that authors must include tests and any adjustments for data censoring (e.g., publication bias, selective reporting).

Based on these results, we think it would not be unreasonable to suggest that if 21 (50%) or more of the Society's members responded that they believed not including this information in a report would be considered unethical or generally must be included that this be viewed as an indication that this reporting practice now approaches the point where best practice becomes an ethical expectation. Below we present the elements of reporting that reached this threshold and present our thinking about why this was the case; in other words, why the choices researchers make at these points can have large effects on the results of their syntheses.

## Aspects of Meta-Analysis Reporting Approaching Ethical Obligation

### The Problem Statement

Three aspects of reporting syntheses that reached our threshold for raising ethical concerns related to the problem statement. More than half of respondents believed that it would be unethical not to include (a) a clear statement of the research question ($n$ = 34; 81%), (b) what were the independent (predictor) and dependent (outcome) variables of primary interest ($n$ = 25; 60%), and (c) a description of the populations to which the question is relevant ($n$ = 23; 55%). For example, a synthesis that claims to examine the relationship between frustration and aggression would need to provide a clear statement of how the variables of interest are

defined conceptually (e.g., frustration involves the blocking of goal attainment; aggression involves the intent to harm), what type of relationship is of interest (associational or causal), and among whom (e.g., animals or humans; children, adolescents, or adults).

A high level of concern regarding these aspects of reporting would probably be evident in a similar survey related to primary research. Without a clear statement of the problem, the variables involved, and the relevant populations, it would be impossible to evaluate the contribution the research makes to the literature, if indeed the relevant literature could be identified.

### The Inclusion Criteria

Three aspects of reporting the criteria for including and excluding studies from the synthesis reached our threshold for raising ethical concerns. More than half of respondents believed it would be ethically problematic not to include in the method section (a) the operational definitions of independent (predictor) and dependent (outcome) variables ($n$ = 23; 55%); (b) a description of the eligible participant populations ($n$ = 24; 57%); and (c) the eligible research design features ($n$ = 28; 67%). Nearly half ($n$ = 20; 48%) gave the highest rating to the need to include any time period restrictions. These concerns about the inclusion and exclusion criteria parallel the conceptual concerns that arose when respondents rated the importance of aspects of the problem statement. There are good reasons for this, and these reasons are especially relevant to research syntheses.

Eligibility criteria take on unique importance because the research designs and characteristics of units sampled in a research synthesis can be considerably more varied than typically is the case for a single primary study. Research synthesists often begin their work with broad conceptual definitions. In the course of searching the literature they may come across numerous operational realizations of the concepts defined in their problem statement. For example, synthesists examining the relation between frustration and aggression might discover studies that used numerous techniques to measure or instill frustration in participants (e.g., asking them to wait in line for a long time, playing a video game that is difficult to win) and numerous ways to measure aggression (e.g., shouting, pushing, hitting). Given this variety, our respondents believed it was of great importance that readers know precisely what the synthesists defined as "in" and "out." Only with this information would readers be able to object, for example, if "shouting" is included as a measure of aggression because they believe verbal attacks are not really meant to harm.

Also, readers might have no objection to the conceptual and operational definitions of the problem but may want to judge whether the concepts and operations fit together well. They may want to determine whether the

operations used in previous research fit the concept definitions used by the synthesists, or whether a broader or narrower conceptualization would be more appropriate. For example, the synthesists might find "shouting" was used as a measure of aggression when initially they had not considered verbal aggression for inclusion. When such a circumstance arises, the synthesists must broaden their conceptual definitions to include these operations, so now aggression includes both physical and verbal assault. If this "refitting" is not done, the conclusions in the synthesis might appear to apply more generally or narrowly than warranted by the data. Readers will not be able to assess fit without a clear statement of the included and excluded operational definitions.

Similarly, synthesists need to clearly tell readers what units were and were not considered relevant to addressing the research question. Without this information, the reader cannot assess to whom the results apply. Nor can the reader object if he or she believes samples from irrelevant populations have been included in the synthesis. One respondent commented that this information was usually included but was vague and "not explicitly defined a priori—also not necessarily considered in relation to external validity." Another commented that in her or his area, people "generally use convenience samples and we agree not to talk about it."

Of equal importance, a clear description of the methodological characteristics of included and excluded studies allows the reader to gauge the fit between the included research designs, how the design was implemented, and the inferences drawn by the synthesists (Valentine & Cooper, 2008). Above, the survey respondents indicated that research synthesists must provide an explicit statement about the type of relationship under study. This aspect of rigorous research synthesis takes center stage when readers consider the correspondence between the design and implementation of individual studies and the desired inferences of the synthesis. For example, the research question "Does frustration cause aggression?" suggests that the research synthesists should focus primarily on summarizing research that used experimental and quasi-experimental designs, whereas the question "Is frustration associated with aggression?" might include cross-sectional designs as well. How readers will evaluate how well the synthesists' inferences correspond with the data will depend on what kinds of designs were admitted as evidence.

We think as well that this same line of reasoning was behind the respondents' frequent use of the highest ratings ($n = 23$; 55%) for the importance of including a thorough description of the study design features that were coded by the synthesists. Here, the description relates only to studies that were included in the synthesis, but the principle is the same. Also, 20 respondents, one short of our criteria, gave the highest rating to the importance of describing the criteria of the quality scale and procedure for its application. This element of a research report might not have reached

our threshold because not all meta-analysts think using quality scales is a good idea (Valentine & Cooper, 2008).

In sum, the inclusion–exclusion aspects of reporting that rose to the level of ethical considerations relate to the readers' ability to evaluate the fit between concepts and operations in research synthesis and the fit between the inferences drawn and what inferences the data can support. Without this information, readers will be unable to decide whether clear and legitimate linkages exist (a) between concepts and operations and (b) between research designs, study implementation, and the interpretation of results (see Cooper, 2010).

## The Parameters of the Literature Search

### *Whether Unpublished Studies Were Included or Excluded*

Not surprisingly, respondents felt strongly about the need to report whether unpublished studies were included in the research synthesis ($n$ = 24; 57%). The concern here is that studies revealing smaller effects will be systematically omitted from the published literature, making relationships appear stronger than if all estimates were retrieved and included. Lipsey and Wilson (1993) compared the magnitudes of effects reported in published versus unpublished studies contained in 92 different meta-analyses. They reported that on average the impact of interventions in unpublished research was one-third smaller than published effects.

A reason frequently given for excluding unpublished research is that it has not undergone the peer-review process and therefore may be of lesser quality. However, researchers often do not publish their results because publication is not their objective (cf. Cooper, DeNeve, & Charlton, 2001); publication does not help them get their work before the audience they seek. For example, some research is conducted to meet degree or course requirements or as evaluations for agencies making decisions about program effectiveness. Also, research is often turned down for journal publication because it is not a novel contribution (although direct replications are of great interest to research synthesists) or because the statistical test fails to achieve standard levels of statistical significance, a problem known as "bias against the null hypothesis" (Rothstein, Sutton, & Borenstein, 2005). Conversely, some low quality research does get published.

For these reasons, it is now "best practice" in the social sciences for research synthesists to include both published and unpublished research. If the synthesists include only published research, their report must include a convincing justification. Our survey responses suggest that providing a clear description and justification for whether and why unpublished

research was or was not included in the synthesis has crossed over into an ethical obligation.

### *Reference and Citation Databases Searched*

The sources of information that provide the most evidence that goes into a research synthesis are likely to be reference databases and citation indexes. Even though reference databases are superb sources of studies, they still have limitations. First, different reference databases restrict what is allowed to enter the system based on their topical or disciplinary coverage. Second, some reference databases contain only published research; others contain both published and unpublished research; and others contain just unpublished research (e.g., dissertation abstracts). Third, there can be a time lag between when a study is completed and when it will appear in the reference database (although technology has reduced this lag dramatically), and this may vary depending on the database. Without information on the databases used, it will be difficult for readers to assess (a) the literature coverage and (b) what studies might have been missed. Equally important, without this information it would be extremely difficult to replicate the results of the synthesis.

## The Measure of Effect

### *The Effect Size Metric(s)*

Although numerous estimates of effect size are available (Cohen, 1988), three dominate the literature: (a) the *d*-index, which is a scale-free measure of the separation between two group means calculated by dividing the difference between the two group means by either their average standard deviation or the standard deviation of the control group; (b) the *r*-index, or correlation coefficient; and (c) the odds ratio, or some variant thereof, applicable when both variables are dichotomous and findings are presented as frequencies or proportions.

The term *effect size* is sometimes used broadly to denote all measures of relationship strength, and sometimes it is used as an alternative label for the *d*-index. This is regrettable because the metrics, although translatable, are not identical. For example, a value of .40 for a *d*-index corresponds to an *r*-index value of .196. Thus, as one respondent noted, "If we don't know the metric then we don't know how to evaluate" the results. Further, it is not always the case that the choice of an *effect size metric* reflects the important design characteristics of the studies from which they are derived (specifically the dichotomous or continuous nature of the variables involved). Therefore, the survey respondents indicated that readers of research syntheses need to be explicitly informed of what indexes were used and why they were chosen ($n = 34$).

### *Effect Size Averaging and Weighting Method(s)*

Once each effect size has been calculated, the meta-analysts next average the effects that estimate the same relationship. It is generally accepted that the individual effect sizes should be weighted by the inverse of the variance (based on the number of participants in their respective samples) before they are averaged. Sometimes, however, unweighted effect sizes are presented. Weighted and unweighted effect sizes can differ in magnitude, the difference depending on the degree of relationship between the size of the effect and the sample size. Therefore, if larger effects are associated with smaller sample sizes (a condition likely obtained if the synthesists were more likely to find studies that produced statistically significant results), an unweighted average effect size would be larger, sometime much larger, than the weighted average. For this reason, the survey respondents found it important that the procedures used to generate average effect sizes were essential to a complete synthesis report.

### *How Studies With More Than One Effect Size Were Handled*

A problem for meta-analysts arises when a single study contains multiple effect size estimates. This is most bothersome when more than one measure of the same construct appears in a study and the measures are analyzed separately. Because the same participants provided multiple outcomes, these measures are not independent, and it would generally be inappropriate to treat them as such when combining the effect sizes across all studies. If they are, studies with more measures would get more weight in an averaged effect size, and the assumption that effect size estimates are independent would be violated in subsequent analyses.

There are several approaches meta-analysts use to handle dependent effect sizes. Some meta-analysts treat the effect size as if it were independent. Alternatively, the study might be used as the unit of analysis, taking the mean or median effect size to represent the study. Another approach is to use a shifting unit of analysis (Cooper, 2010). Here, each effect size associated with a study is first coded as if it were an independent estimate of the relationship. However, for estimating the overall effect size, these are averaged before entry into the analysis, so that the study only contributes one value. However, in analyses that examine moderators, for example, whether physical or verbal aggression is influenced more by frustration, the studies would be permitted to contribute one effect size to the estimate of each category's mean effect size. Finally, more sophisticated statistical approaches also have been suggested as a solution to the problem of dependent effect size estimates (Gleser & Olkin, 2009).

Which of the available techniques the meta-analysts use can have a large impact on the estimated average magnitude of the effect size, the

estimated variance among effect sizes, and the power of tests to uncover moderators of effects? For this reason, a majority of respondents believed meta-analysts might be ethically obligated to report which approach was used in handling nonindependent estimates of effect. One respondent also pointed out that "Specific details are preferable. For instance, simply stating that a 'shifting unit of analysis' approach was used doesn't specify how 'average' effect sizes were computed or how the conditional variances and weights were handled."[5]

### Variation Among Effect Sizes

Three aspects of meta-analysis methodology related to how the variation in effect sizes was treated generated a majority of responses at the high extreme of the scale: whether fixed-effect or random-effect models were used, the justification for the use of a fixed-effect or random-effect model, and how heterogeneity in effect sizes was assessed or estimated.

One important aspect of averaging effect sizes and estimating their dispersion involves the decision about whether a fixed-effect or random-effect model underlies the generation of study outcomes. In a *fixed-effect model*, each effect size's variance is assumed to reflect sampling error of participants only, that is, error solely the result of participant differences. However, other features of studies can be viewed as additional random influences. Thus, in a random-effect analysis, study-level variance is assumed to be present as an additional source of random influence. Hedges and Vevea (1998, p. 3) state that fixed-effect models are most appropriate when the goal of the research is "to make inferences only about the effect size parameters in the set of studies that are observed (or a set of studies identical to the observed studies except for uncertainty associated with the sampling of subjects)." A further statistical consideration is that in the search for moderators fixed-effect models may seriously underestimate error variance and random-effect models may seriously overestimate error variance when their assumptions are violated (Overton, 1998). Schmidt, On, and Hayes (2009) suggest that random-effect models should always be used, whereas Cooper (2010) proposes that both models can be applied to the data and the results interpreted accordingly.

Random-effect models are typically more conservative than fixed-effect models, in the sense that they will estimate more variability around average effect sizes and therefore are less likely to reveal statistically significant effects and moderators of effects. The two models also can generate different average effect sizes, again depending on the relationship between

[5] Another respondent mused theologically that "Cardinal Newman tried to convince people of the existence of God … and he enunciated that it is more compelling to have independent evidence than repeated versions of the same evidence"

the size of effects and the size of samples. A large majority of respondents to our survey ($n$ = 32) believed meta-analysts were ethically obligated to report whether a fixed- or random-effect model was used, why it was chosen ($n$ = 22; 52%), and more generally how heterogeneity in effect sizes was assessed or estimated ($n$ = 23; 55%).

### Tabling Data

Not surprisingly, 26 (62%) respondents believed that meta-analysts were obligated to present a summary table of their results. Related to this table, a near majority of respondents ($n$ = 20; 48%) used the extreme high end of the scale when rating the importance of including information on the number of studies, sample sizes, and subgroup effect sizes for each moderator analysis.

Given these results, perhaps what is surprising is that a table listing the results for the individual studies going into the analyses did not meet our threshold, although 17 (40%) respondents did give this feature of a report the "ethically-obligatory" rating. One respondent commented about tables of individual studies, "If I could have scored this '11' I would have done so!" We suspect that this table did not receive more ratings of 10 because of concerns about space limitations in many reports, especially ones being readied for publication. Two respondents supported this interpretation; one wrote, "But subjected to space limitations," and another noted, "We usually prepare the table for each meta-analysis, however, some journals do not publish it due to the limited number of tables/figures allowed." We will return to this issue below.

## The Interpretation of Results

Finally, a majority of our respondents believed that four of the nine MARS elements of a discussion section were obligatory: a statement of major findings ($n$ = 38; 90%); consideration of alternative explanations for observed results ($n$ = 23; 55%); a discussion of the general limitations (including assessment of the quality of studies included) ($n$ = 23; 55%)[6]; and a discussion of the implications and interpretation of findings for theory, policy, or practice ($n$ = 24; 57%). Another element, populations to whom the results were relevant ($n$ = 19; 45%), approached our threshold. Similar to elements of the introduction, these are likely viewed as obligatory of all good science reporting, not just research synthesis.

[6] One respondent cautioned, "Although these need careful thought to avoid biased interpretations."

## Some Additional Issues Related to the Reporting of Meta-Analysis

### Space Limitations and the Use of Auxiliary Websites

One issue that arises when reporting standards are discussed is the tension between the desire to exhaustively report the background, methods, and results of a study and the space limitations of journals. In fact, this issue generated the most open-ended comments from the Society members responding to our survey. One member wrote:

> Something must be said about practicality and the limitations imposed by editors. [For example] in a recent paper we had 26 pages of included references and the editors wanted us to condense a review of over 200 studies into 40 pages max with all tables. Adding all these potentially important details is impossible in most reports.

Another wrote:

> A major difficulty I've encountered with reporting research syntheses is journal space constraints. It's often infeasible to include sufficient detail about every synthesis phase, especially if there are several studies, complex results (e.g., several distinct effect sizes and multiple moderator analyses). Reporting things like details of each study is essentially impossible, especially in outlets where a certain amount of "didactic" detail about meta-analysis is required for readers unfamiliar with meta-analysis.

And a third wrote:

> We just got a meta-analysis tentatively accepted to … and they are asking us to omit nearly all of the tables and "technical" details. We do not plan to do this totally but probably will need to relegate it to appendices. We are going to resist this as much as possible and will cite JARS/MARS as part of our argument for including it.

Journals have only limited printed pages, and the detail needed to report completely one study conflicts with the desire of the journal to publish as many (worthy) studies as possible. As noted above, in research synthesis this issue arises most frequently when considering whether to publish the table of characteristics and results of individual studies. And, as one of our respondents suggested, sometimes even just the references to these studies can go on for pages.

Today the availability of the Internet eases this tension somewhat. Electronic publishing has largely removed page limitations from what

journals can publish. Still, too much length can be a problem if it leaves authors and readers feeling swamped with data that make it difficult to distinguish what is important and less important in the details of research. For this purpose, many (and an increasing number of) journals now provide auxiliary websites on which material can be placed rather than including it in the print version or the formal electronic version of the article. If the publisher does not make an auxiliary website available, some authors provide this information on their personal web pages and footnote its availability in the article. In electronic versions of articles, the supplemental information that resides on separate web pages can be linked to the article at the point in the report that they would otherwise appear.

It seems then that space limitations should no longer be a justification for the incomplete reporting of meta-analyses. However, when using auxiliary websites, another obligation arises: Authors must provide sufficient documentation accompanying the content of auxiliary websites so that readers can understand how this information is to be interpreted. For example, if meta-analysts want to share the coding sheet they used to retrieve information from studies, the coding sheets need to be completely labeled and contain the code book (which provides definitions and coding conventions) and the coding sheet itself. Similarly, a table (or spreadsheet) that contains the specific codes entered for each study needs to be accompanied by definitions for any abbreviations that are used in the table. Although this seems obvious and straightforward, many readers of web-based articles have experienced frustration when the auxiliary website contained information that was not interpretable. Such presentations do not meet the authors' obligation to report their methods and results thoroughly and clearly.

## Data Sharing and Meta-Analysis With Individual Participant Data and With Aggregate Statistics

At the beginning of this chapter, we defined meta-analysis as the use of aggregate data (AD) from previous research to conduct a research synthesis. Increasingly, meta-analyses are conducted by obtaining and cumulating individual participant data (IPD). Unlike meta-analyses based on AD, IPD meta-analysis involves the collection, checking, and reanalysis of the raw data from each study to obtain combined results.

Cooper and Patall (2009) examine the relative benefits of the two types of meta-analysis. They concluded that if both IPD and AD are equally available, meta-analysis using IPD is the superior approach: IPD meta-analysis permits (a) new analyses of the data, (b) checking of the data and

original analyses for errors, (c) addition of new information to the data sets, and (d) use of different statistical methods. However, Cooper and Patall also point out that because of the cost of IPD meta-analysis and the lack of available individual participant data sets, the best strategy currently is to use both approaches in a complementary fashion; the first step in conducting an IPD meta-analysis might be to conduct a meta-analysis with AD.

Three additional ethical issues become important when we consider the differences between meta-analysis with AD and IPD. These are the issues of data sharing, authorship, and the rights to confidentiality of the participants in the primary research.

With regard to data sharing, Cooper and Patall wrote:

> The incentives for data sharing are increasing while the barriers are coming down. Advances in data storage and ease of data transfer are barriers that have largely been removed. A recent incentive is the development and heightened enforcement of policies requiring or encouraging sharing of data collected with federal funding (National Institutes of Health, 2003). (Cooper & Patall, 2009, p. 174)

Rights to authorship is an issue related to data sharing. Often in medicine, where meta-analysis with IPD is undertaken much more frequently than in the social sciences, multiple primary researchers come together and form a consortium that collects and carries out the meta-analysis. In this case, the meta-analysis may be published under the joint authorship of the consortium with the individual contributors acknowledged in an author note. If such an arrangement is not possible, it is essential that the meta-analysts come to prior agreement with the collectors of the original data regarding how authorship will be handled.

Finally, the reuse of data in an IPD meta-analysis research project also raises issues about the right to confidentiality of the research participants. Here, the ethical issues are no different from those encountered for any secondary use of data. Guidelines covering these uses are fluid. Still, whether an individual's agreement to participate in the original study also made explicit or implied consent to have data included in a secondary analysis is a question that both the original collectors of the data and the IPD meta-analysts must answer. Making data sets available to IPD meta-analysts must occur only under the same rules of confidentiality that applied when the data were first collected. Typically, if the data are not shared until they have been stripped of any and all identifying information, then the investigation is no longer research with human subjects.[7]

[7] The Office for Human Research Protections (OHRP) of the Department of Health and Human Services document *Guidance on Research Involving Coded Private Information or Biological Specimens* states, "With respect to research involving private information and

## Uncovering Duplicate Publication

As a final issue, on occasion meta-analysts may find themselves acting as ethics enforcers. This occurs when in the course of gathering studies and extracting information from them, the meta-analysts identify instances in which researchers have engaged in *duplicate publication* (see point 8.13 in Table 16.1, and also Levin, Chapter 18, this volume).

Sometimes the line between what is and is not duplicate publication is clear. For example, a meta-analyst would not consider it duplicate publication if he or she came across a dissertation, convention paper, and journal article all presenting the same data (although in some fields papers presented at some meetings are considered publications). Likewise, it is certainly an ethical breach when two journal articles present the exact same data without acknowledgment of the earlier publication in the latter publication. The issue is not as clear between these extremes. Is it duplicate publication when two publications use the same data but conduct different analyses? What about publications that re-present the first wave of a longitudinal data collection, already published, in the article presenting the results of the second wave of data collection? An extended discussion of these issues is beyond our scope here, but it is important to make two points. Authors are ethically obligated to make readers (especially the reviewers who will judge the article for its original substantive contribution) aware when they re-present already-published data in subsequent articles. Meta-analysts are ethically obligated to alert the journals involved when they uncover what they consider to be duplicate publication.

## Conclusion

We began this chapter by suggesting, somewhat provocatively, that meta-analysts had it easy relative to primary researcher when it came to the ethical considerations surrounding their work. If readers did not view this assertion with skepticism when we first made it, we hope they do now. Ethical issues surrounding the reporting of methods and results are as, if not more, complex for meta-analysts than for primary researchers.

specimens, the exemption that is most frequently relevant is the exemption under HHS regulations at 45 CFR 46.101(b)(4): 'Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects' " (http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.htm).

Respecting the rights of research participants remains an issue for those meta-analysts using IPD. And more often than primary researchers, meta-analysts may find themselves in a position in which they must report ethical lapses on the part of others in the research community.

Finally, because meta-analysis is a relatively new technique, the standards of reporting are still evolving. The ambiguities in what is and is not important for readers to know make reporting decisions more difficult for research synthesists. We hope that the results of our survey and discussion of other ethical issues will make research synthesists' decisions a bit easier. Perhaps, the broad lesson is that ethical decisions in research—be it primary, secondary, or research synthesis—are never easy and never to be taken lightly.[8]

## References

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. New York: Author.

American Psychological Association. (2010). *Publication manual* (6th ed.). Washington, DC: Author.

APA Publication and Communication Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Brown, B. L., & Hedges, D. (2009). Use and misuse of quantitative methods: Data collection, calculation, and presentation. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics*. Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). New York: Academic Press.

Cook, D. J., Sackett, D. L., & Spitzer, W. O. (1995). Methodologic guidelines for systematic reviews of randomized control trails in health care from the Potsdam consultation on meta-analysis. *Journal of Clinical Epidemiology, 48,* 167–171.

Cooper, H., DeNeve, K., & Charlton, K. (2001). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,* 447–452.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis using individual participant data and aggregate data. *Psychological Methods, 14,* 165–176.

[8] Author note: The authors thank the members of the Society for Research Synthesis Methodology for their participation in the survey reported in this chapter. Correspondence can be sent to Harris Cooper, Department of Psychology & Neuroscience, Box 90086, Duke University, Durham, NC 27708-0086, or cooperh@duke.edu

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Merriam Webster Online. (2007). *Merriam Webster online*. Retrieved from http://www.merriam-webster.com/dictionary

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D., for the QUOROM group. (1999). Improving the quality of reporting of meta-analysis of randomized controlled trials: the QUOROM statement. *Lancet, 354,* 1896–1900.

Moher, D., Tetzlaff, J., Liberati, A., Altman, D. G., & the PRISMA group. (2009). Preferred reporting items for systematic reviews and meta-analysis: The PRISMA statement. *PLoS Medicine, 6*(7): e1000097.

National Institutes of Health. (2003). Final NIH statement on sharing research data. Retrieved from http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* Oxford, UK: Wiley.

Schmidt, F. L., On, I., & Hayes, T. (2009). Fixed vs. random models in meta-analysis: Model properties and comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62,* 97–128.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., … Thancker, S. B. (2000). Meta-analysis of observational studies in epidemiology. *Journal of the American Medical Association, 283,* 2008–2012.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13,* 130–149.

# 17

# *Ethics and Statistical Reform: Lessons From Medicine*

**Fiona Fidler**
*La Trobe University*

In psychology, *null hypothesis significance testing* (NHST; Cumming & Fidler, Chapter 11, this volume) and *meta-analysis* (MA) (Cooper & Dent, Chapter 16, this volume) have occupied advocates of *statistical reform* for decades. Hundreds of psychology journal articles criticize the former and encourage more widespread use of the latter. In medicine, NHST has similarly been admonished and MA promoted. Misuse and misinterpretation of NHST have been widespread in both disciplines. The alternative statistical practices advocated by reformers have been the same in both disciplines, too—estimation (*effect sizes* and *confidence intervals* [CIs]) and, increasingly, MA.

Despite these similarities between the disciplines, changes to statistical practice have been much slower in psychology than in medicine. For example, in 2006, 97% of articles in 10 leading psychology journals still reported NHST as the primary outcome (Cumming et al., 2007). In medicine, by contrast, CIs replaced NHST as the dominant analysis in individual studies in the mid-1980s (Fidler, Thomason, Cumming, Finch, & Leeman, 2004), and they remain a routine feature, being reported in approximately 85% of empirical articles (Cumming, Williams, & Fidler, 2004). In medicine, editorial policy in leading journals (e.g., *BMJ, The Lancet*) now requires that all new trials be placed in the context of previous research and integrated using MA (Young & Horton, 2005). Systematically placing new empirical results in the context of existing quantitative data is far from routine practice in psychology, although MA is certainly increasing.

The dramatic shift from rare to routine reporting of CIs in medical journals in the 1980s was supported by strict individual editorial policies and the institutional support of the International Committee of Medical Journal Editors (Fidler et al., 2004). I have argued elsewhere (Fidler, 2005) that the relative success of medicine's statistical reform occurred partly because medicine framed these statistical issues in ethical terms. Psychologists

and other behavioral scientists, on the other hand, presented mainly only technical and philosophical reasons for the advocated change. *Statistical power*, effect sizes, CIs, and other reform statistics were no longer merely technical issues to be worked out on a calculator or in an analysis software package or relegated to the consultant brought in after data collection, nor were they merely philosophical problems about the nature of evidence or the interpretation of probability. Rather, statistical reform was a practical and ethical concern, with obvious and tangible consequences, for every researcher, statistician or not. In psychology, this context has been largely lacking, the current edited volume being an obvious exception.

In this chapter, I first explicate how an ethical imperative was explicitly used in medicine to discourage NHST and to encourage MA. Next, I discuss two case examples from medicine that have been used to illustrate to practitioners why misuse of these techniques has clear ethical implications. I then provide two parallel examples from psychology that have similar—although comparatively underappreciated—ethical implications. Finally, I discuss reasons why an ethical imperative has been, to date, used in medicine but not psychology, why this is a problem, and how it can be remedied. In so doing, this chapter addresses several questions: Why were the ethical implications of statistical practice so salient to medical reformers but not psychological ones? What gains, in terms of statistical reform, did an ethical imperative afford medicine? What lessons can psychology learn from medicine's reform efforts, as well as from its mistakes?

## In Medicine, Statistical Inference Is an Ethical Concern

One of the main criticisms of typical NHST practice in both medicine and psychology has, over the decades, been the neglect of statistical power. Calls for increased attention to statistical power have been the focus of hundreds of articles in both disciplines. The type of arguments used to promote statistical power, however, provides one of the clearest demonstrations of the differences between the disciplines.

In medicine, neglect of statistical power was identified as an ethical problem from early in the reform process. This is evident in the medical literature of the 1970s, as the following quotations demonstrate:

> One of the most serious ethical problems in clinical research is that of placing subjects at risk of injury, discomfort, or inconvenience in experiments where there are too few subjects for valid results. (May, 1975, p. 23)

> Not every clinician—or even his ethical committee—is acutely attuned to the details of statistical Type II errors. (Newell, 1978, p. 534)

In psychology's reform literature, by contrast, an ethical argument for statistical power has rarely been made explicit. Instead, we have seen analysis of the reporting rates of power (e.g., Fidler et al., 2005; Finch, Cumming, & Thomason, 2001); calculations of the average power of research (e.g., Cohen, 1962; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989); studies of misconceptions about power and sample size (starting with the law of small numbers; Haller & Krauss, 2002; Oakes, 1986; Tversky & Kahneman, 1971); technical explanations of statistical power; and finally philosophical explanations for the neglect of power (e.g., various authors refer to NHST as an incoherent amalgamation of Fisher and Neyman–Pearson, most notably Gigerenzer, 1993). All of these discussions are important in their own right, but none necessarily deals with the ethics of our current practice of neglecting Type II errors.

The ethical framing of this issue within the medical discipline was not an afterthought, nor was it a last-ditch rhetorical effort—rather, it was the primary impetus for statistical reform 3 decades ago. Altman (1982a) explains why:

> A study with an overly large sample may be deemed unethical through the unnecessary involvement of extra subjects and correspondingly increased costs. Such studies are probably rare. On the other hand, a study with a sample size that is too small will be unable to detect clinically important effects. Such a study may thus be scientifically useless, and hence unethical in its use of subjects and other resources. (Altman, 1982a, p. 6)

In the following quotation, Altman (1982b) spells out the consequences of neglecting statistical power:

> (1) The misuse of patients by exposing them to unjustified risk and inconvenience; (2) the misuse of resources, including the researchers' time, which could be better employed on more valuable activities; and (3) the consequences of publishing misleading results, which may include the carrying out of unnecessary further work. (Altman, 1982b, p. 1)

In medicine, particular cases in which flawed statistical practice continued, such as the ongoing neglect of statistical power or lack of attention to effect sizes, became scandals. The attention given to the neglect of statistical power was not seen as statistical nitpicking, but rather as justified criticism of professional misconduct. Here, Altman (1994) encourages researchers to be outraged when they come across misuse of statistics:

> What should we think about a doctor who uses the wrong treatment, either willfully or through ignorance, or who uses the right treatment wrongly (such as by giving the wrong dose of a drug)? Most people would agree that such behaviour was unprofessional, arguably unethical, and certainly unacceptable. What, then, should we think about researchers who use the wrong techniques (either willfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled.… This is surely a scandal. (Altman, 1994, p. 283)

## In Medicine, Meta-Analysis Is an Ethical Concern

The ethical imperative for MA was also explicit in medicine, and its neglect also identified as wasting valuable research time and resources (e.g., the title of the article this quotation is from is "The Scandalous Failure of Science to Cumulate Evidence Scientifically"):

> New research should not be designed or implemented without first assessing systematically what is known from existing research.… The failure to conduct that assessment represents a lack of scientific self-discipline that results in an inexcusable waste of public resources. In applied fields like health care, failure to prepare scientifically defensible reviews of relevant animal and human data results not only in wasted resources but also in unnecessary suffering and premature death. (Chalmers, 2005, p. 229)

In 2005, *The Lancet* made MAs a requirement—new trials must be put in the context of previous research. This innovation was also justified on ethical grounds, with this explicit statement that continuing trials without conducting an MA is both unscientific and unethical:

> The relation between existing and new evidence should be illustrated to an existing systematic review or meta-analysis. When a systematic review or meta-analysis does not exist, authors are encouraged to do their own.… Those who say systematic reviews and meta-analysis are not "proper research" are wrong; it is clinical trials done in the absence of such reviews and meta-analysis that are improper, scientifically and ethically. (Young & Horton, 2005, p. 107)

Perhaps the best example of institutional acceptance of MA in medicine is the Cochrane Collaboration (http://www.cochrane.org), which is dedicated solely to conducting MAs of clinical trials to improve health care.

The Cochrane Collaboration was established in 1993 and has since published thousands of MAs and has clinical trial centers around the world. The Collaboration itself grew out of an ethical concern. Archie Cochrane's (1972) *Effectiveness and Efficiency* laid out the basic principle: Because health care resources would always be limited, the only ethical system was one that practiced only those treatments for which evidence had accrued from systematic, rigorous evaluation. Five years later, Cochrane (1979) laid the final challenge with this now famous quotation: "It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials" (p. 1). These words became a rallying point at the foundation of the Cochrane Collaboration (Chalmers, 2006).

A promising social science parallel, the Campbell Collaboration, began in 1999 (http://www.campbellcollaboration.org). It grew out of the Cochrane Collaboration and shares the goal of increased efficiency through evidence-based decision making. The Campbell Collaboration specializes in meta-analytically reviewing evidence related to education, crime, justice, and social welfare. Unfortunately, this still leaves a lot of clinical and experimental psychology territory uncovered. Another recent MA development in psychology is the Meta-Analytic Reporting Standards (MARS) section in the sixth edition of the American Psychological Association (APA) *Publication Manual* (2010) (see Cooper & Dent, Chapter 16, this volume). However, despite its excellent content, MARS is a mere appendix to the *Manual* and may be easily missed by the casual reader.

In the next section, I outline examples of studies where misinterpretations of NHST have led medical research astray and resulted in both a waste of resources and unnecessary suffering. These case studies also illustrate the importance of cumulative MA in sorting out the confusion left by dichotomous accept–reject decisions made in single experiments. In a subsequent section, I show that parallel examples in psychology also exist but have been less publicized and thus far had less impact on the statistical reform of the discipline.

## Two Medical Case Examples

### Medical Case Example 1: Myocardial Infarction and Streptokinase

Streptokinase is an enzyme that dissolves vascular thrombi, or blood clots caused by atherosclerosis. In the 1950s, medical researchers began to wonder whether it might benefit acute myocardial patients because most cardiac arrests are caused by atherosclerosis—a gradual buildup of a fat-containing substance in plaques that then rupture and form blood

clots on artery walls. Between 1959 and 1988, 33 randomized clinical trials tested the effectiveness of intravenous streptokinase for treating acute myocardial infarction. The majority of these trials (26 of 33) showed no statistically significant improvement at $p < .05$. However, the remaining trials did show a statistically significant improvement, and often a dramatic one. Those improvements were enough to motivate testing to continue, in pursuit of a definitive answer.

If one looks at the results of these trials as CIs, rather than as simply statistically significant or not, it is immediately obvious that those nonsignificant trials have extremely wide CIs. (An excellent graphic can be found in Lau et al., 1992, reprinted with copyright permission in Hunt, 1997.) The CIs of the statistically nonsignificant trials do indeed capture the odds ratio of 1—but they also capture almost every other value on the scale! Wide intervals are an immediate sign that the nonsignificant trials had low statistical power. The seemingly inconsistent results were a simple product of relative power of the trials. The high-powered trials produced statistically significant results; the low-powered trials (in this case, those with small sample sizes) did not.

In 1992, Lau et al. demonstrated that cumulative odds ratio (i.e., the odds ratio produced by an MA after the first two trials, another after the first three trials, and so on) was consistently greater than 1 by the time the fourth clinical trial was added and that the CI around this odds ratio did not capture 1 from the time the seventh clinical trial was added. Recall that there were 33 clinical trials—this result means that there were at least 26 more than there should have been! Overreliance on dichotomous accept–reject decisions from NHST, as well as neglect of statistical power, resulted in the unnecessary testing of 30,000 additional patients over an extra 15 years, half of whom were in the placebo group and therefore denied a treatment already proven to be effective. Presenting the results of individual trials with CIs—and placing those individual trials in the context of previous research by using cumulative MA—would have clearly shown that evidence in favor of the drug was indisputable and that additional subjects and years of further research were redundant.

In a separate publication of the same year, the same team of researchers who demonstrated and emphasized the unethical failings of the above study (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992) presented a comparison of textbook advice on the treatment of people with myocardial infarction and the results of several cumulative MAs. In each case, they showed that advice on lifesaving treatments had been delayed for more than a decade, and, in some, that harmful interventions were promoted long after evidence of their damage had accumulated.

The reports of these researchers on treatment for myocardial infarction have been identified in the medical literature as a "great impetus"

in the widespread recognition of the practical and ethical importance of unbiased, quality scientific reviews: [They] "made it abundantly clear that the failure of researchers to prepare reviews of therapeutic research systematically could have very real human costs" (Chalmers, Hedges, & Cooper, 2002, p. 21). In the same year, the Cochrane Collaboration was born.[1]

### Medical Case Example 2: Antiarrhythmic Drugs

Antiarrhythmic drugs suppress the fast rhythms of the heart and were often prescribed after a cardiac arrest to prolong life. Many clinical trials assessed their safety, and although results were somewhat mixed, the accepted conclusion on an individual study basis was that there was, at worst, no difference in the mortality rate when the drugs were prescribed. This conclusion of "no difference" unfortunately turned out to be an overly optimistic interpretation of the research when an MA was performed. In 1993, an MA was carried out, examining 51 trials of Class I prophylactic antiarrhythmic agents conducted on 23,229 patients (Teo, Yusif, & Furberg, 1993).[2] The results clearly showed a substantially increased mortality rate as a result of the drug in question. Within the drug group there were 660 deaths (5.63% of patients) as opposed to 571 deaths in placebo groups (4.96% of patients).

The ethical implications of this case were swiftly picked up by commentators in the medical literature. Most famously, Moore (1995) commented that the number of deaths from these antiarrhythmic drugs at the peak of their use (in the late 1980s) was comparable with the number of Americans who died in the Vietnam War. Chalmers also discussed the ethics of this case and again argued explicitly that cumulative MA could have saved these lives (e.g., Chalmers, 2005). These comments lent timely support to the argument for the practical importance of cumulative MA, established by the myocardial case above.

---

[1] The other important development in the establishment of the Cochrane Collaboration was a large-scale synthesis of studies relating to pregnancy and childbirth. Chalmers, who was the lead author of the report and later became the founding leader of the Cochrane Collaboration, explained: "International collaboration during this time [the late 1980s] led to the preparation of hundreds of systematic reviews of controlled trials relevant to the care of women during pregnancy and childbirth. These were published in a 1,500-page, two-volume book, *Effective Care in Pregnancy and Childbirth* (Chalmers, Enkin, & Keirse, 1989), deemed an important landmark in the history of controlled trials and research synthesis (Cochrane, 1989; Mosteller, 1993)." (Chalmers, Hedges, & Cooper, 2002).

[2] The meta-analysis also looked at other types of agents, including beta blockers or Class II agents (55 trials), amiodarone or Class III agents (8 trials), or calcium channel blockers or Class IV agents (24 trials), but it is the results of Class I agents that are of particular interest here.

## Comparing Medical Versus Psychology Case Examples

One could be reasonably skeptical about whether such strong linkages between statistical practice and ethics could be forged in psychology, as were achieved in the aforementioned two medical case examples. There may be complicated political and financial pressures associated with pharmaceutical trials that do not apply in psychological research. One could argue, for example, that medical statistics have to be stricter than psychology statistics to prevent unethical behavior by Big Pharma. However, not all cases of cumulative MA from medicine are drug trials. A more recent cumulative MA showed that advice on infant sleeping position, namely, to put newborns to sleep on their backs, was delayed for decades because of a failure to properly assess cumulative results from individual trials.

Advice to put infants to sleep on their fronts for nearly half a century was contrary to evidence available from 1970 that this was likely to be harmful. Systematic review of preventable risk factors for sudden infant death syndrome (SIDS) from 1970 would have led to earlier recognition of the risks of sleeping on the front and might have prevented more than 10,000 infant deaths in the United Kingdom and at least 50,000 in Europe, the United States, and Australasia (Gilbert, Salanti, Harden, & See, 2005).

Gilbert et al. (2005) make practical and ethical implications clear by discussing the consequences of the statistical error in number of thousands of infant deaths—and yet there is no pharmaceutical conspiracy to blame in this case.

However, an extensive search of the reform literature in psychology for equivalently explicit ethical discussions of statistical power and/or misuse of NHST, as well as for explicit framing of consequences of errors in practical terms of damage done (e.g., lives lost, resources wasted, or equivalent harm), was not fruitful. Although some articles infer ethical consequences of poor statistical practice (e.g., Meehl's 1978 article is perhaps the best example), none primarily emphasizes the ethical dimension of these problems as do the case examples in the medical literature. When we hear that psychologists misuse statistical techniques or misinterpret their results, are we "appalled," as Altman urged medical researchers to be? Does a scandal erupt, as it does in the medical literature? Rarely.

And yet, it is possible to find similar (albeit, less publicized) examples of irresponsible statistical analysis from psychology with important ethical implications. Indeed, I present two examples in the following section. These cases demonstrate that it is not simply the medical field's political and financial pressure that leads to ethically concerning statistical practice. The consequences of the poor practice remain an ethical concern, regardless of the discipline. I propose that the fact these psychological cases are comparatively unknown and have not been identified in the

psychology literature, or even in the reform literature, is in itself a scandal of alarming proportions. Additionally, their lack of publicity represents a missed opportunity for psychologists to add an ethical imperative to statistical reform.

## Two Psychology Case Examples

### Psychology Case Example 1: Employment Testing and the Theory of Situational Specificity

Schmidt claimed, as APA Division 5 President in 1996, that "reliance on statistical significance testing … has systematically retarded the growth of cumulative knowledge in psychology" (p. 115). Schmidt's dramatic quotation is now famous. However, I suspect the evidence for it is less well known, much less regarded as a scandal. Hunter and Schmidt (2004) themselves obviously consider the theory of situational specificity (TSS) scandalous (as do I), but it has rarely been advertised as such, even in the statistical reform literature. The evidence is a series of MAs that Schmidt, Hunter, and others conducted throughout the 1970s and early 1980s. These MAs exposed a great flaw in the then orthodox doctrine of organizational psychology.

The TSS pertains to employment tests—professionally developed cognitive ability and aptitude tests that are designed to predict job performance. The theory holds that the correlation between test score and job performance does *not* have general validity; that is, "a test valid for a job in one organization or setting may be invalid for the same job in another organization or setting" (Schmidt & Hunter, 1981, p. 1132). The validity of the tests, it seemed, depended on more than just the listed tasks for a given position description. The theory proposed that the validity of any one test depended on the cognitive information-processing and problem-solving demands of the job and perhaps even the social and political demands of the workplace. In other words, TSS proposed that there is a distinct context for each job, and that a general employment test may not predict specific job context performance.

How did the TSS come about? The belief in "situational specificity" grew out of the considerable variability observed from study to study, even when the jobs and/or tests were similar. Some studies found statistically significant correlations, whereas others found none. "Situational specificity" explained the inconsistency in the statistical significance of empirical results by generating potential moderating variables. Another obvious factor that could also explain why one study found a statistically significant result and another study did not was the varied and usually

low statistical power of the studies. This alternative explanation, however, went unnoticed for several decades.

The TSS grew structurally complex, with addition of many potential moderating variables, including organization size, gender, race, job level, and geographic location. In fact, the search for such moderating variables became the main business of industrial or organizational psychology for decades. Researchers sought to shed further light on the "specific" nuances of the theory, despite the fact that the variability that they were working to explain was illusory. Not until 1981, when Hunter, Schmidt, and their colleagues carried out an MA using the results of 406 previous studies, did it finally become clear that the difference in allegedly inconsistent results could be exclusively accounted for by the low statistical power of the studies.

> If the true validity for a given test is constant at .45 in a series of jobs … and if sample size is 68 (the median over 406 published validity studies…) then the test will be reported to be valid 54% of the time and invalid 46% of the time (two tailed test, $p = .05$). This is the kind of variability that was the basis for theory of situation-specific validity. (Schmidt & Hunter, 1981, p. 1132)

As Schmidt and Hunter finally revealed, the reporting of individual results as "significant" or "nonsignificant" had created the illusion of inconsistency, even though almost all the obtained effect sizes were in the same direction.

How long did organizational psychology pursue this misdirected theory and its associated research program? In 1981, toward the end of their MA series, Hunter and Schmidt wrote: "the real meaning of 70 years of cumulative research on employment testing was not apparent [until now]" (p. 1134). Of the use of NHST in this program, they wrote: "The use of significance tests within individual studies only clouded discussion because narrative reviewers falsely believed that significance tests could be relied on to give correct decisions about single studies" (p. 1134).

The case of the TSS provides evidence that NHST, as typically used—with little regard for statistical power and overreliance on dichotomous decisions—can seriously damage scientific progress. In this case, an important research program was led astray by a search for moderating variables to explain illusory differences. Years of empirical data were seen to support a theory, for which there was, in fact, no empirical evidence. A program of this scale going astray represents an enormous waste of public funds and scientific resources, including person hours, dollars, careers, and other research not conducted at the expense of this program and/or because the real findings were obscured. These losses are themselves serious ethical concerns. However, Schmidt and Hunter (1981) also hint at another, perhaps more disturbing, level of damage:

> Tests have been used in making employment decisions in the United States for over 50 years.… In the middle and late 1960s certain theories about aptitude and ability tests formed the basis for most discussion of employee selection issues, and in part, the basis for practice in personnel psychology.… We now have.… evidence.… that the earlier theories were false. (pp. 1128–1129)

In other words, the false TSS findings influenced the success of companies that relied on it (including missing out on potentially valuable employees who were rejected on the basis of test results, and vice versa), not to mention the careers of uncounted jobseekers. Despite the disturbing ethical implications of Schmidt and Hunter's findings, their debunking of the TSS failed to motivate widespread statistical reform regarding MA in psychology. Unlike the parallel cases in medicine—whose ethical implications helped launch the Cochrane Collaboration—this less publicized scandal about employment test validity had strikingly little impact on statistical practice in psychology.

### Psychology Case Example 2: Learned Helplessness and Depression

The second psychology case example concerns *learned helplessness,* a concept pioneered by Seligman. The phenomenon was first isolated in dogs (Seligman, Maier, & Geer, 1968), much in the tradition of Pavlov. Caged dogs were given random electric shocks from which they could not escape. Later they were placed in different cages with separate compartments that they could use to escape from the shocks. They were again administered shocks. Surprisingly, around two thirds of the 150 dogs did not try to escape. They remained in the shock compartment and did not attempt to move. Seligman concluded that the dogs had learned that they were helpless. Immediately, Seligman and his colleagues began to wonder what links learned helplessness (or pessimistic explanatory style) might have with depression and illness.

Throughout the 1970s and early 1980s, strong links between pessimistic explanatory style–learned helplessness and depression and illness soon were made. For example, the effects of helplessness on growth of cancerous tumors and death rates were first observed in rats, and later experiments demonstrated the links in human subjects. Seligman and his colleagues published at least 25 articles on the topic between 1969 and 1977, as well as a book *Helplessness: On Depression, Development and Death* (Seligman, 1975). However, other researchers had trouble replicating the experimental results linking explanatory style to depression—or rather, they had trouble replicating the statistical significance of the results (see the 1978 special issue of the *Journal of Abnormal Psychology*). Eventually, several MAs showed that the inconsistencies in the literature were an artefact of NHST.

Specifically, an MA by Sweeney, Anderson, and Bailey (1986) combined 104 studies, excluding those from Seligman's lab, and for the first time found results consistent with Seligman's. Second, a series of statistical power analyses by Robins (1988) pointed out that only 8 of 87 previous individual studies on depression and explanatory style (or "attributions" as Robins calls them) had an a priori power of .80 or better for detecting the small population effect. Robins explained that the situation was so poor that "even adopting the assumption of a larger true effect, which I term *medium* (e.g., $r = .30$), only 35 of the 87 analyses had the desired chance of finding such an effect" (p. 885).

In sum, the misinterpretation of statistically nonsignificant results produced by underpowered learned helplessness studies caused several decades of ongoing debate and confusion where there should have been none. The academic damage in this instance could have been that a valid theory was lost for all time. Still, for at least a decade, important theoretical developments and clinical interventions based on relationships between learned helplessness, explanatory style, depression, and illness were delayed. And yet, no media scandal resulted, as likely would have been the case in the medical literature. No discussion of the ethics of statistical inference ensued, as also would have likely been the case in the medical literature.

## Why Medicine and Psychology Approach Statistical Reform Differently

Why is there such a stark contrast in the way the two disciplines have dealt with the misapplication of NHST? The following section outlines three hypotheses that may explain the difference between medicine and psychology. I stress that these hypotheses may explain the difference in responsiveness to statistical reform in these two disciplines. To put it another way, belief in these three hypotheses is perhaps sufficiently widespread to have impeded statistical reform in psychology. I am *not* arguing that there is a difference in the need for ethical statistical practice in the two disciplines. On the contrary, my aim is to demonstrate that ethical practice is equally important in both.

1. *The proximity of experimental outcomes to utilitarian consequences*. In medicine, trials are usually designed with some particular utilitarian outcome in mind—that is, to test whether a certain specific intervention improves health care in some particular way. By contrast, many psychological trials are designed with the purpose

of improving our understanding of how the mind works, rather than whether one particular intervention improves its function. There are utilitarian ethical arguments to be made here, too, of course, but the ethical consequences of "our theory is wrong" are considerably different than the ethical implications of "our treatment doesn't work" or "our drug causes harm." (There are exceptions in both disciplines, of course.)

2. *The stakes.* In medicine, the stakes can be life or death. Perhaps more often medical studies offer opportunities to enhance health care—for example, through injury prevention, minor improvements to quality of life, and decreases in hospital visits or length of stay. Although not as high, these stakes are still tangible, and medical trials are directed toward measuring these precise outcomes. In psychology, too, there may well be high stakes—opportunities to implement clinical, developmental, or educational interventions, and studies with implications for legal decisions in areas such as child custody or employment—but these outcomes are usually several steps removed from the health results measured by medical trials and experiments.
3. In medicine, distrust of pharmaceutical companies and their motives has led to increased vigilance and helped create a healthy skepticism. In psychology, there is rarely a big, special interest company to blame or substantial potential financial gain by pursuing fallacious results. As a result, it is perhaps more difficult to get a handle on why the statistical errors in psychology should be conceptualized as ethical, as well as academic, concerns.

## Costs to Psychology of Not Ethically Motivating Statistical Reform

Whatever the reason for the difference, it is difficult to deny that psychology would be better off if statistical inference was an ethical concern and not just a technical one. There are *costs to science*, as well as costs beyond science, in making ongoing resistance to statistical reform an increasing ethical concern. Below I list some of these costs; there are no doubt others I have not listed.

Costs to science of overreliance on NHST, and neglect of statistical power and MA:

- Research programs may go astray while attempting to explain illusory variability in results (e.g., the employment testing case)

- Unnecessary, prolonged debate; delayed progress and implementation of interventions (e.g., the learned helpless case)
- Time and resources wasted on incorrect, weak, or trivial research programs that happen to turn up statistically significant results
- Potentially useful research programs or directions come to an end because of the inability to produce "consistent" (e.g., statistically significant) results
- Alternatively, research programs may never get started because of their inability to jump the statistical significance hurdle

Costs to public welfare beyond scientific knowledge itself:

- The delayed release of useful interventions and applications (e.g., those based on understanding links between illness, depression, and learned helplessness)
- The implementation of interventions that have little or no impact (in cases where statistical significance has been achieved by overpowered experiments)
- The implementation of harmful interventions because the adverse effects are not detected in low-powered studies
- The implementation of interventions based on misguided theory (e.g., workplace-specific employment tests based on the TSS)
- Various economic costs, including running extra studies in search of statistical clarification when research resources could be better used elsewhere

## Conclusion

Thus far, I have discussed the swifter improvements in statistical practice that accompanied the use of an ethical imperative in the medical field. This discussion can be consolidated into the following three main lessons for psychology, from medicine.

### Lesson 1: Statistical Reform Needs to Be Ethically, and Technically and Philosophically Motivated

Despite struggling with reform debates for an extra 2 decades, psychology still relies almost exclusively on NHST. Repetition of the technical and

philosophical arguments has done little to motivate change, but psychological researchers may well respond to ethical arguments for statistical reform, as did medical professionals.

### Lesson 2: Changes in Statistical Reporting Are Just the First Step; Thinking and Interpretation Need to Change, Too

Thus far I have argued that medicine has improved its practice by treating statistical practice as an ethical issue. I now turn to the more subtle distinction between statistical reporting practice (which medicine has improved dramatically) and *statistical thinking* and interpretation (which has changed far less). Reporting practices in medical journals have changed to be sure, but here is a lesson that psychology can learn from what medicine did not do! Despite the dramatic increase in CI reporting in medical journals, there was little change the way researchers interpret and discuss their findings.

In our own survey of medical journals, we found many articles that did not include any *p* values but still had discussions focused on "statistical significance" (Fidler et al., 2004). Savitz, Tolo, and Poole (1994) also found this in their survey of the *American Journal of Epidemiology*: "The most common practice was to provide confidence intervals in results tables and to emphasize statistical significance tests in result text" (p. 1047). Poole lamented the fact that change in reporting had not led to a change in thinking:

> The reporting of confidence intervals really hasn't changed the way people think: 99% of the people that now report CIs, 20 years ago would have reported *p* values or asterisks, or *s* and *ns* and they aren't thinking differently to that. They have this vague idea that they are reporting more information with CIs, because they read that somewhere in something Ken Rothman [who made editorial changes at *AJPH* and *Epidemiology*] wrote. But basically they are only reporting CIs because Rothman was an authority figure, and he and others encouraged them—well, his journals insisted on it. (Poole, personal communication, September 2001)

Psychology has the chance to make a substantial reform, one that involves changes in the way researchers approach analysis, and interpret and think about data, as well as what they report in the tables and figures of their journal articles. Substantial reform requires cognitive change and empirical evidence about which presentations of statistics communicate most clearly. Psychology itself is perfectly positioned to collect such empirical data through research into statistical cognition and to advocate for evidence-based reform.

### Lesson 3: Statistical Reform Should Be Integrated: Estimation and *Cumulative Meta-Analysis* Go Hand in Hand

Importantly, all four of the case studies given above feature MA. It is one of the best tools available for telling us when there is enough research on a topic for us to stop throwing resources at it. MAs, with their superior power and ability to pin down effect sizes, can help emphasize thinking in terms of estimation rather than hypothesis testing—they stop us from falling into the trap of trying to "explain" the difference between "inconsistent" results of NHST studies. In medicine, editorial policies instituting CIs were a phenomenon of the 1980s, whereas policies about cumulative research and MA came decades later (e.g., 2005 in *The Lancet*). In psychology we should aim to make the shift to CIs and estimation inseparable from cumulative MA: CIs and estimation should be reported as the primary outcome of individual studies, and cumulative MAs should be updated with each new study. The ethical advantages of the two practices in combination are great: CIs make the uncertainty of each trial explicit, and MAs help guard against unnecessary further studies when a question has been adequately answered.[3]

## References

Altman, D. G. (1982a). How large is a sample? In D. G. Altman & S. M. Gore (Eds.), *Statistics in Practice* (pp. 6–8). London: BMJ Books.

Altman, D. G. (1982b). Misuse of statistics is unethical. In D. G. Altman & S. M. Gore (Eds.), *Statistics in Practice* (pp. 1–2). London: BMJ Books.

Altman, D. G. (1994). The scandal of poor medical research. *British Medical Journal, 308,* 283–284.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Journal of the American Medical Association, 268,* 240–248.

Chalmers, I. (2005). The scandalous failure of science to cumulate evidence scientifically. *Clinical Trials, 2,* 229–231.

Chalmers, I. (2006). Archie Cochrane (1909–1988). The James Lind library. Retrieved from http://www.jameslindlibrary.org

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25,* 12–37.

[3] Author note: This research was supported by the Australian Research Council.

Chalmers, I., Enkin, M., & Keirse, M. J. N. C. (Eds.). (1989). *Effective care in pregnancy and childbirth*. Oxford, UK: Oxford University Press.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services.* London: Nuffield Provincial Hospitals Trust.

Cochrane, A. L. (1979). 1931–1971: A critical review, with particular reference to the medical profession. In G. Feeling-Smith & N. Wells (Eds.), *Medicines for the year 2000*. London: Office of Health Economics.

Cochrane, A. L. (1989). Foreword. In I. Chalmers, M. Enkin, & M. J. N. C. Keirse (Eds.), *Effective care in pregnancy and childbirth*. Oxford, UK: Oxford University Press.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., … Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology*. (Unpublished doctoral dissertation). University of Melbourne, Australia.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., … Schmitt, R. (2005). Evaluating the effectiveness of editorial policy to improve statistical practice: The case of the *Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology. 73,* 136–143.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15,* 119–126.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gilbert, R., Salanti, G., Harden, M., & See, S. (2005). Infant sleeping position and the sudden infant death syndrome: Systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International Journal of Epidemiology, 34,* 874–887.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1–20.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed). Thousand Oaks, CA: Sage.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine, 327,* 248–254.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

May, W. W. (1975). The composition and function of ethical committees. *Journal of Medical Ethics, 1,* 23–29.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Moore, T. (1995). *Deadly medicine*. New York: Simon and Schuster.

Newell, D. J. (1978). Type II errors and ethics. *British Medical Journal, 5,* 534–535.

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: J. Wiley & Sons, Inc.

Robins, C. J. (1988). Attributions and depression: Why is the literature so inconsistent? *Journal of Personality and Social Psychology, 54,* 880–889.

Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.

Savitz, D. A., Tolo, K., & Poole, C. (1994). Statistical significance testing in the *American Journal of Epidemiology*, 1970–1990. *American Journal of Epidemiology, 139,* 1047–1052.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–315.

Seligman, M. E. P. (1975). *Helplessness: On depression, development, and death*. New York: W. H. Freeman.

Seligman, M. E. P. (1990). *Learned optimism.* New York: Knopf.

Seligman, M. E. P., Maier, S. F., & Geer, J. (1968). The alleviation of learned helplessness in dogs. *Journal of Abnormal Psychology, 73,* 256–262.

Sweeney, P. D., Anderson, K., & Bailey, S. (1986). Attributional style in depression: A meta-analytic review. *Journal of Personality and Social Psychology, 50,* 974–991.

Teo, K. K., Yusif, S., & Furberg, C. F. (1993). Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction: An overview of results from randomized controlled trials. *Journal of the American Medical Association, 270,* 1589–1595.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.

Young, C., & Horton, R. (2005). Health module page. *The Lancet, 366*, 107.

# 18

# *Ethical Issues in Professional Research, Writing, and Publishing*

**Joel R. Levin**
*University of Arizona*

In the final chapter of this impressive volume, I consider the professional research, writing, and publication process, primarily within the field of psychology, focusing on a host of ethical (mal)practices in the process. Such (mal)practices will be framed in the context of several actual-case examples that I have encountered over the years, initially as a young researcher, then as a journal reviewer, editorial board member, journal associate editor, and editor, and most recently as Chief Editorial Advisor for the Publications Office of the American Psychological Association (APA). It is hoped that such examples, which relate directly to various ethical principles that can be found in the most recent edition of APA's *Publication Manual* (APA, 2010), will help readers appreciate the relevance of these issues as they engage in their own research and publishing activities. In addition, in the final section of the chapter I offer a set of summary recommendations that follow from the actual-case examples.

## A Personal Introduction to Research Ethics

Early on in my academic career, I acquired some lifelong lessons about conducting, analyzing, and reporting research. The overarching theme of these lessons was "integrity" of various kinds. My story begins as a graduate student in educational psychology at the University of California, Berkeley, at which time I was fortunate to have apprenticed with two outstanding research mentors, Bill Rohwer and Len Marascuilo.

### Experimental Rigor

Rohwer, originally trained as an experimental psychologist who studied theoretical issues of learning, memory, and cognition, had turned his attention to analogous problems in the field of education. From him I received comprehensive tutelage in the appropriate methods and procedures that characterize "scientifically credible" (Levin, 1992, 1994) psychology research, including rigorous experimental designs that invariably incorporated *to the letter* Campbell and Stanley's (1966) internal validity standards.

### Analytic Appropriateness and Accuracy

Marascuilo, originally trained as a biostatistician, was applying his trade to quantitative methods in education, including the development of statistical tests for analyzing different types of data with varying distributional assumptions. Under his instruction, Cook and Campbell's (1979) notions of statistical conclusion validity—manifesting itself to combat what I later termed statistical "bugs" (Levin, 1985)—were drummed into my head daily. Included in such drumming was that data analyses in manuscripts submitted for publication need to be carefully checked, rechecked, and rechecked again with respect to their accuracy and validity.

Around that time as well, I came to believe in the sanctity of the publication process, which included the holiness of each written word and data fragment that graced the pages of our professional journals. This also included an introduction to the heretofore foreign (to me) concept of a journal "hierarchy" within each academic discipline and subdiscipline. That is, I learned that within each field of specialization journals could be ordered with respect to their prestige, visibility, and even the believability of the results reported therein—somewhat akin to what nowadays can be quantified as a journal's "impact factor." Without telling tales out of school, I was advised that in academic departments at most "research" universities, publishing in certain journals would be "worth" more to one's career (in terms of annual merit evaluations, promotion, and tenure) than would publishing in other journals. That is, a publication was not a publication was not a publication. In fact, I was further cautioned that at the leading research institutions publication in a journal of questionable repute might even result in points being *subtracted* from a researcher's total research productivity "score."

Thus, high-minded and equipped with a headful of research dos and don'ts (in the context of other "Berkeley in the '60s" ongoing events), I ventured off to my first "real" job as an assistant professor in the University of Wisconsin–Madison's Department of Educational Psychology. From that

point on, it did not take long for me to discover that what happens in the "real world" of professional research and publishing does not always correspond to the practices that are preached to "wide-eyed and bushy-tailed" young aspiring researchers.

### Researcher Integrity (And Specifically, a Lack Thereof)

I still recoil from the shocks I experienced in first learning of unethical behaviors associated with the research and publication process. Most salient among these remembered violations are:

- An announcement that appeared in a University of Wisconsin–Madison newsletter, indicating that a PhD degree (of a former student in psychology, no less) was being revoked because it had been discovered that the student had plagiarized his dissertation.
- A similar announcement that a chemistry graduate student was being dismissed from the program because he had "dry labbed" (fabricated) his doctoral research.
- Reports that (a) knighted psychologist Cyril Burt had doctored his universally cited twin-IQ data (e.g., Dorfman, 1978), and (b) the much heralded production of cold fusion by two University of Utah scientists was more likely a bold illusion (*Wisconsin State Journal*, 1991). Such reports revealed that scientific missteps are not restricted to the deviousness of graduate students with graduation and employment motivations but that they are similarly manifested in the published work of prominent professional researchers.

With respect to data analysis accuracy, I experienced another horrifying reaction when I first heard about the results of an informal journal survey conducted by Arthur Jensen and William Rohwer (summarized by Levin, 1985). In examining the published data analyzed and reported in several creditable psychology journals, Jensen and Rohwer found that a sizable percentage of the articles (25–33%) contained serious data analysis errors—errors so serious in fact that reanalyses of the published data completely changed the nature of the authors' results and conclusions. As I also noted: "[T]he numerical errors discovered [by Jensen and Rohwer] include only data-analysis errors that could be easily checked. It is a sobering thought to speculate how many other undiscovered errors were associated with these articles" (Levin, 1985, p. 230, footnote 8).

So much for the sanctity of the words and data produced for professional journals by seasoned researchers. Yet, in the "ethical" spirit of this book, the extent to which such discovered data analysis misadventures

consisted of authors' carelessness or ignorance, rather than intentional efforts to misrepresent their data (typically, to present their data in a more favorable light), is unknown. Issues and examples associated with the credibility of reported data are considered in this volume (see, e.g., Rosnow & Rosenthal's Chapter 3) and in various published sources, including ones that I have routinely assigned to students in my graduate research methodology and professional issues course over the years (Barber, 1973; MacCoun, 1998; Rosenthal, 1966).[1]

Two personal close encounters with research integrity violations came my way while I was serving in an editorial capacity for APA's *Journal of Educational Psychology*. The first was a case of data fabrication, and the second was a case of plagiarizing another author's work. Each of these encounters is summarized in turn—as will be a number of other unethical research and publishing behaviors throughout this chapter.

### Illustrative Case 1: Data Too Good to Be True

A colleague to whom I had sent an empirical manuscript for review wrote back that the manuscript contained "interesting" results. The focus of the study was on outcome differences among several experimental conditions (in both mean patterns and levels), differences that would be interpreted as either supporting or not supporting theory-based predictions that the authors were testing. In each of some five experiments, not only did the data fit the theory-based predictions almost perfectly, but from one experiment to the next, the different mean values, condition by condition, were also virtually *identical*. "Too good to be true" were these inter-experimental means that they aroused the reviewer's suspicions about the research's integrity, much in the way that suspicions were aroused in readers observing Sir Cyril Burt's reported correlation coefficients. Those correlations, one after another, were "right on the money"—down to the second decimal place—in terms of what Burt's favored genetic theory predicted they should be. In his review of the manuscript under consideration here, the reviewer indicated that it would be a good idea (specifically, a wise precautionary action) for me to request the research participants' actual data protocols from the senior author before proceeding with the review process.

[1] Not presented in the present chapter are issues related to the ethical and proper treatment of human participants per se (e.g., recruitment, informed consent, confidentiality, and deception) because these are covered extensively in a variety of other sources, including APA's *Ethical Principles of Psychologists and Code of Conduct*—cited throughout the most recent (6th) edition of the *Publication Manual of the American Psychological Association* (APA, 2010) and retrievable from http://www.apa.org/ethics—as well as books by Fisher (2003) and Sales and Folkman (2000).

As an important aside, an editor's request for raw data during the manuscript review process (for any number of reasons) is nothing out of the ordinary. It is explicitly addressed in the APA *Publication Manual*:

> Researchers must make their data available to the editor at any time during the review and publication process if questions arise with respect to the accuracy of the report. Refusal to do so can lead to rejection of the submitted manuscript without further consideration. (APA, 2010, p. 12)

Equally important, and as is noted both for this example and later in the chapter, authors are expected to retain their studies' raw data "for a minimum of five years after publication of the research" (APA, 2010, p. 12).

Now back to the manuscript in question. When a raw data request was made of the manuscript's senior author, his initial response was that he had just relocated to a different university and that he would need some time to find the box containing the study's data. After a long time, the author wrote back, indicating that a graduate student coauthor must have the data in his possession and so he would be contacted. Here's an Andy Rooney-esque question to ponder: Why in such sticky situations is it inevitably a case of *Cherchez l'étudiant!* (which, in the present context, can be loosely translated as "Assign blame to the graduate student coauthor!")? To make this long story shorter, several additional months passed, and the missing data were never found. Were the "lost" data truly lost? I leave that for the reader to ponder, and we will return to this case in the section entitled "Unethical and Illegal Research and Writing Behaviors: Crimes and Punishments."

### Illustrative Case 2: Excuse Me, Haven't I Met You Somewhere Before?

A second up-close-and-personal experience that hit me squarely between the eyes also occurred during my journal-editing days. In that instance, I was conducting an editor's standard read-through of the literature review section of a manuscript that had been recommended for publication by journal reviewers. In that section, I came across a few paragraphs and turns of phrase that genuinely appealed to me. Initially the appeal was largely a result of my resonating to the substance and style of the author's writing. I found the flow of ideas particularly easy to follow, almost as though I had written the text myself. Reading on a little further, I realized that I was mistaken. I hadn't *almost* written the material myself: I *had* written the material myself—in fact, several paragraphs in a previously published article of mine. Just to be clear: The paragraphs in the author's manuscript were not reported in direct quotes or as paraphrases of my earlier work with accompanying citations. The paragraphs were verbatim

copies of mine, with nary a quotation mark nor cite of my earlier work in sight. Although some readers may feel somewhat sympathetic toward the author for having the misfortune of finding his plagiarized passages end up in the hands of the plagiarizee (the journal editor, to boot!), I surely hope that sympathy is not the prevalent reaction. Plagiarism is one of the most serious ethical violations associated with the publication process. Let me again refer to the APA *Publication Manual* to see how this type of offense is regarded in professional circles:

> Researchers do not claim the words and ideas of another as their own; they give credit where credit is due (APA ethics code Standard 8.11, Plagiarism [see APA, 2002]). Quotation marks should be used to indicate exact words of another. *Each time* you paraphrase another author … you need to credit the source in the text …. This can extend to ideas as well as written words. (APA, 2010, pp. 15–16)

Believe it or not, over the years I have come to learn that I am not alone in this "could not have said it better myself" experience. Other colleagues have reported uncovering similar plagiarism episodes. And what about the type of punitive action that should be taken against such plagiarists, including the one involved in my own "up-close-and-personal" encounter? That will be revealed in the later "Crimes and Punishments" section. More on the topic of plagiarism appears throughout this chapter, along with extensions to its close cousins, "self-plagiarism" and "duplicate publication."

## Questionable Behaviors Observed as APA's Chief Editorial Advisor

Under the watchful eye of APA Publisher Gary VandenBos, the association's Publications and Communications Board selects the editors, typically for a 6-year term, for its many prestigious professional journals. In addition, for several years now, APA's Publications Office has engaged the services of an academic researcher (and generally a past APA journal editor) to be its Chief Editorial Advisor (CEA). Among a variety of official duties, a major responsibility of the CEA is to serve in an ombudsperson capacity, helping to resolve editorial conflicts between editors and authors. In addition, consultation with editors on suspected ethical and legal violations by authors comprises a substantial portion of the CEA's caseload.

When I was Editor of the *Journal of Educational Psychology* from 1991–1996, Martha Storandt (Washington University in St. Louis) served capably as

APA's CEA. Through Martha's regular reports to APA's journal editors of "shady" research and publishing practices that she had adjudicated during that period, I came to realize that my own idiosyncratic editorial encounters with research and publishing misbehaviors were not so idiosyncratic at all—and much more prevalent than I had initially envisioned. Such realizations were further bolstered between 2003 and 2008, during my own stint as APA's CEA.

And what was the frequency of alleged researcher and author misconduct cases of which I became aware throughout my CEA tenure?[2]

- One source of data comes from the 25–30 cases that were brought to my attention each calendar year. My records indicate that typically between 30% and 40% of that caseload (or some 8–12 cases per year) involved unethical behavior on the part of a researcher or author. During my 6-year CEA term, I likely dealt with an astounding 50–70 "ethical violations" cases.
- A second data source comes from an informal survey that I conducted 5 years ago with editors of APA's major primary research journals (numbering about 30 at the time). The study focused on plagiarism per se and included various forms of plagiarism discovered by journal editors and their reviewers that were either handled by the journal editor or discovered and reported to me. Based on an approximate 67% response rate, the modal number of plagiarism cases during the targeted 1-year period was 1 per journal (ranging from 0 to 5), with the majority of those transgressions consisting of self-plagiarism (to be discussed later in this chapter).

Complementing my own CEA encounters with researcher misbehavior is a commentary published in the widely read scientific journal *Nature*. In that commentary, Titus, Wells, and Rhoades (2008) report the results of a survey that they conducted with 4,298 researchers from 605 academic institutions who had received funding from the National Institutes of Health (NIH). The focal questions of the survey were whether and how frequently the researchers had observed any instances of research misconduct (defined as data fabrication, data falsification, or plagiarism) among

[2] Two aspects of this question call for clarification. First, of the vast number of alleged author misdoings that were brought to my attention, all but a handful of them were substantiated. Second, the incidence figures that I am about to provide must be regarded as substantial underestimates. That is because the reported misconduct statistics are based almost exclusively on journal reviewers', editors', APA staffers', and—believe it or not—a few coauthors' discovered and reported ethical violations. The incidence of undiscovered and unreported (and, therefore, untabulated) unethical practices is anybody's guess.

colleagues in their departments during the most recent 3-year period. Based on a 51% response rate:

> One hundred ninety-two scientists (8.7%) indicated that they had observed or had direct evidence of researchers in their own department committing one or more incidents of suspected research misconduct over the past three academic years. The 192 scientists described a total of 265 incidents. (Titus et al., 2008, p. 980)

Of the 265 reported incidents, 201 met the federal definition of "research misconduct," with 60% of those incidents consisting of data fabrication or falsification and 36% consisting of plagiarism. The study's authors offer a sobering conclusion regarding the estimated frequency of research misconduct at academic institutions:

> In our survey, 201 cases were observed over three years by 2,212 respondents, essentially 3 cases per 100 people per year. Most conservatively, we assumed that non-responders (roughly half of our sample) did not witness any misconduct. Thus, applying 1.5 cases in 100 scientists to 155,000 researchers suggests that there could be, minimally, 2,325 possible research misconduct observations in a year. If 58% of these cases were reported to institutional officials as in our survey, approximately 1,350 would have been reported whereas almost 1,000 could be assumed to go unreported to any official. (Titus et al., 2008, p. 981)

And what is a possible take-home message of the Titus et al. (2008) survey? How about, "Believing the results of NIH-funded research can be dangerous to your health!"?

## Questionable Behaviors by Researchers and Authors

In this section, I provide a soup-to-nuts sampling of problematic issues to which I was exposed during my service as APA's CEA.[3] For many of these instances, after summarizing the context in which the issue arose, I will include what the 6th edition of APA's *Publication Manual* (2010) has to say in the way of proscriptions against such research and publishing "no-nos." I state at the outset that not all of these "don'ts" are equally egregious, willful, or even unethical—a topic that in fact is the focus of the final major section of this chapter. Although the majority of the offenses that I address here bear on the actions of researchers and authors, I will

[3] The details associated with these issues will be reported with as much accuracy as my memory allows.

also mention a number of questionable behaviors engaged in by editors and reviewers.

### *Simultaneous Submission of the Same Manuscript to Different Journals*

Submitting a manuscript to a professional journal does not assure publication in that journal. For example, a look at the *American Psychologist's* annual report of APA journal operations will reveal that for any given journal, the submission figures do not equal the acceptance figures. Far from it, as APA journals' acceptance rates, across the various content domains, are typically in the 20–30% range. In short, the vast majority of manuscripts submitted to an APA journal are not ultimately accepted for publication in that journal. The same is true, in varying degrees, for manuscripts submitted to academic journals associated with other professional organizations and publishers. So what is a researcher to do when his or her manuscript is rejected by a journal editor? Although a few different options are available, one option is for the researcher to submit the manuscript (either in its original or a revised form) to a different journal. In the research and publication process, that is a perfectly natural, and acceptable, thing to do. What is not acceptable, however, is when, in anticipation or fear of rejection—or simply to maximize one's publication probability—a researcher submits a manuscript to two (or more) different journals *simultaneously*. Curiously buried only in the *Publication Manual's* "Checklist for Manuscript Submission" questions (Section 8.07) can be found the following prohibitive signal to authors: "Is a cover letter included with the manuscript? Does the letter … state that the manuscript is original, not previously published, and not under concurrent consideration elsewhere?" (APA, 2010, p. 243).

The previous (2001) edition of the APA *Publication Manual* made the "simultaneous submission" prohibition policy much more explicit and conspicuous:

> The same manuscript must not be submitted to more than one [journal] at the same time. Submission of a manuscript implies commitment to publish in the journal if the article is accepted for publication. Authors submitting manuscripts to a journal should not simultaneously submit them to another journal. (p. 352)

And what is the rationale for this policy? Primarily, that not adhering to it can create burdensome situations (in terms of time, energy, resources) for editors and reviewers. In addition, it can create awkward predicaments for authors in cases where a manuscript ultimately is accepted for publication by more than one journal.

### *Piecemeal (or Fragmented) Publication*

Another research and publishing practice that is frowned on is one of publishing (or submitting) two or more manuscripts that contain different "pieces" of information from the same general project or data source. As is stated in the APA *Publication Manual*:

> Authors are obligated to present work parsimoniously and as completely as possible within the space constraints of journal publications. Data that can be meaningfully combined within a single publication should be presented together to enhance effective communication. Piecemeal, or fragmented, publication of research findings can be misleading if multiple reports appear to represent independent instances of data collection or analyses; distortion of the scientific literature, especially in reviews or meta-analyses, may result. Piecemeal publication of several reports of the results from a single study is therefore undesirable unless there is a clear benefit to scientific communication.… Whether the publication of two or more reports based on the same or on closely related research constitutes fragmented publication is a matter of editorial judgment. (APA, 2010, p. 14)

Piecemeal publication considerations on the part of authors and editors often boil down to judgments about the magnitude of the particular manuscript's contribution: Is the information provided in the manuscript sufficiently important or "newsworthy" that it merits publication? Or would the journal's readership be better informed if that information were combined with "something more" (in the form of related research hypotheses, different participant populations, other outcome measures, an additional experiment or two, and the like)? These judgments are especially relevant when it is clear that the author in question has first-hand access to such additional information (specifically, comparable findings based on his or her previously or not-yet-published work). Whether a given manuscript constitutes a "piecemeal" attempt (intentional or not) is, unfortunately, not a cut-and-dried determination. For example, data associated with multiple-site and multiple-year longitudinal studies, as well as multivariable studies based on the same participants, but that represent fundamentally different aspects of the research problem under consideration could represent reasonable exceptions to the general rule.

Authors should be keenly aware of piecemeal publication policies and, in cases where there is any doubt, they are advised to consult directly with journal editors concerning the suitability of manuscripts reporting the results of narrowly focused individual experiments. The alternative to publishing reports of individual experiments is to publish "multiple-experiment packages" (Levin, 1991a), an alternative that typically provides more meat on the bones of the work being reported, thereby "telling

a more compelling and complete story" of the research issues in question. At the same time, an oft-heard, and readily appreciated, "downside" of crafting multiple-experiment packages is the conflict it creates (with respect to the investment of time and effort) for tenure-seeking researchers at the beginning of their academic careers where, more often than not, publication quantity weighs heavily alongside publication quality. An editor of this volume also correctly notes that another potential negative byproduct of (necessarily longer) manuscripts reporting multiple experiments is that readers may be less inclined to "wade" through them.

### *Duplicate Publication of the Same Work*

An author submitting a previously (or about-to-be) published manuscript to additional publication outlets—referred to as "duplicate publication"—is unacceptable from a publishing ethics standpoint. The preceding statement applies equally well to manuscripts that overlap *substantially* with previously published works (including both journal articles and book chapters) and therefore might be characterized as "overlapping" publication. According to the APA *Publication Manual*:

> Misrepresentation of data as original when they have been published previously is specifically prohibited by APA ethics code Standard 8.13, Duplicate Publication of Data [APA, 2002]. Authors must not submit to an APA journal a manuscript describing work that has been published previously in whole or in substantial part elsewhere, whether in English or in another language. More important, authors should not submit manuscripts that have been published elsewhere in substantially similar form or with substantially similar content. (APA, 2010, p. 13)[4]

And why not? In addition to the editorial workload issues already noted:

> Duplicate publication can give the erroneous impression that findings are more replicable than is the case or that particular conclusions are more strongly supported than is warranted by the cumulative evidence. Duplicate publication can also lead to copyright violations; authors cannot assign the copyright for the same material to more than one publisher. (APA, 2010, p. 13)

Two other circumstances that might be considered "duplicate publication" by some professional organizations should also be noted. One is submitting or publishing material that has previously been published

[4] Here's a *Guinness Book of World Records* aside: In my professional experience, I have seen the same article written by the same author, in the same form with almost exactly the same content—and without acknowledgment or citation of the previously published work—in *five* different publication outlets!

(typically in a book) as proceedings from a professional conference. The second is the publication of universally available Internet-based materials. Different professional organizations, editors, and publishers have different views on the legitimacy of publishing such previously disseminated information, and so authors are advised to consult with individual journal editors before proceeding.

### *Self-Plagiarism*

As was just discussed, authors publishing the same work more than once (duplicate publication) is a professionally unacceptable behavior. The same can be said of authors incorporating selections from their previously published work into their subsequent publications, if such incorporation is done in an inappropriate manner. By "inappropriate," I mean authors including portions of their previous work (whether large or small) without explicit reference to it through the use of quotation marks in the case of exact duplications of text, along with specification of the authors, dates, page numbers, and the complete citation of the previous work in the reference list. Verbatim copying (which has become an especially painless process in the modern "cut and paste" era of writing) is an obvious manifestation of self-plagiarism, but so too is paraphrasing previous text segments and ideas—or what I have previously referred to as "*paraphragiarism*" (Levin, 1991a; Levin & Marshall, 1993). That is, the repetition of large sections of parallel idea development and structure in multiple documents generated by the same author(s) is generally not acceptable (but see below). Again, I refer to the relevant section of the APA *Publication Manual* for APA's self-plagiarism stance:

> The general view is that the core of the new document must constitute an original contribution to knowledge, and only the amount of previously published material necessary to understand that contribution should be included, primarily in the discussion of theory and methodology. When feasible, all of the author's own words that are cited should be located in a single paragraph or a few paragraphs, with a citation at the end of each. Opening such paragraphs with a phrase like "as I have previously discussed" will also alert readers to the status of the upcoming material. (APA, 2010, p. 16)

Yet, for past several years, APA's Council of [Journal] Editors and its Publications and Communications Board have discussed, been amenable to, and now endorse the notion that in selective situations, authors' verbatim reuse of their same previously written words—absent the just-stated self-plagiarism safeguards—*is* acceptable. Specifically, what might be termed the "boilerplate-language" proviso appears in the most recent edition of the APA *Publication Manual* in the following form:

> There are, however, limited circumstances (e.g., describing the details of an instrument or an analytic approach) under which authors may wish to duplicate without attribution (citation) their previously used words, feeling that extensive self-referencing is undesirable or awkward. When the duplicated words are limited in scope [and typically restricted to repeated descriptions of theoretical rationales, materials, procedures, measures, and analysis strategies], this approach is permissible. (APA, 2010, p. 16)

The logic behind this proviso is that for certain technical expositions, authors may have previously invested considerable cognitive resources in crafting "just right" descriptions. So, the argument goes, why force these authors through the same laborious process again to come up with similar, although not identical or paraphrased, "just right" descriptions? The argument sounds "just right" to me, and so I am pleased to see the boilerplate-language proviso included in the most recent edition of the APA *Publication Manual* (2010).

### *Plagiarism*

The serious research and publication offense known as plagiarism—appropriating parts (and in some cases, even wholes) of another author's work without attribution—was initially presented here in the context of Illustrative Case 2. While serving as APA's CEA, I cautioned researchers about the gravity of this ethical violation in an open letter that APA editors make available to all authors who submit manuscripts to their journals:

> Imitation may be the "sincerest form of flattery," but in professional writing imitation without appropriate attribution (e.g., Colton, 1820–22, cited in Bartlett, 1992, p. 393) is not acceptable. Authors should cite the sources of their ideas and methods, as well as put quotation marks around phrases taken from another source. The change or reordering of a few words in a sentence does not relieve authors of the obligation to quote and recognize appropriately the source of their material. As recent cases inform us, authors need to be scrupulous in their notetaking (especially in electronic form) and careful about using those notes in their own manuscripts. (Levin, 2003)

It should be noted that professional plagiarism applies not just to unattributed copying or paraphrasing of other researchers' printed words and thoughts, but also to "stealing" their materials, instruments, or any other aspects of their research.[5] It is also worth noting that one can readily

[5] Using others' copyrighted materials without explicit permission (and, in many cases, fees) opens a Pandora's box of potential legal violations, which takes us beyond the scope of the present chapter.

identify occurrences of plagiarism in all academic fields, in relation to authors of popular books, and in almost daily reports by the mass media. In recent years, for example, well-known historians Doris Kearns Goodwin and Stephen Ambrose (and even Martin Luther King, Jr., years earlier) have not escaped the plagiarism-charge net. One "ethical violations" manila file that I have amassed over the years measures approximately 4 inches in depth and consists primarily of specific plagiarism offenses. Thus, we are not talking about a small—and clearly not a rare—problematic behavior here. I will now summarize a few illustrative plagiarism cases that I experienced during my tenure as APA's CEA.

*Illustrative Case 3: Do as I Say, Not as I Do*

Undoubtedly the most ironic instance of plagiarism that I ever encountered came my way while I was reading two contiguous chapters in the earlier-mentioned APA-published book, *Ethics in Research with Human Participants* (Sales & Folkman, 2000). In the chapters, entitled "Authorship and Intellectual Property" and "Training," the general topic of assigning *authorship credit* (discussed here in a later section) was being considered. In the former chapter the following text appeared:

> In practice, however, students come to the dissertation with varying degrees of experience and expertise. Indeed, in some instances, the dissertation represents the student's first research project.… [I]t is generally expected that the student would have the opportunity to be the principal author on any publication resulting from the dissertation.… Prior to the initiation of the dissertation research, the faculty advisor and student should explicitly discuss and reach an agreement on publication requirements.… (McGue, 2000, p. 81)

Then, in the latter chapter appeared the following text:

> In practice, students come to the dissertation with varying degrees of prior experience and expertise. In fact, in some instances, the dissertation represents the student's first research project … Throughout the dissertation process, researcher-supervisors are encouraged to discuss and evaluate together with the student the appropriateness, if any, of supervisor co-authorship, based on their relative contributions to the project. There is a presumption that the student will be listed as principal author on any multiple-authored article that is substantially based on the student's dissertation or thesis.… (Tangney, 2000, p. 103)

The first two sentences of the respective chapters are practically identical. The remaining sentences are close paraphrases. What is going on here? Both authors properly cite a journal article by Goodyear, Crego, and Johnston (1992) as the source of their respective material, but "Who's

(on) first?" Why doesn't one chapter author properly cite the other author with respect to the uncannily similar language and ideas? Are those similarities merely a "coincidence?" Or is one set of excerpts—taken from different-authored chapters in a book on research and publishing *ethics*—simply a plagiarized version of the other? We will return to this illustrative case later.

*Illustrative Case 4: Pardon My French!*

The just-presented plagiarism example focused on an author improperly "borrowing" several words and thoughts from another author. Further along the plagiarism continuum is one author stealing virtually *all* of another author's work. In my capacity as APA's CEA, I was informed of a plagiarism situation in which the author of a chapter for a book (copublished by APA) had obviously plagiarized a previously written entire biography of a noted French psychologist. Moreover, an ongoing investigation revealed that the plagiarized biography in question was not the only one of concern. Two additional biographies in the same volume, "written" by the same author, were also suspect. An interesting twist to this case was that the earlier biographies had been written in French, whereas the ones for the current book were in English. One has to wonder whether the plagiarizer imagined that copying the original words from one language to another would help to camouflage his offense, thereby decreasing the likelihood of the crime being discovered. Once again, more on this case later.

And now for the grand finale to this section: Years ago, an outgoing editor of the journal *Child Development* generated a humorous set of "action letters that I wish I had written" and distributed them to his editorial board and other colleagues. Most of these concocted letters to authors were rejection letters, pointing out various "defects" in the manuscript, including the research topic itself, the writing style, organization, data analyses, etc.—and likely even defects traced to the author's parentage! One of the letters, however, was singularly praiseworthy with respect to the author's manuscript, and went something like this:

> Dear Author:
>
> Your manuscript reports one of the most groundbreaking pieces of research that I have ever seen. Moreover, it is written impeccably. My reviewers were overwhelmed by the manuscript's excellence and offered no suggestions for its improvement. I have thoroughly examined the manuscript myself and agree wholeheartedly with the reviewers' assessments. Because of your study's exceptional quality and unquestionable impact on the field, I have therefore decided to act on your manuscript in a totally unprecedented way as Editor of *Child Development*. I am accepting the manuscript in its present form,

> unconditionally, and without the need for any revision whatsoever. I just used the phrase "without the need for any revision whatsoever," but that is not completely accurate. There *is* one very minor editorial change that you will be required to make in order for the manuscript to be published in *Child Development*. The required change is that you remove your name from the "Author" byline and that you replace it with my name!
>
> **Sincerely,**
> **The Editor**

This fictitious letter was, of course, obviously tongue in cheek and it undoubtedly comprised a source of great hilarity to its many recipients. The following illustrative case, patterned after the name-change action in the *Child Development* editor's letter, is neither fictitious nor tongue in cheek, but rather a true-to-life occurrence that comprised a source of extreme discomfort and grief to everyone associated with it.

*Illustrative Case 5: Research and Writing Made Easy*

In perusing the contents of an electronic journal, a researcher came across an article reporting the results of an experiment that was of special interest to him. And why exactly was that? Could it be because the research reported came directly, entirely, and in near-verbatim form from an article that the individual himself had coauthored and had previously published in an APA journal? In fact, to quote the understating complainant: "As far as we can tell, the only differences between our article and the one published in [the electronic journal] are (1) the title, (2) the formatting, and (3) occasional pronoun replacements. Clearly something is not right about this." Clearly. Bear in mind that this was not simply a piece of expository writing that had been "lifted" by the offender, but rather a primary research report, complete with participants, procedures, data collection, analyses, and all. I became aware of this incident because one of the original authors immediately reported the situation to the editor of the APA journal in which the article was original published, who in turn contacted me for advice about how to proceed. Those proceedings proceeded with immediate haste. And how did all of this end, including the fate of the plagiarism perpetrator? Stay tuned for the "rest of the story" (à la the late Paul Harvey), to be reported later.

*Avoiding Plagiarism and Fingering Plagiarists*

In my experience, I have found that—in contrast to the blatant author-charade example just discussed—many instances of publication plagiarism are unwitting, resulting from the lack of complete information or education on the part of the plagiarizer concerning the nature and specific characteristics of the offense. Student research assistants often

fall into this category, as do researchers and authors from various foreign countries, where the norm is to use the exact words of respected authorities rather than dare to change them. What such researchers need to be told is that using exact words (sparingly) is certainly permissible, as long as they are enclosed in quotation marks with the exact source cited. With respect to paraphrasing, let me repeat some previously offered suggestions about how researchers can safeguard against plagiarism:

> [W]hen delegating the responsibility of a literature review to a novice student coauthor, one may be flirting with the possibility of [plagiarism]. The recent advice of *Parade* magazine's savant, Marilyn vos Savant (1992) is well taken here: She recommended not attempting to "paraphrase" someone else's thoughts while looking directly at the source; close the book and then paraphrase [with appropriate referencing] (p. 27). (Levin & Marshall, 1993, p. 6)

Similarly stated, it is a dangerous practice to paraphrase a passage while reading it. My own recommended four-step plagiarism-prevention practice is (a) read the passage; (b) digest what you have read; (c) close the source book or article; and then (d) summarize the passage in your own words. Also be aware that when plagiarism is traced to a single author of a multiple-authored paper, unfortunately in most cases all authors must accept some degree of responsibility for the offense.

In this chapter and also in my experience, the plagiarism cases that surfaced and that later were documented resulted from *human surveillance techniques*. That is, plagiarizers have been caught red handed through the vigilance of knowledgeable readers or, as happened in Illustrative Cases 2 and 5, by the hapless victims of the crimes that were perpetrated on them. Fortunately, over the past several years, more systematic (dare I say foolproof?) means of "catching a thief" have been developed and validated. One such approach is Glatt's "Plagiarism Screening Program" (2010), based on Barbara Glatt's dissertation at the University of Chicago and follow-up research by Glatt and Haertel (1982). The approach, which requires the alleged plagiarizer to submit to a performance test (to "prove" his or her innocence), is adapted from what in the reading-comprehension literature is known as the "cloze technique" (e.g., Bormuth, 1969). Specifically, the test-taker is provided with a section or more of the presumed plagiarized material in which every fifth word of the original text has been removed and replaced with a blank. The test-taker's task is to fill in all the blanks. Under the very reasonable (and empirically supported) assumption that one can reconstruct one's own previously constructed sentences more accurately than can an individual who did not originally construct the sentences, a plagiarizer typically fails the test

by not exceeding a preestablished accuracy cutoff score. In fact, in the many cases of which I am aware in which the test has been administered in other settings (e.g., with university students accused of stealing term papers from the Internet or copying those of other students), the alleged plagiarizer makes a quick "self-discovery" of his or her futile test attempt and admits to the offense.

A second approach to uncovering plagiarism is to take advantage of the enormous *databases* that nowadays can be readily accessed through the speed and power of *computer technology*. The databases in question would include as much professional literature as is relevant to the presumed plagiarized text. Exhaustive comparisons between the databases and the subject text are then made.[6] What follows is an intriguing detailed description of how one of the earliest versions of this type of plagiarism-detection software works:

> [Walter Stewart and Ned Feder's] method is to feed pages from the suspect article into a scanner that can read a variety of typefaces and convert them into electronic form. The electronic version of the text is then broken down by the computer program into strings of 30 characters each, including letters, numbers and spaces. The first string begins with the first word of the first paragraph, the second begins with the second word, and so forth, building overlapping strings throughout the article.… To compare all the strings in one text with all the strings in the rest of a field's scientific literature would take an inordinate amount of computer time. Instead, the program sorts all the strings in a computer equivalent of alphabetical order, thus putting identical pairs next to each other, which the computer then prints in boldface.…
>
> After doing thousands of such runs, Mr. Stewart said: "The most surprising thing is how unique human language is. We find very, very few duplicates—even in highly technical text talking about the same thing." As it turns out, in the 7,000 or so manuscripts that he has looked at so far, Mr. Stewart has found that only about one string in 200 may be duplicated by chance alone. That rate is about five "millifreemans" in the units of plagiarism created by Dr. Feder and Mr. Stewart.…
>
> The basic unit, one freeman, "refers to the ultimate case, the theft of an entire document word for word, changing only the author's name," said Mr. Stewart. The unit is named after an individual whom he regards as having committed large-scale plagiarism. Attempts to reach the individual by telephone were unavailing.…

[6] The currently popular Turnitin website (http://www.turnitin.com), with its purchasable software, is a manifestation of this approach, as applied to the term papers of university and high school students who are suspected of plagiarizing from other sources.

> Mr. Stewart says that at the level of 10 millifreemans and above, "There is serious reason to look at two documents to see if there is plagiary or the identical passages have been properly attributed." (Hilts, 1992)

The message of this extensive section to would-be text thieves? Plagiarists beware: If the eyeball does not get you, then the cloze test or text-matching computer software will!

### *Data Falsification*

Earlier in the context of Illustrative Case 1, along with tales of suspected twin-IQ and cold fusion studies, I introduced the research transgression of data falsification (also referred to as *data fabrication* and *data fudging*).[7] According to the APA *Publication Manual*:

> [P]sychologists do not fabricate or falsify data (APA ethics code Standard 8.10a) [APA, 2002]. Modifying results, including visual images … to support a hypothesis or omitting troublesome observations from reports to present a more convincing story is also prohibited…. (APA, 2010, p. 12)

Despite the comparatively brief attention paid to data falsification in the *Manual*, make no mistake that the offense is an extremely serious one, typically resulting in a severe form of punishment.

While serving as APA's CEA, I encountered a few documented instances of data falsification by a researcher–author, each of which had profound repercussions for the offender. In one case, a multiauthored research study, the tainted results had already been published in an APA journal. To deal with that unfortunate situation and to "correct" the literature, a formal *retraction* of the article was published in the original journal while punitive action against the falsifying coauthor was undertaken. In another case, the source of intentionally altered data was discovered, which was followed by a reanalysis of the correct data that appeared as a corrected version of the original article (a *corrigendum*).

### *Assignment of Authorship Credit*

When it comes to professional publications, the answer to the question, "Who deserves authorship credit when, and in what order?" is not an easy (or universally accepted) one. The APA has a set of authorship standards

[7] One wonders whether the latter term comes from a likely-to-be-heard expression uttered by the falsifier on being fingered: "Oh fudge!"

and expectations, of which I have excerpted parts here from Section 1.13 "Publication Credit" of the APA *Publication Manual*:

> Individuals should only take authorship credit for work they have actually performed or to which they have substantially contributed.… As early as practicable in a research project, the collaborators should decide on which tasks are necessary for the project's completion, how the work will be divided, which tasks or combination of tasks merits authorship credit, and on what level credit should be given (first author, second author, etc.).… Principal authorship and the order of authorship credit should accurately reflect the relative contributions of persons involved. (APA, 2010, pp. 18–19)

A coherent set of principles and recommendations for authorship credit also appears in a superb article by Fine and Kurdek (1993). In my own professional experience, I have observed—and have counseled aggrieved parties about—authorship conflicts.[8] Frequently such conflicts may be traced to "power" or "status" issues involving senior (typically faculty supervisors or advisors) and junior (typically students) players in a research project—see also Goodyear et al. (1992). In such situations, it is not difficult to imagine who (at least initially) generally comes out on the "short end of the stick." Fine and Kurdek provide several instructive examples related to authorship (including authorship order) as a function of an individual's involvement in various aspects of the research and writing process.

### *Data Sharing*

A recent addition to the most recent edition of the APA *Publication Manual* (2010) focuses on authors "sharing" their data with well-meaning editors and professional colleagues with well-meaning intentions. Well-meaning intentions would include editors or reviewers requesting data during the submission process for purposes of corroborating an author's analyses and interpretations of the data.[9] Such intentions would also include complying with legitimate data requests from colleagues (a) who may be involved in a similar field of inquiry and wish to examine the same research question using different data sources; or (b) who are conducting a meta-analysis and need to include information not reported in the published article but that could be extracted from the data. In fact, the *Manual* now clearly states:

[8] Perhaps surprisingly, some conflicts have consisted of one author's disputes about, or accusations of, some form of unethical behavior on the part of a coauthor.

[9] A more extreme version of data sharing is Wicherts, Borsboom, Kats, and Molenaar's (2006) proposal that manuscripts published in professional journals *must* be accompanied by an ASCII (text) file containing the raw data.

> [O]nce an article is published, researchers must make their data available to permit other qualified professionals to confirm the analyses and results.… To avoid misunderstanding, it is important for the researcher requesting data and the researcher providing data to come to a written agreement about the conditions under which the data are to be shared. Such an agreement must specify the limits on how the shared data may be used (e.g., for verification of already published results, for inclusion in meta-analytic studies, for secondary analysis). (APA, 2010, p. 12)

Authors refusing to comply with legitimate data-sharing requests are committing an ethical research and publishing violation, at least from APA's viewpoint: "Authors are expected to comply promptly and in a spirit of cooperation with requests for data sharing from other researchers" (APA, p. 12).

### *Other Questionable Researcher and Author Behaviors*

I conclude this section by listing, in summary fashion and with minimal additional explanation, several questionable (and in varying degrees, unethical) research and publishing behaviors on the part of authors that I experienced while serving as APA's CEA:

- Authors' misrepresentation of others' work, in some cases intentionally and in others not.
- Personal (*ad hominem*) attacks on an author (as opposed to professionally acceptable criticism of an author's work), followed by accusations of defamation.
- Authors listing articles as "in press" that have not received final acceptance; until a final acceptance letter has been received from the editor, authors should continue to refer to the work as either "Manuscript submitted for publication" or "Unpublished manuscript"—see the APA *Publication Manual* (2010, p. 211).
- Author conflicts of interest, particularly in relation to commercially available tests or software; such conflicts frequently involve some sort of monetary connection to a product or a service being reported in the manuscript (APA, 2010, p. 17); in the *Manual*, on p. 231 and in Figure 8.3, it is indicated that a no-conflict-of-interest statement must be signed on by each author.[10]

[10] Reports in *The New York Times* indicate that the conflict-of-interest sin frequently operates in conjunction with the sin of plagiarism (as manifested in the form of ghostwriting), specifically in regard to medical researchers attaching their names to documents written by anonymous authors representing pharmaceutical companies (Harris, 2009; Wilson, 2009; Wilson & Singer, 2009). A recent conflict-of-interest situation (which also includes suspicious data and ethical mistreatment of participants) has been alleged of a medical

- Comments on "in press" articles before they appear in print; this can happen when a researcher has access to an unpublished version of a manuscript and "jumps the gun" in submitting a response to it.
- Authors who cannot take "no" to an editor's decision to reject their manuscript outright; that is, despite the finality of the editor's decision, the authors resubmit the manuscript anyway, generally following some sort of revision attempt and often with an argumentative appeal to the editor for a reconsideration of his or her decision.[11]
- Authors surreptitiously submitting a previously rejected manuscript to an incoming editor of the same journal (i.e., without informing the new editor that the manuscript had been rejected by the former editor).

### Questionable Behaviors by Journal Reviewers and Editors

The research and publishing ethical considerations that I have presented thus far have been directed at researchers and authors. However, and perhaps contrary to popular belief, journal reviewers—and even editors—are also human beings, and sometimes fallible human beings at that! Present page limitations do not permit a complete account of the questionable behaviors of reviewers and editors that I have witnessed (and, in some cases, about which I have intervened) as a publication ethics watchdog, and so I will provide an abbreviated sampling of them here:

- Reviewer conflicts of interest, as was previously noted with respect to authors' conflicts of interest.
- Reviewers not respecting the principle that a manuscript they receive to review is a "privileged" document (APA *Publication Manual*, 2010, p. 18), in that they must not (a) share the manuscript with others; (b) redirect the manuscript to a colleague for review without first notifying (and receiving permission from) the editor; (c) enlist the assistance of a student reviewer without similar permission from the editor; or (d) "steal" ideas that are presented in the manuscript.

researcher who, in a now-retracted controversial 1998 article in *The Lancet*, reported a link between a combined measles–mumps–rubella vaccination and the incidence of childhood autism. In particular, "Part of the costs of [the investigator's] research were paid by lawyers for parents seeking to sue vaccine makers for damages [and the investigator] was also found to have patented in 1997 a measles vaccine that would succeed if the combined vaccine were withdrawn or discredited" (Harris, 2010).

[11] Appeals of this genre are typically accompanied by authors' claims of "ignorance" or "misunderstanding of the topic" on the part of the reviewers.

- Reviewers providing a review of a manuscript that they had previously reviewed (often negatively for another journal) without informing the editor, in that such action would serve to place the author in *double jeopardy*.[12]
- Reviewers challenging an editorial decision that goes against their own personal recommendations.
- Editor conflicts of interest in handling (a) manuscripts of associate editors that were submitted to their own journal; or (b) manuscripts of collaborators or institutional colleagues.
- Editors rejecting manuscripts outright (i.e., prior to initiating the formal review process) with inadequate justification.
- Editors being responsible for unwarranted delays in manuscript reviews.
- Editors applying "bait-and-switch" tactics in their editorial decisions; by this I mean informing an author that if X, Y, and Z are attended to in the way of revision, then the manuscript will be accepted for publication, but yet when X, Y, and Z *are* attended to by the author, the editor either requests additional revision or rejects the manuscript.
- Editors posting "uncivil" (Sternberg, 2002) reviews on the journal's website.

So, with a warning about incivility, this concludes my (dirty) laundry list of unethical and other questionable behaviors that I have personally observed in the professional research and publication enterprise. In the final section of this chapter, I offer some thoughts on (a) how such misbehaviors might be represented, and (b) selecting an appropriate punishment to fit the "crime."

## Unethical and Illegal Research and Writing Behaviors: Crimes and Punishments

I have found it both instructive and pragmatically useful to position specific ethical violations cases within a two-dimensional framework

[12] Scott E. Maxwell, a contributing author to this volume (Chapter 7) and the current Editor of *Psychological Methods*, applied the "double jeopardy" term to this publication situation (S. E. Maxwell, personal communication, April 25, 2008). An opposing view is that a rereview allows a previously negative reviewer to put a more positive "spin" on the manuscript by taking a look at a potentially revised and improved version of the original manuscript.

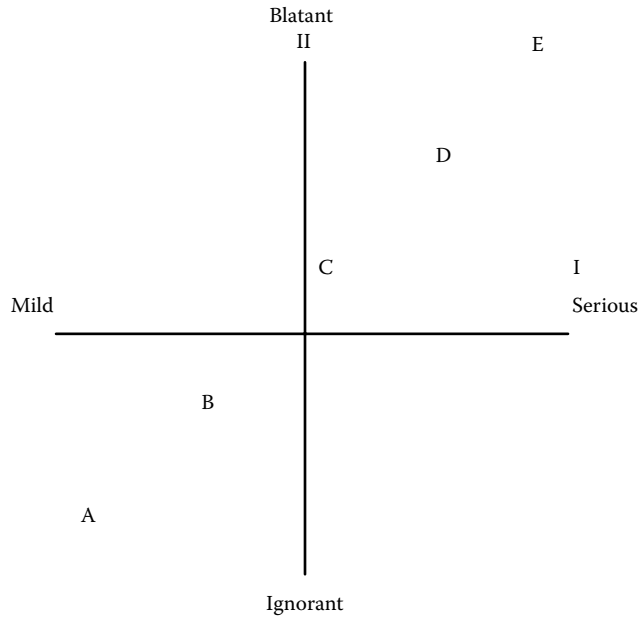



**FIGURE 18.1**
Two-dimensional representation of research and publishing crimes and punishments.

(adapted from Levin, 1991b, and displayed in Figure 18.1). The two dimensions are represented by (a) the *severity* of the offense (the horizontal axis of Figure 18.1), anchored by *mild* on the low end of the continuum and *serious* on the high; and (b) the degree of the offender's *intentionality* (the vertical axis of Figure 18.1), anchored by *ignorant* on the low end of the continuum and *blatant* on the high.

The severity dimension is fairly straightforward to characterize. Offenses such as carving up a multiple-experiment study into several pieces—what someone once called *LPUs*, or "least publishable units" (original source forgotten)—and submitting them to different journals (piecemeal publication) would be on the "mild" side in Figure 18.1, whereas data falsification and the "simply change the author's name" scenario summarized in Illustrative Case 5 would be far along the "serious" side.[13] In contrast, the intentionality dimension often requires assessing (or inferring) the offender's motivations, and, in some cases, ascertaining the level of the offender's research and publishing experience and knowledge. For example, if

[13] It should be noted that many cases in the latter ("serious") domain go beyond the CEA's jurisdiction and are referred to APA's legal office.

importing other authors' words or ideas by fledgling researchers (and, in my experience, by authors for whom English is not their primary language) can be attributed simply to naiveté (i.e., "not knowing the rules"), then the punishment should not be as severe as it would be in cases of blatant and willful plagiarism. As has previously been suggested:

> We believe that the particular punishment for [researchers' unethical behavior] boils down to the question of intentionality on the part of the offender, with "ignorant" or honest mistakes (stemming from an improper education on these matters) being treated more leniently than purposeful [misdeeds]. (Levin & Marshall, 1993, p. 6)

Thus, as can be seen in Figure 18.1A–E, the severity of the offense and the offender's intentionality should be considered jointly when determining an equitable punishment.[14] According to the present framework, for example, the punishment for a new PhD recipient carving up his or her dissertation into multiple LPUs (often following the advice of a dissertation advisor) might be represented by A ("Point out") in Figure 18.1, whereas an experienced researcher engaging in piecemeal publication (and especially, without informing the editor of the related manuscripts) might occupy a position in Figure 18.1's upper left quadrant and would deserve a somewhat harsher punishment (e.g., a "hand slap"). Similarly, and as will be seen shortly, the blatantly plagiarizing individuals described in Illustrative Cases 4 and 5 might have landed them in location E ("Take legal action") of Figure 18.1, whereas a plagiarizing novice author might be positioned in the lower right quadrant of Figure 18.1 and would deserve something less in the way of punishment. But enough of these idle hypotheticals! Let us proceed to some real live cases.

### Reconsideration of Illustrative Cases From a "Crimes and Punishments" Perspective

In this section, I describe the specific punitive actions that were taken for each of the five illustrative cases that have been discussed throughout this chapter. All actions should be interpreted with the two dimensions of Figure 18.1's framework in mind. In reviewing the various cases, I ask the reader to understand that ostensibly the same ethical crimes differ in their particulars, and so they cannot be cleanly fit into the same punishment "holes." In addition—and unfortunately—practical constraints and circumstances sometimes prevent the "ideally" prescribed punishments of Figure 18.1's framework from being administered.

[14] Much appreciated are conversations with attorney Allan Koritzinsky, who elucidated the similarities of this framework and our legal system's determination of equitable punishments.

### *The Rest of the Story for Illustrative Case 1: Data Too Good to Be True*

This was the case of a submitting author whom a reviewer "caught" with data that fit "theoretical" predictions a little too perfectly. After repeated requests for the "lost" raw data from the editor, the author (an experienced and widely published researcher) voluntarily withdrew his manuscript from consideration by the journal. Thus, an editorial decision that likely would have been to reject the manuscript outright—or even worse from a Figure 18.1 perspective if the data in question had been proven or admitted (rather than suspected) to be fraudulent—was averted by the author's ultimate action. As was noted by one of this volume's editors, the author surely ended up with an "easy out" in this particular case.

### *The Rest of the Story for Illustrative Case 2: Excuse Me, Haven't I Met You Somewhere Before?*

In this situation, the manuscript authors had the extraordinarily bad luck of their work being reviewed by someone who had previously authored several paragraphs that were included in their manuscript. That someone (JRL) was coincidentally the editor of the journal to which the manuscript had been submitted. I apprised the author of the problem, which the senior author attributed to a student coauthor who had been assigned the task of doing the literature review. He further apologized for his negligence and carelessness in not catching the plagiarism in the student's writing.[15] Adopting the department store philosophy that "the customer is [sometimes] right" (original source disputed), I accepted the senior author's claim of naiveté on the part of the student coauthor and was willing to receive a revised version of the manuscript with the plagiarized passages removed. An eerily similar plagiarism situation also occurred (i.e., someone submitting a plagiarized article of mine to a journal for which I was the editor) with a novice foreign author. Again after a plea of ignorance on the author's part, along with a "justification" stating in effect that "in my country we are taught to respect authority to the extent that we do not dare to alter an authority's words" (but without any source attribution at all?), I decided on the educative "Point out" punishment of Figure 18.1 and allowed the author to resubmit a manuscript that was written in accordance with "our" country's standards and expectations.

### *The Rest of the Story for Illustrative Case 3: Do as I Say, Not as I Do*

Here we have an APA book on research and publishing *ethics* that contained two chapters with virtually identical paragraphs. In that the book

[15] This particular scenario (including the senior author's response) is curiously similar to those of historians Doris Kearns Goodwin and Stephen Ambrose, alluded to earlier.

had already been published and I was not directly involved in this case but rather discovered it fortuitously after the fact, I did not take any action or recommend any action to be taken. I did, however, point out to the APA Publications Office the unfortunate irony of the situation. At the same time, although editors of a book with contributed chapters (or of special issues of a journal) would not be expected to recognize stolen passages from previously published works—other than passages stolen from the editors themselves!—at least they ought to be able to recognize familiar-sounding passages written by different contributors to the very volume that they are editing.

### *The Rest of the Story for Illustrative Case 4: Pardon My French!*

A French author "wrote" three biographies for a book for which APA was a copublisher. Somewhere along the way it was found that the author had basically translated the original biographies (written by a previous author in French) to English versions of them for the APA book. In that the book had not yet appeared in print, immediate action was taken by APA to "go after" the plagiarizer and, of course, to pull the three stolen biographies from the volume. I have no additional information about the situation, including whether the offender denied the allegations and attempted to argue for the originality of his contributions. If the author's arguing for originality in fact occurred, it would be a relatively straightforward operation to apply one of the plagiarism-fingering devices discussed earlier in this chapter—specifically, plagiarism-detection computer programs and a French-to-English adaptation of a cloze-based test procedure.

### *The Rest of the Story for Illustrative Case 5: Research and Writing Made Easy*

Recall the blatant and intentional plagiarism case, where a previously published article was republished in an electronic journal with the two original authors' names replaced with the plagiarizer's name. With reference to Figure 18.1's entry E, here is the rest of the story. After 6 weeks worth of correspondence between the dean of the plagiarizer's college and me, followed by a flurry of activity on the part of APA's legal staff (insofar as copyright infringement was a critical issue in this case), the plagiarizer—with the vise tightening on him—acknowledged committing the offense and admitted to having "made a mistake." Soon thereafter he resigned his university position and moved to a different state. No material financial implications were involved and so no related action was taken by APA. Fortunately, in this particular case, the wheels of justice moved quickly and efficiently toward as equitable a resolution as could have been expected.

## Conclusion

In this chapter I have described, through specific detail, anecdotes, and illustrative examples, several research and publishing malpractices that I have encountered in our academic community. I conclude with the following summary recommendations to individuals who are committed to engaging in the publication process according to professionally acceptable ethical principles:

1. Be well acquainted with the ethical guidelines contained in the APA *Publication Manual* (APA, 2010) and the APA ethics code (APA, 2002). In my experience, although most would-be authors become schooled in the "ins and outs" of proper *manuscript preparation* behaviors (contained primarily in Chapters 2–7 of the current edition of the *Manual* [APA, 2010]), they are not as familiar with the *professional conduct* behaviors (contained in Chapter 1 and on pp. 169–174) that are a fundamental part of the research and publishing process—behaviors that are expected of every participant in that process.
2. Faithfully adhere to these ethical standards in your own research and writing. Perhaps you may succeed in "getting away" with a transgression or two, but as is the case for most criminals, you eventually will be "caught." Whenever that occurs, whatever professional reputation you may have managed to earn will be forever sullied.
3. A recommendation that reflects common courtesy as much as it does professional ethics relates to our previous discussion of assigning authorship credit. In particular, and as I have stated previously: "[E]xplicitly acknowledging the others who contributed in various ways to your own research … accomplishments is a good idea.… Be generous in your giving appropriate credit to the colleagues, students, and other players who are instrumental in helping you achieve your own academic successes. Exhibiting a little more humility than hubris will serve you well in this profession" (Levin, 2004, p. 182).
4. Finally, if you should require clarification or further information regarding research and publishing ethical issues (apart from what is accessible from the APA *Publication Manual* and the APA ethics code), please address your questions to Harris Cooper, a contributing author to this volume (see Cooper & Dent, Chapter 16) and APA's CEA (whose current term runs at least through 2010). And if you should contact Dr. Cooper, it is not necessary for you to

provide a direct reference to anything that I have written in the present chapter!

## References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Barber, T. X. (Ed.). (1973). *Pitfalls in research*. Chicago: Aldine-Atherton.

Bartlett, J. (1992). *Bartlett's familiar quotations* (16th ed.). Boston: Little, Brown.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly, 4,* 358–365.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.

Dorfman, D. D. (1978). The Cyril Burt question: New findings. *Science, 201,* 1177–1186.

Fine, M. A., & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist, 48,* 1141–1147.

Fisher, C. (2003). *Decoding the ethics code: A practical guide for psychologists*. Thousand Oaks, CA: Sage.

Glatt, B. S. (2010). *Glatt plagiarism screening program.* Retrieved from http://www.plagiarism.com

Glatt, B. S., & Haertel, E. H. (1982). The use of the cloze testing procedure for detecting plagiarism. *Journal of Experimental Education, 50,* 127–136.

Goodyear, R. K., Crego, C. A., & Johnston, M. W. (1992). Ethical issues in the supervision of student research: A study of critical incidents. *Professional Psychology: Research and Practice, 23,* 203–210.

Harris, G. (2009, November 19). Academic researchers' conflicts of interest go unreported. *New York Times*, p. A17.

Harris, G. (2010, February 3). Journal retracts 1998 paper linking autism to vaccines. *New York Times*, p. A9.

Hilts, P. J. (1992, January 7). Plagiarists take note: Machine's on guard. *New York Times*, pp. B5, B9.

Levin, J. R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C. J. Brainerd (Eds.), *Cognitive learning and memory in children* (pp. 205–233). New York: Springer-Verlag.

Levin, J. R. (1991a). Editorial. *Journal of Educational Psychology, 83,* 5–7.

Levin, J. R. (1991b, April). *Flair and* savoir faire *in research and publishing*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Levin, J. R. (1992). On research in classrooms. *Mid-Western Educational Researcher, 5,* 2–6, 16.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review, 6,* 231–243.

Levin, J. R. (2003). *Open letter to authors of manuscripts submitted to APA journals*. Retrievable between 2003 and 2008 from http://www.apa.org/journals/authors/openletter.pdf

Levin, J. R. (2004). Random thoughts on the (in)credibility of educational-psychological intervention research. *Educational Psychologist, 39,* 173–184.

Levin, J. R., & Marshall, H. H. (1993). Publishing in the *Journal of Educational Psychology*: Reflections at midstream. *Journal of Educational Psychology, 85,* 3–6.

MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology, 49,* 259–287.

McGue, M. (2000). Authorship and intellectual property. In B. D. Sales & S. Folkman (Eds.), *Ethics in research with human participants* (pp. 75–95). Washington, DC: American Psychological Association.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Sales, B. D., & Folkman, S. (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association.

Sternberg, R. J. (2002). On civility in reviewing. *APS Observer*, *15,* 3–4.

Tangney, J. (2000). Training. In B. D. Sales & S. Folkman (Eds.), *Ethics in research with human participants* (pp. 97–105). Washington, DC: American Psychological Association.

Titus, S. L., Wells, J. A., & Rhoades, L. J. (2008). Repairing research integrity. *Nature, 453,* 980–982.

vos Savant, M. (1992, November 15). Ask Marilyn. *Parade*, p. 27.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61,* 726–728.

Wilson, D. (2009, November 23). Medical schools quizzed on ghostwriting. Retrieved from http://www.nytimes.com/2009/11/18/business/18ghost.html?scp=3&sq=plagiarism&st=cse

Wilson, D., & Singer, N. (2009, November 23). Ghostwriting is called rife in medical journals. Retrieved from http://www.nytimes.com/2009/09/11/business/11ghost.html?ref=business

*Wisconsin State Journal* (1991, March 24). "Fusion" scientist faces uncertain university future. p. 6A.

# *Author Index*


**A**

Abbott, J., 17
Abelson, R., 314
Abramson, P. R., 256
Agresti, A., 111
Aiken, L. S., 1, 2, 131
Aitken, C. G. G., 77
Albert, M., 314, 330
Algina, J., 215
Allen, M. J., 95
Allison, P. D., 359
Altman, D. G., 143, 165, 172, 179, 376, 421, 447, 448
Anderson, K., 456
Angold, A., 280
Angrist, J. D., 398, 399
Antman, E. M., 450
Arkes, H. R., 152
Aronowitz, R., 75
Asch, S. E., 41
Ashby, F. G., 141
Aurelius, M., 21
Avin, C., 409

**B**

Bacchetti, P., 163, 168
Bafumi, J., 87
Bailar, J. C., 144
Bailey, S., 456
Balke, A., 389, 389*n*6
Banaji, M. R., 150, 151, 152
Baraldi, A. N., 371
Barber, T. X., 466
Barbour, V., 144
Baron, R. M., 402
Bartlett, J., 475
Bassler, D., 175
Bath, P. M. W., 159
Bath-Hextall, F. J., 159
Bauer, D. J., 283, 284
Bayer, R., 41
Becker, G., 130
Beecher, H. K., 41
Behnke, S., 45
Beins, B. C., 5
Belia, S., 302
Bellack, A. S., 40
Benjamini, Y., 105
Bennett, B. E., 334
Bennett, D. A., 409
Berk, R. A., 275
Berlin, J. A., 159, 165, 168
Bernhard, J., 363
Bernstein, J., 159
Bersoff, D. N., 5
Bertin, J., 102
Best, J., 108
Beunckens, C., 360
Beutler, L. E., 298
Beyth-Marom, R., 300
Bezeau, S., 159, 173
Bickel, P. J., 111
Bickman, L., 191, 198
Biemer, P., 274
Biesanz, J. C., 176
Birnbaum, A., 133
Blacker, D., 314
Blackman, N. J.-M., 141
Blanck, P. D., 40, 42, 50
Blanton, H., 152
Blyth, C. R., 112
Bogomolny, A., 22
Bok, S., 45
Bolger, N., 282, 402
Borenstein, M., 162, 433
Bormuth, J. R., 479
Borsboom, D., 482*n*9
Boruch, R. F., 188, 189, 190, 192, 244
Bosson, J. K., 151
Bossuyt, P. M., 144
Bowles, R. P., 314, 330
Bowley, A. L., 268

Bragger, J. D., 47*n*5
Braunholtz, D., 167
Breakwell, G. M., 5
Brennan, R. L., 129, 132, 133, 134, 135, 233, 234
Brent, R., 402
Brighton, H., 94
Broekaert, E., 256
Brooks, D., 18
Brooks, J. O., 178
Brown, B. L., 418, 430
Brown, C. H., 175, 402
Brown, R., 251
Brown, T. A., 314, 316, 320, 333
Browne, M., 314, 316, 319, 327
Browne, W. J., 345
Bryk, A. S., 255
Buchanan, M., 77
Bucuvalas, M. J., 243, 244
Bunch, M., 221, 238
Bunch, W., 256, 257
Bunn, F., 159
Burgess, S., 345
Burghardt, J., 256
Burtless, G., 193

**C**

Cai, L., 142
Camilli, G., 211, 218
Campanelli, P., 349
Campbell, B. J., 198
Campbell, D. T., 92, 139, 163, 187, 188, 196, 241, 244, 248, 252, 464
Campbell, S. K., 69
Cannista, S. A., 175
Canter, M. B., 334
Carlsmith, J. M., 160, 165
Carman, J. G., 243, 244
Carroll, J. B., 84
Carroll, R. J., 360
Cassel, C., 269
Catania, J. A., 253
Cattell, R. B., 315, 316, 330, 331, 332, 335
Ceci, S. J., 42
Chalmers, I., 448, 449, 451, 451*n*1
Chalmers, T. C., 193, 450
Chambers, R. L., 274
Champod, C., 77
Chan, A., 172, 179
Chapman, J. P., 85
Chapman, L. J., 85
Charlton, K., 433
Charter, R. A., 130
Chedd, G., 151
Cheng, Y., 276*n*6
Chitwood, D. D., 253
Christ, S., 274
Christie, C., 152
Christman, J., 248
Cizek, G. J., 211, 220, 221, 238
Clark-Carter, D., 159, 173
Coates, T. J., 253
Cochrane, A. L., 449, 451*n*1
Coffman, D. L., 142
Cohen, J., 2, 131, 162, 295, 434, 447
Cohen, P., 131
Collins, L. M., 363, 371
Cook, D. J., 421
Cook, S. W., 42
Cook, T. D., 139, 163, 187, 188, 190, 196, 241, 243, 244, 248, 252, 256, 259, 464
Cooper, H. M., 417, 432, 433, 435, 436, 439, 440, 451
Cortina, J., 87
Cotton, D., 248
Coulson, M., 302
Cousins, J. B., 244
Coverly, D., 31
Cox, D. R., 387, 400
Crego, C. A., 476
Crepaz, N., 144, 376
Crocker, L., 215
Cronbach, L. J., 133, 225, 248, 256
Cudeck, R., 314, 316, 319, 327
Cumming, G., 293, 296, 297, 300, 301, 302, 303, 309, 445, 447
Cumsille, P. E., 366, 367
Curran, P. J., 283, 284

**D**

Daniel, L. G., 135
Davies, H. T. O., 197
Davis, A., 282

Davis, S. F., 5
Dawes, R. M., 89, 93, 118, 119
Dawid, A. P., 394, 396*n*11
Dawkins, N., 248
Delaney, H. D., 131
DeMets, D. L., 175
Demirtas, H., 373, 374
DeNeve, K., 433
Denworth, L., 258
Des Jarlais, D. C., 144, 376
Dickinson, K., 159
Dietz, S., 365
DiMatteo, M. R., 51
DiStefano, C., 142
Dixon, J., 365
Dolan, C. V., 142
Donaldson, S. I., 365
Dorfman, D. D., 465
Dorn, S., 350
Dreyfus, H. L., 258
Dreyfus, S. E., 258
Dunivant, N., 131
Dunn, D. S., 5
Dunn, G., 279
Dunn-Rankin, P., 109
du Toit, M., 276*n*6
du Toit, S. H. C., 276*n*6
Dwyer, J. H., 409

**E**

Ebner-Priemer, U. W., 282
Eckert, W. A., 194
Edmonds, D., 38
Edwards, M. C., 142
Edwards, P., 159
Edwards, S. J. L., 167, 169
Efron, B., 100, 347
Egger, H. L., 280
Eid, M., 282
Eidinow, J., 38
Elashoff, J. D., 162
Elwert, F., 410*n*16
Emanuel, J., 163
Embretson, S. E., 320, 333
Enders, C. K., 357, 358, 359, 360, 363, 365, 367, 368, 371, 373, 374, 375
Enkin, M., 451*n*1
Erlebacher, A., 244
Esanu, J. M., 29
Etzioni, D. A., 159
Eyssell, K. M., 198

**F**

Fabrigar, L. R., 2
Fairchild, A. J., 402
Fairchild, A. L., 41
Fang, G. Y., 374
Farrington, D. P., 190
Faulkner, C., 303
Faulkner, S., 303
Fazio, R. H., 152
Feldt, L. S., 129, 130
Festinger, L., 160, 165
Fidler, F., 293, 295, 296, 297, 300, 301, 302, 303, 445, 447, 459
Fienberg, S. E., 268*n*1
Fife-Schaw, C., 5
Figueredo, A. J., 363
Finch, S., 445, 447
Fine, M. A., 482
Fisher, C. B., 5, 6
Fisher, C. F., 248, 466*n*1
Fisher, G. G., 322
Fisher, R. A., 314
Fitch, J., 21
FitzGibbon, C. T., 351
Fitzpatrick, A. R., 133, 134
Flores, A., 16
Folkman, S., 5, 39, 466, 476
Follmann, D., 282, 285
Foster, E. M., 374
Fowler, F., 253
Frankel, M. R., 275
Frankel, M. S., 19
Franzén, T., 39
Fredericks, K. A., 243, 244
Freedman, D. A., 68, 99
Freedman, K. B., 159
Freedman, L. S., 409
Freeman, M. A., 47*n*5
French, M. T., 257
Fridell, M., 256
Friedman, L. M., 175

Friendly, M., 102
Fritz, M. S., 402
Furberg, C. D., 175
Furberg, C. F., 451

**G**

Gaissmaier, W., 106
Galton, F., 91
Garb, H. N., 119
Garcia, L. T., 5
Gardenier, J. S., 6, 7, 15, 28, 30
Gardner, C., 178
Gawande, A., 86
Geer, J., 455
Gelman, A., 87
Gersten, R., 190, 191
Gibson, D. R., 253
Gigerenzer, G., 66, 78, 94, 106, 108, 117, 159, 295, 447
Gilbert, J. P., 190, 191, 197, 248
Gilbert, R., 452
Ginsburg, A., 244, 255
Glatt, B. S., 479
Gleser, G. C., 133
Gleser, L. J., 435
Glymour, C. N., 384*n*1
Gödel, K., 39
Goldstein, H., 341, 342, 344, 345, 346, 346*n*1, 347, 348, 349, 350, 352
Gong, G., 100, 347
Gonzales, P. M., 152
Goodman, S. N., 159, 168
Goodyear, R. K., 476, 482
Gotzsche, P. C., 172
Gould, S. J., 212
Grady, C., 163
Graham, J. W., 2, 358, 359, 360, 363, 364, 365, 366, 367, 368, 377
Graubard, B. I., 409
Graves, R., 159, 173
Green, S. B., 177
Greene, J. C., 193, 196, 203, 206
Greenland, S., 391, 404
Greenwald, A. G., 150, 151, 152
Grimes, D. A., 167, 168
Grimm, K. J., 314
Grizzard, F. E., Jr., 21
Grob, G., 250
Groves, R. M., 281
Guba, E. G., 248
Guenther, W. C., 170
Gunsalus, C. K., 34
Guyatt, G. H., 175

**H**

Haahr, M. T., 172
Haertel, E. H., 129, 130, 132, 133, 141, 479
Hafeman, D. M., 410*n*16
Hägglund, G., 132
Haller, H., 301, 303, 447
Halpern, S. D., 165, 166, 168
Hambleton, R. K., 134, 215, 223
Hammel, E. A., 111
Hammond, S., 5
Hann, J., 201
Harden, M., 452
Hardisty, D. J., 116
Harlow, L. L., 2, 296, 298, 331
Harris, C., 89
Harris, G., 46*n*4, 483*n*10, 484*n*10
Haslam, A., 5
Hayes, T., 436
Hays, W. L., 23, 61, 62
Healey, M., 302
Heckman, J. J., 283
Hedges, D., 418, 430
Hedges, L. V., 131, 436, 451
Heisey, D. M., 304
Henry, G. T., 190, 191, 203
Herson, H. J., 268
Hesse, M., 256
Hill, C. L., 159
Hill, N., 350
Hilts, P. J., 481
Hintze, J. L., 162
Hitchcock, J., 190, 191
Hoaglin, D. C., 48, 52
Hochberg, Y., 105
Hoenig, J. M., 304
Hofer, S. M., 365
Hoffman, J. M., 402
Holden, C., 119
Holland, P. W., 386, 398, 400*n*15

Holt, D., 274
Holtgrave, D. R., 256
Holton, G., 39
Hope, T., 298
Horn, J. L., 320, 321, 324, 332
Horsten, L., 22
Horton, R., 445, 448
House, E. R., 249, 259
Howe, K. R., 249, 259
Hox, J., 342
Hoyle, R. H., 2, 131
Hròbjartsson, A., 172
Hu, F., 200
Hubbard, R. L., 257
Huff, D., 101
Huitema, B. E., 131
Hunt, M., 450
Hunter, J. E., 159, 167, 301, 453, 454

**I**

Iglewicz, B., 48
Imai, K., 405, 407, 409
Imbens, G. W., 201, 398
Irvine, W. B., 21

**J**

Jaccard, J., 152
Jackson, J., 167
Jaffe, E., 277
James-Burdumy, S., 256
Jansen, I., 360
Jennison, C., 174
Jennrich, R. I., 328
Jensen, A., 269
Jiroutek, M. R., 161, 170
Johnson, E. J., 116
Johnston, M. W., 476
Jones, H. H., 41
Jones, L. V., 170
Jones, N., 249
Jones, R. W., 134
Jones, S. E., 334
Jöreskög, K. G., 131, 316
Judd, C. M., 402
Julnes, G. J., 203
Juster, F. T., 322

**K**

Kadlec, K. M., 322
Kahn, M., 144
Kahnman, D., 108, 114, 115, 116, 447
Kaier, A. N., 268
Kaiser, H., 332
Kam, C.-M., 363
Kane, M., 138, 139, 149, 220, 226
Kaplan, A., 54
Karlawish, J. H. T., 165, 168
Kats, J., 482*n*9
Kazdin, A. E., 5
Keele, L., 405, 409
Keen, H. I., 159
Keirse, M. J. N. C., 451*n*1
Keith-Spiegel, P., 5, 6, 41*n*2, 128
Kelley, K., 159, 162, 170, 178
Kelley, T. L., 94
Kendall, M. G., 87
Kenny, D. A., 92, 131, 402
Kenward, M. G., 360
Kerr, N. L., 172
Kettel-Khan, L., 248
Kimmel, A. J., 5, 41*n*2, 42, 45
Kish, L., 269, 275
Kleindienst, N., 282
Kline, R. B., 296, 298, 300
Knezek, G. A., 109
Knoll, E., 258
Ko, C. Y., 159
Koch, S., 41
Koehler, J. J., 69
Koenig, B. A., 175
Kolata, G., 67, 74, 75
Kolen, M. J., 233, 234
Koocher, G. P., 5, 6, 41*n*2, 42, 128
Kosciulek, J. F., 159, 173
Kosslyn, S. M., 51*n*7
Kovac, M. D., 256
Koval, J. J., 141
Kraemer, H. C., 162, 178, 179
Krämer, W., 78
Krauss, S., 301, 303, 447
Kruskal, W., 270
Kupelnick, B., 450
Kupper, L. L., 161, 170

Kurdek, L. A., 482
Kurz-Milcke, E., 106

**L**

Lahi, M. G., 345
Lai, J., 301, 309
Lau, J., 450
Lautenschlager, G. J., 140
Lavery, J. V., 174, 175
Lavrakas, P., 253
Lawley, D. N., 315, 316, 319
Leckie, G., 348, 352
Lecoutre, B., 331
Lecoutre, M.-P., 331
Leeman, J., 445
Lees-Haley, P. R., 136
Lehman, R. S., 315
Leithwood, K. A., 244
Lemieux, T., 201
Leonardi-Bee, J., 159
Levant, R. F., 298
Levin, J. R., 463, 464, 465, 472, 474, 475, 479, 486, 487, 490
Leviton, L. C., 241, 243, 244, 248, 250, 253, 256, 258, 259
Lewis, C., 316
Li, Y., 409
Liberati, A., 421
Lilford, R. J., 166, 167
Lilienfeld, S. A., 119
Lincoln, Y., 248
Lipsey, M. W., 162, 166, 244, 248, 252, 253, 254, 433
Little, R. J. A., 275, 282, 287, 357, 358, 359
Liu, J. H., 159
Lockwood, C. M., 402
Loftus, G. R., 301
Lohr, S. L., 274
Lok, L., 402
Lord, F. M., 129, 133
Lykken, D. T., 334
Lyles, C., 144, 376

**M**

MacCallum, R. C., 2, 321
MacCoun, R. J., 466
Mace, A. E., 170
MacKinnon, D. P., 365, 402, 409
Maggard, M. A., 159
Maier, S. F., 455
Maillardet, R., 302, 309
Mallinckrodt, C., 360
March, J., 255
Marcus-Roberts, H. M., 141
Margot, P.-A., 77
Mark, M. M., 185, 190, 195, 198, 203
Marshall, H. H., 474, 479, 487
Marshall, M., 143
Martin, M., 7
Maughan, B., 345, 346
Mauland, W., 159
Maxwell, A. E., 315, 316, 319
Maxwell, S. E., 2, 131, 159, 162, 170, 172, 177, 307, 447
May, W. W., 446
Maydeu-Olivares, A., 142
McArdle, J. J., 313, 314, 315, 320, 321, 322, 324, 328, 329, 330, 332
McCartney, K., 167
McClelland, G. H., 176
McConnell, B., 345
McCulloch, C. E., 163
McDonald, R. P., 314, 315, 316, 319, 320
McGarty, C., 5
McGhee, D. E., 150, 152
McGroder, S. M., 244
McGue, M., 476
McKnight, K. M., 363
McKnight, P. E., 363
McPeak, B., 190
McPeek, B., 248
Meade, A. W., 140
Meehl, P. E., 92, 294, 295
Mellenbergh, G. J., 140
Mels, G., 276*n*6
Meredith, W., 321
Messick, S., 136, 139, 140, 224
Michell, J., 141
Miller, D. T., 167
Mills, J. L., 166, 171
Millsap, R. E., 1
Mitcham, C., 272
Miyazaki, I., 212

Moher, D., 144, 179, 376, 421
Molenaar, D., 482*n*9
Molenberghs, G., 287, 360
Mone, M. A., 159, 173
Montague, M., 365
Montori, V. M., 175
Moore, T., 451
Morgan, S. L., 410*n*16
Morris, M., 247
Mortimore, P., 345, 346
Morton, S., 25
Moss, M. B., 314
Mosteller, F., 7, 52, 144, 190, 248, 270, 450
Mueller, G. C., 159
Mueller, P. S., 175
Mulaik, S. A., 2, 296, 314, 315, 331
Muller, D., 402
Muller, K. E., 161
Mun, E. Y., 232
Murphy, K. R., 162
Murray, D. M., 255
Myors, B., 162

**N**

Nagy, T. F., 334
Nakashian, M., 253, 254
Nanda, H., 133
Necowitz, L. B., 2, 321
Needels, K., 256
Nesselroade, J. R., 321, 330
Neustel, S., 214
Newell, D. J., 165, 447
Newell, J. D., 5
Neyman, J., 268, 386, 395
Nickerson, R. S., 298
Nock, M., 272, 282, 283*n*11, 367
Noden, P., 344
Norcross, J. C., 298
Nosek, B. A., 152
Novick, M. R., 129, 133
Nutley, S. M., 197

**O**

Oakes, M. W., 447
O'Brien, R. G., 162
O'Connell, J. B., 159
O'Connell, J. W., 111
Olchowski, A. E., 366
Olkin, I., 131, 435
Olsen, M. K., 282
Olson, M. A., 152
O'Muircheartaigh, C., 349
On, I., 436
Ouston, J., 345, 346
Overton, R. C., 436

**P**

Pan, Z., 170
Panter, A. T., 1, 2
Park, D., 87
Parker, R. A., 253
Parry, P. I., 45, 53
Pashler, H., 89, 93, 100
Patall, E. A., 439, 440
Patton, M. Q., 248
Paz, A., 393
Pearl, J., 383, 384*n*1, 385, 387, 389, 389*n*6, 390, 391, 391*n*8, 392, 392*n*9, 393, 394, 396, 397, 398*n*13, 399, 399*n*14, 403, 404, 405, 406, 407, 408, 409, 410*n*17
Pennebaker, J. W., 151
Peters, D., 42
Petersen, M. L., 405, 410
Peterson, R. A., 277
Peugh, J. L., 360, 375
Pfeffermann, D., 274
Pickles, A., 279
Pile, K., 159
Pinkerton, S. D., 256
Pitoniak, M. J., 223
Pitts, S. C., 176
Plotkin, J., 42
Poehlman, T. A., 151
Poitevineau, J., 331
Polanyi, M., 38
Poole, C., 459
Popper, K. R., 38*n*1
Porter, A. C., 219
Prentice, D. A., 167
Prentice, R., 168

Prescott, C. A., 328, 329
Preskill, H., 249
Prinstein, M. J., 272, 282

**R**

Rachal, J. V., 257
Radelet, M. L., 111
Rafaeli, E., 282
Rajaratnam, N., 133
Rapp, R., 256
Rasch, G., 133
Rath, T., 350
Raudenbush, S. W., 255
Rausch, J. R., 162, 170, 178
Rawls, J., 249
Raymond, M. R., 214
Rea, L. M., 253
Reich, S., 191, 198
Reichardt, C. S., 191
Reis, H. T., 282
Reisch, L. M., 254
Reise, S. P., 320, 333
Rescher, N., 389*n*6
Rhett, N., 244, 255
Rhoades, L. J., 469
Ribisl, K. M., 363
Rivera, R., 250
Roberts, F. S., 141
Roberts, I., 159
Roberts, S., 93, 100
Robins, C. J., 456
Robins, J. M., 391, 395, 399, 404, 410*n*16
Robinson, T. N., 178
Robinson, W. S., 86, 344
Rodgers, J. L., 296
Rog, D., 248
Rosenbaum, P., 393, 397, 398*n*13
Rosenberger, W. F., 200
Rosenthal, R., 5, 37, 38*n*1, 40, 46, 47, 50, 51, 52, 53, 54, 162, 167, 198, 357, 366, 369, 375, 378, 466
Rosnow, R. L., 5, 37, 38*n*1, 40, 42, 46, 47, 47*n*5, 50, 51, 357, 366, 369, 378
Rosoff, P. M., 164
Rossi, J., 159, 447
Rossi, P. H., 193
Rotheram-Borus, M. J., 40, 42
Rothstein, H., 162, 433
Royall, R. M., 268
Roznowski, M., 2, 321
Rubin, D. B., 51, 275, 275*n*4, 357, 358, 359, 363, 364, 386, 391*n*8, 393, 395, 397, 398, 398*n*13, 409
Russell, B., 38, 39
Russell, J. T., 87
Rutter, M., 345, 346

**S**

Sack, K., 75
Sackett, D. L., 421
Salanti, G., 452
Sales, B. D., 5, 39, 466, 476
Sambursky, S., 37
Sammons, P., 344
Sarndal, C., 269
Savitz, D. A., 459
Saxe, L., 45
Schafer, J. L., 2, 282, 358, 359, 360, 363, 364, 365, 373, 374, 377
Schatzkin, A., 409
Scheffe, H., 313, 335
Scheines, R., 384*n*1
Schmidt, F. L., 159, 167, 301, 436, 453, 454
Schneider, J. A., 409
Schoenbaum, M., 202
Schooler, N. R., 40
Schorr, L. B., 248
Schuh, R. G., 250
Schuler, H., 41*n*2, 42
Schulz, K. F., 167, 168, 179, 376
Schwartz, J. K. L., 150, 152
Schwartz, L. M., 106
Schwartz, S., 410*n*16
Schwarz, N., 322
Scott-Jones, D., 50
Scriven, R., 194, 195
Sears, D. O., 277
Sedlmeier, P., 159, 447
See, S., 452
Segal, M. R., 163

Seligman, M. E. P., 455
Seltzer, W., 7
Selvin, H. C., 86
Shadish, W. R., 163, 173, 187, 188, 189, 192, 198, 201, 241, 243, 244, 248, 252, 256, 259
Shiffman, S., 282
Shor, B., 87
Shpitser, I., 393, 409
Shrout, P. E., 402
Sidani, S., 363
Simon, H. A., 389*n*6
Simpson, E. H., 111
Singer, J., 53
Singer, N., 53, 107, 483*n*10
Sinisi, S. E., 45
Skinner, C. J., 274
Skleder, A. A., 47*n*5
Slutsky, A. S., 174, 175
Smith, A., 345, 346
Smith, M. B., 39
Smith, R. A., 5
Smith, T. M. F., 268, 274, 275
Smithson, J. L., 219
Sobel, M. E., 398, 400*n*15
Soltysik, R. C., 118
Sox, H., 144
Spearman, C. E., 317, 324
Spiegelhalter, D. J., 350
Spielmans, G. I., 45, 53
Spirtes, P., 384*n*1
Spitzer, W. O., 421
Stabenow, S., 282
Stam, A., 118
Stanley, J. C., 129, 187, 188, 464
Stapleton, L. M., 322
Stapulonis, R. A., 256
Starobin, P., 45
Steiger, J. H., 2, 296, 331
Steininger, M., 5, 6
Steinley, D., 110
Stephan, F., 268
Sterba, S., 1, 2
Sterba, S. K., 1, 4, 258, 267, 268, 269, 272, 274, 277, 278, 280, 282, 288
Sterling, T. D., 165
Sternberg, R. J., 160, 485
Stevens, J. P., 304
Stewart, P. W., 161
Stigler, S. M., 204
Stoop, I. A., 281
Stout, D., 79
Strohmetz, D. B., 47*n*5
Stroup, D. F., 421
Strube, M. J., 170, 175
Sugden, R. A., 268, 275
Sutton, A. J., 179, 433
Suzman, R., 322
Swaminathan, H., 215
Swann, W. B., 151
Sweeney, P. D., 456
Szymanski, E. M., 159, 173

**T**

Tangney, J., 476
Tanur, J. M., 268*n*1
Tanzi, R., 314
Taroni, F., 77
Tatsuoka, M. M., 332
Taylor, B. J., 366, 367
Taylor, H. C., 87
Telfair, J., 250
Teo, K. K., 451
Tetlock, P. E., 152
Tetzlaff, J., 421
Thaler, R. H., 74
Thiemann, S., 162
Thijs, H., 360
Thissen, D., 142
Thomason, N., 445, 447
Thompson, B., 135
Thompson, J., 151
Thompson, S. G., 360
Thompson, S. J., 229
Thorndike, E. L., 86
Thorndike, R. L., 129
Thurlow, M. L., 229
Tian, J., 393, 399
Tiedeman, D. V., 332
Tierney, J., 151
Tingley, D., 409
Titus, S. L., 469, 470
Todd, P., 201
Tolo, K., 459
Townsend, J. T., 141

Trattner, W. I., 259
Trull, T. J., 282
Tucker, L. R., 316
Tufte, E. R., 51*n*7, 98, 99, 102
Tukey, J. W., 25, 49, 52, 102, 170, 314, 336
Turnbull, B. W., 174
Tversky, A., 108, 114, 115, 116, 447
Tymms, P., 351

**U**

Uchino, B. N., 2
Uhlir, P. F., 29
Uhlmann, E., 151

**V**

Valdiserri, R. O., 253, 256
Valentine, J. C., 432, 433
Van der Klaauw, W., 201
van der Laan, M. J., 405
Vanderplasschen, W., 256
VanderWeele, T. J., 408, 410*n*16
Vazquez-Barquero, J. L., 279
Vedantam, S., 151
Verbeke, G., 287
Vevea, J. L., 436
Visscher, A., 350
von Eye, A., 232
vos Savant, M., 479
Vul, E., 89

**W**

Wagstaff, D. A., 256
Wainer, H., 51*n*7, 61, 81, 93, 94, 102
Wallace, S. R., 109
Walter, I., 197
Wang, W., 402
Ware, J. H., 175
Weaver, C. S., 159
Webb, N. L., 214, 219
Weber, E. U., 116
Wegener, D. T., 2
Weiss, C. H., 243, 244, 252, 259
Wells, J. A., 469
Wells, K. B., 201
Welsh, B. C., 190
Wendler, M. D., 163
Wentz, R., 159
Wermuth, N., 400
West, S. G., 1, 131, 176
Wheeler, L., 282
Whitaker, C. F., 113
White, I. R., 360
Whitehead, A. N., 38, 39
Whitmore, E., 250
Whittaker, J., 400
Wholey, J. S., 248, 249
Wicherts, J. M., 482*n*9
Wilkinson, L., 2, 51, 102, 144, 169, 272, 276, 296, 334, 375, 401
Wilks, S. S., 26
Williams, J., 302, 445
Williams, W. W., 160
Wilson, D., 53, 244, 248, 252, 253, 433, 483*n*10
Wilson, M., 320
Winkielman, P., 89
Winship, C., 410*n*16
Wirth, R. J., 142
Wittgenstein, L., 38
Wolf, L. E., 163
Woloshin, S., 106
Wood, A. M., 360
Wood, J. M., 119
Wretman, J., 269
Wright, B. D., 216
Wright, S., 387
Wright, T. A., 128*n*1
Wright, V. P., 128*n*1
Wu, M., 282, 285

**Y**

Yamamoto, T., 405
Yang, M., 350, 351, 352
Yarnold, P. R., 118
Yates, A., 327
Yen, W. M., 95, 133, 134
Yesavage, J. A., 178
Young, C., 445, 448
Yule, G. U., 84, 86, 87, 111, 112
Yusif, S., 451
Yzerbyt, V. Y., 402

**Z**

Zarkin, G. A., 257
Zelen, M., 175
Zhang, S., 109
Zola, E., 77
Zumbo, B. D., 225

# *Subject Index*


**A**

AAAS. *See* American Association for the Advancement of Science (AAAS)
Absolute risk reduction, 107
Accountability, 42
Accuracy
  of data analysis, 464–466
  in parameter estimation, 160, 161, 165, 168, 169, 170, 171, 180
Actuarial prediction vs. clinical prediction, 92–94
ACYF. *See* Administration on Children, Youth and Families (ACYF)
Adaptive designs, 175
Adaptive randomization, 200
Adaptive sample size planning, 199
Administration on Children, Youth and Families (ACYF), 185
AERA. *See* American Education Research Association (AERA)
Alignment analyses, 219
American Association for the Advancement of Science (AAAS), 15
American Education Research Association (AERA), 145
*American Journal of Epidemiology*, 459
American Psychological Association (APA), 1, 39, 127, 146, 169, 334
  Chief Editorial Advisor
    questionable behaviors observed as, 468–470
  ethical standards adopted by, 44–45
  Journal Article Reporting Standards (JARS), 376
  meta-analysis reporting standards, 420–421
  and moral sensitivities, 40–45
*American Psychologist*, 43, 471
American Statistical Association (ASA), 6, 341
*American Statistician, The*, 6, 102
Analysis of variance (ANOVA) model, 63, 131
Analytic appropriateness, 464–465
ANOVA. *See* Analysis of variance (ANOVA) model
APA. *See* American Psychological Association (APA)
Apophenia, 86
Applied ethics framework, 19
Applied research studies, and program evaluations, 202–203
Appropriateness, 223
  analytic, 464–465
APS. *See* Association for Psychological Science (APS)
Arguing from a vacuum, 119
Argument from ignorance, 119
*Argumentum ad ignorantiam*, 119
ASA. *See* American Statistical Association (ASA)
Assignment of authorship credit, 481–482
Associational concept, 384–385
Association for Computing Machinery
  *Code of Ethics and Professional Conduct*, 20
Association for Psychological Science (APS), 43
Authors, questionable behaviors by, 470–484
  assignment of authorship credit, 481–482
  data falsification, 481
  data sharing, 482–483
  duplicate publication of the same work, 473–474
  piecemeal publication, 472–473
  plagiarism, 475–481
  self-plagiarism, 474–475

simultaneous submission of same manuscript to different journals, 471
Authorship credit, assignment of, 481–482
Autonomy, 248–249
Auxiliary websites, 438–439

**B**

Back-door criterion. *See* Covariate selection
Balance of representation, 220
Basic sampling model, 79–83
Bayes' rule, 67, 73–78
and confusion of conditional probabilities, 76–78
for screening of rare events, 73–76
Behavioral science, standards for, 144–148
*Belmont Report, The*, 40, 42, 46, 50, 189, 191
Beneficence, 248
Berkeley Graduate School Admissions Data
aggregate, 111
six largest majors, 112
Bias
confirmation, 86
lead time, 107, 117
against null hypothesis, 433
overdiagnosis, 107
Biomedical research, origins in, 143–144
Blowing the whistle, 33
Boilerplate-language, 474
Bonferroni inequality, 67
Borrowing strength model, 347

**C**

*Calculated Risks*, 66
Campbell Collaboration, 449
Cancer, colorectal, 66
Categorical concurrence, 219
Causal concept, 384–385
Causal modeling, 383–410
confounding and causal effect estimation, 391
controlled direct effects, 403–404
covariate selection, 391–399
potential outcomes, 395–396
problem formulation, 396–399
structural models, counterfactual analysis in, 393–394
direct vs. total effects, 401–402
distinctions, 384–387
coping with change, 384
formulating, 384–385
ramifications of, 385–386
untested assumptions and new notation, 386–387
Mediation Formula, 407–410
methodological dictates and ethical considerations, 399–401
causal assumptions, explicating, 400–401
target quantity, defining, 399–400
natural direct effects, 404–405
natural indirect effects, 405–407
structural equations as oracles for causes and counterfactuals, 387–391
Causation
token, 389
type, 389*n*7
CC. *See* Colorectal cancer (CC)
Central limit theorem (CLT), 80
CFA. *See* Confirmatory factor analysis (CFA)
Chi-square ($\chi^2$) distribution, 316
CIs. *See* Confidence intervals (CIs)
Citation indexes, 434
Classical test theory (CTT), 129, 215
estimation in statistical analysis, 131–132
estimation of reliability coefficient, 130
Clinical prediction vs. actuarial prediction, 92–94
*Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, 92
Cloze technique, 479

CLT. *See* Central limit theorem (CLT)
Cluster sampling, 271
Cochrane Collaboration, 448–449, 451*n*1
*Code of Fair Testing Practices in Education*, 212, 217, 218, 227, 228, 229, 232, 236
*Code of Professional Ethics & Practices*, 272
*Code of Professional Responsibilities in Educational Measurement*, 212
*Code of Standards and Ethics*, 272
Codes of ethics, 19
Coefficients of generalizability, 133
*Colbert Report, The*, 120
Collective cost, 47
Collective utility, 47
Colorectal cancer (CC), 66
Committee on Standards in Research (CSR), 42
Common-item nonequivalent groups design, 233
Common sense ethics framework, 17
Complex sampling features, 270
Compositional effects, 347
Comte, Auguste, 38*n*1
Concurrence, categorical, 219
Conditional ignorability, 397, 398
Conditional probability, 65
Conditional standard error of measurement (CSEM), 215
Conduct, ethical, 258
Confidence intervals (CIs), 80, 294, 297, 445
Confirmation bias, 86
Confirmatory factor analysis (CFA), 142, 315, 321
Consequence data, 223
Consolidated Standards of Reporting Trials (CONSORT), 143
CONSORT. *See* Consolidated Standards of Reporting Trials (CONSORT)
Construct irrelevant variance, 226
Construct underrepresentation, 226
Content standards, 220
Controlled direct effects, 403–404
Convergent validity, 139
Correlated errors, 320
Correlated specifics (CS), 321
Correlation, 83
  ecological, 86–87
  illusory, 85–86
  odd, 89–90
  between original variables, 83
  restriction of range for, 87–88
Cost, collective, 47
Cost neutrality, 251
Cost–utility model, 357
*Cours de Philosophie Positive*, 38
Covariate selection, 391–399
  potential outcomes, 395–396
  problem formulation, 396–399
  structural models, counterfactual analysis in, 393–394
Crimes, 485–489, 486
Criterion validity, 139
CS. *See* Correlated specifics (CS)
CSEM. *See* Conditional standard error of measurement (CSEM)
CSR. *See* Committee on Standards in Research (CSR)
CTT. *See* Classical test theory (CTT)
Cumulative research, 460
Cut score, 220

**D**

Data
  fabrication, 466–467, 488
  falsification, 481
  impact, 223
  mining, 52, 104
  misreporting, 106–109
  missing, 357–379
  normative, 223
  reality, 223
  sharing, 482–483
Data analysis
  accuracy of, 465–466
  ethical principles in
    and American Psychological Association (APA), 40–45
    ethical and technical standards, intersection, 49–54

risk–benefit process, in research, 46–49
Data analysis, sample selection in, 267–288
narrowing gap between methodological guidelines and practice, 279–287
partially observed selection feature, investigation of, 280–287
sample selection information, recording, 279–280
random and nonrandom, 268–269
reporting about, 269–274
current practice, 272–273
ethical guidelines, 270–272
as ethical issue, 273–274
methodological guidelines, 269–270
statistically accounting for, 274–279
current practice, 277–278
ethical guidelines, 276–277
as ethical issue, 278
methodological guidelines, 274–276
Database match probability (DMP), 70
Data collection, ethical issues, 362–367
auxiliary variables, role of, 363–365
documentation of reasons, 365
planned designs, 365–367
prevention of problem, 363
Data presentation and analyses, 95–101
graphical presentation, 101–102, 103
misreporting data, 106–109
multiple testing, problem of, 102–106
multivariable systems, 99–101
software implementations, pitfalls of, 109–110
Data snooping, 174
Data World framework, 79
Decision-plane model, 47, 48
*Declaration on Professional Ethics*, 272
Deductive disclosure, 235
Degree of freedom (*df*), 96, 316
Dependent variable (DV), 313
Depth of knowledge, 219–220
Descriptive statistics, 23
Designs
adaptive, 175
matrix, 347
planning missingness, 363, 378
rotation, 347
*Developmental Psychology*, 272
Diagnoses, 352
Dichotomous thinking, 293
DIF. *See* Differential item functioning (DIF)
Differential item functioning (DIF), 218
Direct effects
controlled, 403–404
natural, 404–405
vs. total effects, 401–402
Direct maximum likelihood, 360–361
Disproportionate selection, 271, 274, 275
DMP. *See* Database match probability (DMP)
*DNA Technology in Forensic Science*, 70
Duplicate publication, 473–474
DV. *See* Dependent variable (DV)

**E**

EBM. *See* Evidence-based medicine (EBM)
EBP. *See* Evidence-based practice (EBP)
Ecological correlation, 86–87
Ecological fallacy, 86, 344
*Economist, The*, 196
Editors, questionable behaviors by, 484–485
EFA. *See* Exploratory factor analysis (EFA)
*Effectiveness and Efficiency*, 449
Effect size, 445
averaging and weighting methods, 435
metrics, 434
variation among, 436–437
EM. *See* Episodic memory (EM)
Episodic memory (EM), 326
Equating, 233–234
Equipercentile equating,234

Error
correlated, 320
false-negative, 224
false-positive, 224
fundamental attribution, 117
measurement, 26, 27–28
nonsampling, 27
sampling, 27
Ethical and technical standards, intersection, 49–54,
Ethical conduct vs. evaluation quality, 258
Ethical guidelines, 19
*Ethical Guidelines for Statistical Practice*, 15, 20, 31, 272
Ethical issues
in analysis, 254–255
to improve quality and serve ethics, 251–252
in interpretation and reporting of results, 255–257
in measurement and design, 252–253
in quantitative data collection, 253–254
in quantitative evaluations, 251–257
Ethical principles
application to evaluation, 248–250
definitions of, 248–250
*Ethical Principles in the Conduct of Research with Human Participants*, 39
*Ethical Principles of Psychologists and Code of Conduct*, 127, 272, 334, 418, 419
Ethical reporting standards, 143
Ethics
vs. ideology, 259
professional ethics, principles of, 31–34
in quantitative methodology
research ethics resources, 3–4
textbooks about ethics, 4–6
in quantitative professional practice, 15
research. *See* Research ethics
and sample size planning, 159
scientific research, 27–30
*Evaluation of Forensic DNA Evidence, The*, 70
Evaluation quality
vs. ethical conduct, 258
vs. technical quality, 258–259
Event-contingent selection, 282
Evidence-based medicine (EBM), 309
Evidence-based practice (EBP), 294
Exclusion restriction, 397
Experimental rigor, 464
Exploratory factor analysis (EFA), 315, 327–328
External evidence, 224

**F**

Fabrication, falsification, and plagiarism (FF&P), 29
Face validity, 139
Factor analysis, 313–336
case study, 322–331
confirmatory, 142, 315, 321
exploratory, 315, 327–328
item, 142
methodological issues, 313–315
one-factor concept, variations on, 320–321
prior work, 331–333
statistical basis of, 315–316
structural, 314
expanding, 321–322
initial, 316–320
Factor loadings, 316, 318
Factor rotation, 322
Fairness, 249
Fallacy of the Transposed Conditional, 76
False dilemma, 120
False-negative error, 224
False-positive error, 224
Falsification. *See* Data falsification; Fabrication, falsification, and plagiarism (FF&P)
Family Educational Rights and Privacy Act (FERPA), 212, 235
Fecal occult blood test (FOBT), 66
Federal Judicial Center, 192, 193, 194, 196, 197

FERPA. *See* Family Educational Rights and Privacy Act (FERPA)
FF&P. *See* Fabrication, falsification, and plagiarism (FF&P)
Fidelity, procedural, 223
*Fifteen Thousand Hours*, 345, 346
"File drawer" effect, 119, 178
FOBT. *See* Fecal occult blood test (FOBT)
Formative evaluation, 248
Fragmented publication. *See* Piecemeal publication
Full information maximum likelihood, 360–361
Fundamental attribution error, 117

**G**

Generalizability theory (GT), 132–133
Generic $2 \times 2$ contingency table, 65
Geomin criterion, 327–328
Graduate Record Examination (GRE) scores, 160
Graduate Record Examination–Verbal (GRE-V) scores, 90
GRE. *See* Graduate Record Examination (GRE) scores
GRE-V. *See* Graduate Record Examination–Verbal (GRE-V) scores
GT. *See* Generalizability theory (GT)
GT decision study (D-study), 133
*Guinness Book of World Records*, 473

**H**

HARKing, 172
Health and Retirement Study (HRS)
  cognitive measurement in, 322–331
Heat maps, 102
*Helping Doctors and Patients Make Sense of Health Statistics*, 106
*Helplessness: On Depression, Development and Death*, 449
Hierarchical linear model, 343
High school grade point averages (HSGPAs), 100
High-stakes assessment, and psychometric methods, 211
*How to Lie with Statistics*, 102
HRS. *See* Health and Retirement Study (HRS)
HSGPAs. *See* High school grade point averages (HSGPAs)
Hypothesis testing. *See* Null hypothesis significance testing (NHST)

**I**

IAT. *See* Implicit Association Test (IAT)
Ideology
  vs. ethics, 259
  of limits, 39
IFA. *See* Item factor analysis (IFA)
Ignorability
  conditional, 397, 398
  demystification of
    potential outcomes and, 395–396
    problem formulation and, 396–399
Illusory correlation, 85–86
Impact data, 223
Implicit Association Test (IAT), 150
Inclusive analysis strategy, 363, 370–371
Indirect effects, natural, 405–407
Individual participant data (IPD), 439–440
Inductive-hypothetico-deductive spiral, 332
Inferential statistics, 23
Institutional review boards (IRBs), 1, 42, 164, 205
Integrity, 50
  researcher, 465–466
Internal evidence, 223
International Committee of Medical Journal Editors, 445
International Statistical Institute (ISI), 269, 341
*Introduction to the Theory of Statistics, An*, 84, 86, 112
Invariance, measurement, 140
Inversion fallacy, 76

IPD. *See* Individual participant data (IPD)
IRBs. *See* Institutional review boards (IRBs)
Irrelevant variance, 139
  construct, 226
IRT. *See* Item response theory (IRT) model
ISI. *See* International Statistical Institute (ISI)
Item factor analysis (IFA), 142
Item response theory (IRT) model, 133–134, 215
Item–task construction and evaluation, 217–219

**J**

*Journal of Abnormal Psychology*, 272, 455
*Journal of Educational Psychology*, 272, 466, 468
*Journal of Personality and Social Psychology*, 150, 272
*Journal of the American Medical Association*, 165
Journal reviewers, questionable behaviors by, 484–485
Justice, 50
  social, 249–250

**K**

Kelley's Paradox, 94

**L**

*Lancet, The*, 445, 448, 460, 484
Latent variable path analysis (LVP), 329
Law of large numbers (LLN), 80
Law of the instrument, 54
Lead time bias, 107, 117
Learned helplessness, 455
Least publishable units (LPUs), 486
Likelihood ratio test (LRT), 319
LLN. *See* Law of large numbers (LLN)
LPUs. *See* Least publishable units (LPUs)
LRT. *See* Likelihood ratio test (LRT)
LVP. *See* Latent variable path analysis (LVP)

**M**

Magnetic resonance imaging (MRI), 366
Major, John, 350
MANOVA. *See* Multivariate analysis of variance (MANOVA)
MAR. *See* Missing at random (MAR) mechanism
MARS. *See* Meta-analysis reporting standards (MARS)
Mathematics, assumptions in, 22–23
Matrix designs, 347
Maturity, 18
MAUP. *See* Modifiable areal unit problem (MAUP)
Maximum likelihood estimation (MLE), 315, 360–361
MCAR. *See* Missing completely at random (MCAR) mechanism
Mean equating, 234
Measurement error, 26, 27–28
Measurement invariance, 140
Mediation Formula, 407–410
Medicine
  research, ethics and sample size planning in, 164–171
  use of ethical imperative for statistical reform in
    changes in statistical thinking and interpretation, 459
    estimation and cumulative analysis, 460
    ethical, technical, and philosophical motivation, 458–459
Mental status (MS), 326
Meta-analysis, 445, 448–449
  APA standards, 420–421, 422–425
  approaching ethical obligation, 430–437
    effect, measure of, 434–436
    effect sizes, variation among, 436–437

inclusion criteria, 431–433
literature search parameters, 433–434
problem statement, 430–431
tabling data, 437
auxiliary websites, use of, 438–439
duplicate publication, uncovering, 441
with individual participant data and aggregate statistics, 439–440
reporting, 417–442
results interpretation, 437
Society for Research Synthesis Methodology survey, 421, 426–430
space limitations, 438–439
Meta-analysis of Observational Studies in Epidemiology (MOOSE), 421
Meta-analysis reporting standards (MARS), 420–421, 422–425, 449
Methodologists, ethical framework for
quantitative professional practice, ethics in, 15
applied professional ethics, principles of, 31–34
general frameworks, 16–22
mathematics, assumptions in, 22–23
scientific research ethics, 27–30
statistics, assumptions in, 24–27
MI. *See* Modification indices (MI)
Mismatched framing, 108
Misreporting, data, 106–109
Missing at random (MAR) mechanism, 358, 378
Missing completely at random (MCAR) mechanism, 358, 378
Missing data, 357–379
data analysis, ethical issues, 368–374
analysis options, 371–374
imputation, 370
inclusive analysis strategy, 370–371
quantity of missing data, 368–370
sensitivity analysis, 374
design and data collection, ethical issues, 362–367
auxiliary variables, role of, 363–365
documentation of reasons, 365
planned designs, 365–367
prevention of problem, 363
mechanisms, 357–359
reporting, ethical issues, 375–378
overstating benefits of analytic technique, 376–378
standards, 375–376
techniques, 359–362
atheoretical methods, 359–360
MAR-based methods, 360–361
MCAR-based methods, 360
MNAR-based methods, 361–362
Missing not at random (MNAR) mechanism, 358, 378
Mixed model, 343
MLE. *See* Maximum likelihood estimation (MLE)
MNAR. *See* Missing not at random (MNAR) mechanism
Modifiable areal unit problem (MAUP), 87
Modification indices (MI), 320
Monte Hall problem, 113, 114
MOOSE. *See* Meta-analysis of Observational Studies in Epidemiology (MOOSE)
Moral outrage, 18
Morton, Sally, 25
MRI. *See* Magnetic resonance imaging (MRI)
MS. *See* Mental status (MS)
Multilevel modeling, 341–353
case history, 349–352
data analyst, role of, 348–349
hierarchical structures, designing studies with, 345–346
substantive issues, clustered designs for, 346–348
Multiple imputation, 360–361
Multiple membership models, 345

Multiple testing, problem of, 102–106
Multivariable systems, 99–101
Multivariate analysis of variance (MANOVA), 105

**N**

National Institutes of Health (NIH), 19, 469
National Research Council (NRC), 70
  Committee on DNA Technology in Forensic Science, 70
National Science and Technology Council, 29
National Science Foundation (NSF), 19
National Society of Professional Engineers
  *NSPE Code of Ethics for Engineers*, 20
Natural direct effects, 404–405
Natural indirect effects, 405–407
*Nature*, 469
*New York Herald Tribune*, 69
*New York Times, The*, 67, 75, 77, 483
Neyman–Rubin potential–outcome framework, 394
NHST. *See* Null hypothesis significance testing (NHST)
NIH. *See* National Institutes of Health (NIH)
Noncontinuous scores, 141–143
Nonmaleficence, 248
Nonsampling error, 27
Nonsuicidal self-injury (NSSI) behaviors, 281
Normative data, 223
NRC. *See* National Research Council (NRC)
NSF. *See* National Science Foundation (NSF)
NSSI. *See* Nonsuicidal self-injury (NSSI) behaviors
Null hypothesis significance testing (NHST), 445
  basic argument, 294–298
    APA *Publication Manual*, 296–297
    prospects for change, 297–298
  cognitive evidence for, 300–301
  vs. estimation, 293–310
  example, 305–309
  formulation of arguments, 298–300
  misinterpretation of, 445, 449–457
  misuse of, 445
  neglect of statistical power, 446–448
    costs to public welfare, 458
    costs to science of overreliance, 457–458
  statistical cognition research, 303–305
Number needed to treat, 107

**O**

ODA. *See* Optimal data analysis (ODA)
Odd correlation, 89–90
Office of Research Integrity (ORI), 46*n*4
One-factor concept, variations on, 320–321
"Opportunity to learn" concept, 214
Optimal data analysis (ODA), 118
*Opus Majus*, 37
ORI. *See* Office of Research Integrity (ORI)
Ought implies can principle, 273
Overdiagnosis bias, 107

**P**

*Parade*, 113
Parameter estimation
  accuracy in, 160, 161, 165, 168, 169, 170, 171, 180
  cognitive evidence about, 301–303
  vs. NHST. *See under* Null hypothesis significance testing (NHST)
Paramorphic representation, 93
Paraphragiarism, 474, 476–477, 488–489
Pareidolia, 86
Parents charter, 350
Participatory evaluation, 250
Paternalism, 249
Path diagram, 388
Path-specific effects, 406
Pattern mixture model framework, 361

Pearson's product moment correlation coefficient, 83
Performance standards, 220
Personal introduction to research ethics, 463
*Perspectives on Psychological Science*, 89
Piecemeal (fragmented) publication, 472–473
Plagiarism, 467–468, 475–481, 488–489
  avoiding, 478–481
  blatant, 489
  intentional, 489
  paraphragiarism, 474, 476–477, 488–489
  self-plagiarism, 474–475
Plagiarism Screening Program, 479
Planning missingness designs, 363, 378
*PLoS Medicine*, 53
Positive predictive value (PPV), 73
Positivism, 38*n*1
Postintervention distribution, 389
Potential outcome approach, 395
Potsdam consultation on meta-analysis, 421
Power analysis, 160, 162, 169, 170
PowerPoint (PP), 98
PP. *See* PowerPoint (PP)
PPV. *See* Positive predictive value (PPV)
Prediction, 90–95
  effects in selection, 95
  reliability corrections in, 94–95
Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA), 421
*Primer on Regression Artifacts, A*, 92
Principal stratification, 409
*Principia Mathematica*, 38
Principled discovery, 203
Principles of professional ethics, 31–34
PRISMA. *See* Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA)
Probability
  conditional, 65
  database match, 70
  random match, 69
Probability theory, 63–67
  Bayes' rule, 73–78
    and confusion of conditional probabilities, 76–78
    for screening of rare events, 73–76
  probabilistic generalizations, 71–73
  probabilities, assignment of, 68–71
Problem of "three caskets," 114
Procedural fidelity, 223
Professional ethics, principles of, 31–34
Program evaluation, ethics in
  distinguishing ethical, technical, and ideological issues in, 257–259
  ethical issues
    in analysis, 254–255
    to improve quality and serve ethics, 251–252
    in interpretation and reporting of results, 255–257
    in measurement and design, 252–253
    in quantitative data collection, 253–254
    in quantitative evaluations, 251–257
  ethical principles
    application to evaluation, 248–250
    definitions of, 248–250
  overview of, 242–246
  literature linking ethics to, 246–247
*Program Evaluation Standards*, 246
Prosecutor's fallacy, 65, 76
*Psychological Bulletin*, 162
*Psychological Science*, 116
*Psychological Science in the Public Interest*, 106
Psychology, use of ethical imperative for statistical reform in
  changes in statistical thinking and interpretation, 459
  estimation and cumulative analysis, 460
  ethical, technical, and philosophical motivation, 458–459

*Psychology: A Study of a Science*, 41
Psychometric methods, and high-stakes assessment, 211
item–task construction and evaluation, 217–219
psychometric model, 215–217
standard setting, 220–224
test administration, 227–238
conditions, 228–230
test development, 213
identification of, 213–215
test form, development of, 219–220
test scoring, 227–238
comparability, 233–234
confidentially, 234–235
integrity, 235–238
procedures, 230–232
report, 232–238
validation, 224–226
Psychometric model, 215–217
Psychometric properties, definition of, 145
*Publication Manual of the American Psychological Association*, 145–146
Publication process, sanctity of, 464
Punishments, 485–489
Pursuit of happiness framework, 21

**Q**

QAV. *See* Quantitative assignment variable (QAV)
Quantitative assignment variable (QAV), 201
Quantitative professionalism, 31
Quantitative professional practice, ethics in, 15
general frameworks, 16–22
mathematics, assumptions in, 22–23
professional ethics, principles of, 31–34
scientific research ethics, 27–30
statistics, assumptions in, 24–27
*Quarterly Journal of Political Science*, 87
Quasi-experiments, 186–189
ethicality of, 205–206
in field settings, 185, 189
methodological advances in, 199–202
principled discovery, 203–204
"Queen of the sciences," 39
Questionable behaviors
by journal reviewers and editors, 484–485
observed as APA's chief editorial advisor, 468–470
by researchers and authors, 470–484
QUOROM Statement (Quality of Reporting of Meta-analysis), 421

**R**

Random encouragement design, 201, 202
Random groups design, 233
"Random intercept" model, 343
Randomized experiments, 186–189
ethical argument for, 189–192
ethical criticisms of, 192–198
ethicality of, 205–206
in field settings, 185, 189
methodological advances in, 199–202
principled discovery, 203–204
value-based outcomes, 204–205
Random match probability (RMP), 69
Random sampling, 24
Rangefinding, 231
Range of knowledge correspondence, 220
Range restriction problem, 88
RCR. *See* Responsible Conduct of Research (RCR)
Reality data, 223
Reference databases, 434
Regression–discontinuity design, 201, 202
Regression toward the mean, 82, 91–92
Regressive fallacy, 91
Relative risk reduction, 107
Reliability, 129–135. *See also* Validity
classical test theory (CTT), 129

estimation in statistical analysis, 131–132
estimation of reliability coefficient, 130
corrections in prediction, 94–95
estimation, applications of in statistical analysis, 131–132
generalizability theory (GT), 132–133
item response theory (IRT) model, 133–134
Reliability coefficient, 129
estimation of, 130
Reporting standards
data analysis, 269–274
ethical, 143
meta-analysis, 420–421, 422–425, 449
scientific, 143
Requirement, definition of, 420
Research
biomedical, origins in, 143–144
cumulative, 460
ethical dimension of, 163–164
ethics
personal introduction to, 463
resources, 3–4
Researchers
integrity, 465–466
questionable behaviors by, 470–484
assignment of authorship credit, 481–482
data falsification, 481
data sharing, 482–483
duplicate publication of the same work, 473–474
piecemeal publication, 472–473
plagiarism, 475–481
self-plagiarism, 474–475
simultaneous submission of the same manuscript to different journals, 471
Research misconduct, 29, 46*n*4, 469–470
Research quality, as ethical issue, 198–199
Responsible Conduct of Research (RCR), 3
*Rights and Responsibilities of Test Takers: Guidelines and Expectations*, 212, 228, 229, 235, 236
Risk, 46
Risk–benefit process, in research, 46–49
RMP. *See* Random match probability (RMP)
RMSEA. *See* Root mean square error of approximation (RMSEA)
Root mean square error of approximation (RMSEA), 304
*Rosemary's Baby*, 108
Rotation designs, 347
Royal Statistical Society (RSS), 341
RSS. *See* Royal Statistical Society (RSS)
Rule of total probability, 67


## S


Sample selection
conditionally ignorable, 275
in data analysis. *See* Data analysis, sample selection in
equal/unequal probabilities of, 271
ignorable, 275
mechanism of, 282–287
multiple phases of selection, 271
nonignorable, 275–276
Sample size planning
adaptive, 199
and ethics, 159
inadequate resources, 176–179
medical research, perspective from, 164–171
research, ethics in, 163–164
statistical significance, 172–176
Sampling
cluster, 271
complex features, 270
distributions, 80
error, 27
frame, 25, 271
random, 24
stratified, 271
units, 271
SAT. *See* Scholastic Aptitude Test (SAT) scores

Scholastic Aptitude Test (SAT) scores, 100
School league tables, 349–352
*Science and Engineering Ethics*, 30
Scientific record, 28
Scientific reporting standards, 143
Scientific research ethics, 27–30
Selection mechanism, definition of, 268
Selection model framework, 361
Selective reporting, 166
Self-plagiarism, 474–475. *See also* Plagiarism
SEM. *See* Structural equation modeling (SEM)
Sensitivity analysis, 73, 283–287, 374
SFA. *See* Structural factor analysis (SFA)
Shared parameter model, 282
"Shifting unit of analysis" approach, 436
SIDS. *See* Sudden infant death syndrome (SIDS)
Signal-contingent selection, 282
Signal sensing, 406
Simple random sample, 270
Simpson's Paradox, 111–113
"Simultaneous submission" prohibition policy, 471
Situational specificity, 453
Smoking, dangers of, 103
Social justice, 249–250
Society for Research Synthesis Methodology
  survey, meta-analysis, 421, 426–430
Software implementations, pitfalls of, 109–110
Specificity, 73
*Sports Illustrated*, 92
Standard, definition of, 420
*Standards for Educational and Psychological Testing*, 135, 212, 217, 220
*Standards for Reporting on Empirical Social Science Research in AERA Publications*, 375
Standards for the Reporting of Diagnostic Accuracy Studies (STARD), 144
Standards of practice, 19
STARD. *See* Standards for the Reporting of Diagnostic Accuracy Studies (STARD)
Statistical conclusion validity, 464
Statistical guide for ethically perplexed
  actuarial vs. clinical prediction, 92–94
  basic sampling model, 79–83
  correlation, 83
    ecological, 86–87
    illusory, 85–86
    odd, 89–90
    restriction of range for, 87–88
  data presentation and analyses, 95–101
    graphical presentation, 101–102
    misreporting data, 106–109
    multiple testing, problem of, 102–106
    multivariable systems, 99–101
    software implementations, pitfalls of, 109–110
  prediction, 90–95
    effects in selection, 95
    unreliability corrections in, 94–95
  probability theory, 63–67
    Bayes' rule, 73–78
    probabilistic generalizations, 71–73
    probabilities, assignment of, 68–71
  regression toward the mean, 91–92
  Simpson's Paradox, 111–113
Statistical independence, 63
Statistical inference, 446–448
*Statistical Methods in Psychology Journals: Guidelines and Explanations*, 272
Statistical power, neglect of, 446–448
Statistical prediction. *See* Actuarial prediction
Statistical reform
  ethical imperative of, 445–460
    costs to public welfare, 458

costs to science of overreliance, 457–458
lack of publicity, 457
linkages between statistical practice and ethics, 452–453
meta-analysis, 448–449
null hypothesis significance testing, misinterpretation of, 449–451
poor practice, 453–456
proximity of experimental outcomes to utilitarian consequences, 456–457
stakes, 457
statistical inference, 446–448
Statistical significance, 459
results, 172–176
Statistics
assumptions in, 24–27
descriptive, 23
inferential, 23
Stop rule, 199–200, 202
Storandt, Martha, 468–469
Stratified sampling, 271
Structural equation modeling (SEM), 304, 315–316
Structural factor analysis (SFA), 314
Sudden infant death syndrome (SIDS), 64, 452
Sufficient set, 391
Summative evaluation, 248
Systematic error. *See* Irrelevant variance

**T**

Technical and ethical standards, intersection, 49–54
Technical quality vs. evaluation quality, 258–259
*Technical Recommendations for Psychological Tests and Diagnostic Techniques*, 212
Test administration, 227–238
conditions of, 228–230
Testing accommodations, 229
Testing modifications, 229
Test scoring, 227–238
comparability, 233–234
confidentially, 234–235
integrity, 235–238
procedures, 230–232
report, 232–238
Texas Sharpshooter fallacy, 86
Textbooks, about ethics, 4–6
Theory of situational specificity (TSS), 453
Theory World framework, 79
Token causation, 389
Total effects vs. direct effects, 401–402
Trait underrepresentation, 139
Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) statement, 143
TREND. *See* Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) statement
*Trials of War Criminals Before the Nuernberg Military Tribunals*, 40
Truthiness, 120
TSS. *See* Theory of situational specificity (TSS)
Two-part model, 282
Type causation, 389*n7*

**U**

UGPA. *See* Undergraduate grade point average (UGPA)
UN. *See* United Nations (UN)
Undergraduate grade point average (UGPA), 88
Underrepresentation
construct, 226
trait, 139
Unified view of validity, 225
United Nations (UN)
Economic and Social Council, 269
Units of analysis problem, 346
Universe of generalization, 132
Universe score variance, 133
U.S. Office of Research Integrity, 29

USPHS. *See* U.S. Public Health Service (USPHS)
U.S. Public Health Service (USPHS), 41
Utility, collective, 47

**V**

Validation, 224–226
  samples, 232
Validity, 129, 135–137, 224. *See also* Reliability
  convergent, 139
  criterion, 139
  definition of, 136
  face, 139
  recommendations, 140
  score interpretation, 137–140
    extrapolation, 139
    generalization, 138–139
    implication, 139
    scoring, 138
  score relevance, 140
  statistical conclusion, 464
  threats to, 139–140
  unified view of validity, 225
Variance
  irrelevant, 139, 226
  universe score, 133
Variance components model, 343
Verbatim copying, 474
Verificationism, 38*n*1

**W**

*Wall Street Journal*, 21
Wechsler Adult Intelligence Scale, 238
Wherry's shrinkage formula, 101
Whistleblowers, 29

**Y**

Yule's Q, 85